Promoting access to public research data for scientific, economic, and social development

Organisation de Coopération Et de Développement Économiques (ocde)

Promoting Access to Public Research Data for Scientific, Economic, and Social Development

OECD Follow Up Group on Issues of Access to
Publicly Funded Research Data

Final Report

March 2003

Acknowledgements

Participants from the United States are being funded through NSF Grant ACI-9619020 to help coordinate this group's activities and develop a series of case studies to explore data access issues. Additional support for the group's activities comes from Netherlands' Ministry of Education, Culture and Science, and governments and agencies of the Follow-up Group's members.

The CODATA Secretariat hosted a meeting of the Follow-up Group in March 2002 in Paris. CODATA is the interdisciplinary Committee on Data for Science and Technology of the International Council of Scientific Unions (ICSU).

The Polish State Committee for Scientific Research hosted the 4th Global Research Village Conference (October 2002), which provided a forum to discuss the Follow-up Groups' Interim Report.

The European Science Foundation (ESF) is a partner in this activity. ESF hosted a meeting of representatives of member funding agencies in October 2002 in Strasbourg, France. The meeting provided a forum for discussion of the Interim Report of the Follow-up Group, and resulted in continued discussions at the funding agencies themselves and, in some cases, new policies (see Box 1). In addition, ESF helped solicit input on agency policies for this report.

Finally, The Ministry of Education, Culture and Science from The Netherlands funded a web survey of current data policies as well as a number of case studies. These efforts were directed by Paul Wouters of NIWI-KNAW, The Netherlands. Their publication in the series *The Public Domain of Digital Research Data* also contributed to the activities of this Follow-up Group on Issues of Access to Publicly Funded Research Data.

Executive Summary

It is now commonplace to say that information and communications technologies are rapidly transforming the world of research. We are only beginning to recognize, however, that management of the scientific enterprise must adapt if we, as a society, are to take full advantage of the knowledge and understanding generated by researchers. One of the most important areas of information and communication technology (ICT)-driven change is the emergence of e-science, briefly described as universal desktop access, via the Internet, to distributed resources, global collaboration, and the intellectual, analytical, and investigative output of the world's scientific community.

The vision of e-science is being realised in relation to the outputs of science, particularly journal articles and other forms of scholarly publication. This realisation extends less to research data, the raw material at the heart of the scientific process and the object of significant annual public investments.

Ensuring research data are easily accessible, so that they can be used as often and as widely as possible, is a matter of sound stewardship of public resources. Moreover, as research becomes increasingly global, there is a growing need to systematically address data access and sharing issues beyond national jurisdictions. The goals of this report and its recommendations are to ensure that both researchers and the public receive optimum returns on the public investments in research, and to build on the value chain of investments in research and research data.

To some extent, research data are shared today, often quite extensively within established networks, using both the latest technology and innovative management techniques. The Follow Up Group drew on the experiences of several of these networks to examine the roles and responsibilities of governments as they relate to data produced from publicly funded research. The objective was to seek good practices that can be used by national governments, international bodies, and scientists in other areas of research. In doing so, the Group developed an analytical framework for determining where further improvements can be made in the national and international organization, management, and regulation of research data.

The findings and recommendations presented here are based on the central principle that ***publicly funded research data should be openly available to the maximum extent possible.*** Availability should be subject only to national security restrictions; protection of confidentiality and privacy; intellectual property rights; and time-limited exclusive use by principal investigators. Publicly funded research data are a public good, produced in the public interest. As such they should remain in the public realm. This does not preclude the subsequent commercialization of research results in patents and copyrights, or of the data themselves in databases, but it does mean that a copy of the data must be maintained and made openly accessible. Implicitly or explicitly, this principle is recognized by many of the world's leading scientific institutions, organizations, and agencies. Expanding the adoption of this principle to national and international stages will enable researchers, empower citizens and convey tremendous scientific, economic, and social benefits.

Evidence from the case studies and from other investigation undertaken for this report suggest that successful research data access and sharing arrangements, or regimes, share a number of key

attributes and operating principles. These bring effective organization and management to the distribution and exchange of data. The key attributes include: openness; transparency of access and active dissemination; the assignment and assumption of formal responsibilities; interoperability; quality control; operational efficiency and flexibility; respect for private intellectual property and other ethical and legal matters; accountability; and professionalism. Whether they are discipline-specific or issue oriented, national or international, the regimes that adhere to these operating principles reap the greatest returns from the use of research data.

There are five broad groups of issues that stand out in any examination of research data access and sharing regimes. The Follow Up Group used these as an analytical framework for examining the case studies that informed this report, and in doing so, came to several broad conclusions:

- Technological issues: Broad access to research data, and their optimum exploitation, requires appropriately designed technological infrastructure, broad international agreement on interoperability, and effective data quality controls;
- Institutional and managerial issues: While the core open access principle applies to all science communities, the diversity of the scientific enterprise suggests that a variety of institutional models and tailored data management approaches are most effective in meeting the needs of researchers;
- Financial and budgetary issues: Scientific data infrastructure requires continued, and dedicated, budgetary planning and appropriate financial support. The use of research data cannot be maximized if access, management, and preservation costs are an add-on or after-thought in research projects;
- Legal and policy issues: National laws and international agreements directly affect data access and sharing practices, despite the fact that they are often adopted without due consideration of the impact on the sharing of publicly funded research data;
- Cultural and behavioural issues: Appropriate reward structures are a necessary component for promoting data access and sharing practices. These apply to both those who produce and those who manage research data.

The case studies and other research conducted for this report suggest that concrete, beneficial actions can be taken by the different actors involved in making possible access to, and sharing of, publicly funded research data. This includes the OECD as an international organization with credibility and stature in the science policy area. The Follow Up Group recommends that the OECD consider the following:

- Put the issues of data access and sharing on the agenda of the next Ministerial meeting;
- In conjunction with relevant member country research organizations,
    - Conduct or coordinate a study to survey national laws and policies that affect data access and sharing practices;
    - Conduct or coordinate a study to compile model licensing agreements and templates for access to and sharing of publicly funded data;
- With the rapid advances in scientific communications made possible by recent developments in ICTs, there are many aspects of research data access and sharing that have not been addressed sufficiently by this report, would benefit from further study, and will need further clarification. Accordingly, further possible actions areas include:
    - Governments from OECD expand their policy frameworks of research data access and sharing to include data produced from a mixture of public and private funds;

- OECD consider examinations of research data access and sharing to include issues of interacting with developing countries; and
- OECD promote further research, including a comprehensive economic analysis of existing data access regimes, at both the national and research project or program levels.

National governments have a crucial role to play in promoting and supporting data accessibility since they provide the necessary resources, establish overall polices for data management, regulate matters such as the protection of confidentiality and privacy, and determine restrictions based on national security. Most importantly, national governments are responsible for major research support and funding organizations, and it is here that many of the managerial aspects of data sharing need to be addressed. Drawing on good practices worldwide, the Follow Up Group suggests that national governments should consider the following:

- Adopt and effectively implement the principle that data produced from publicly funded research should be openly available to the maximum extent possible;
- Encourage their research funding agencies and major data producing departments to work together to find ways to enhance access to statistical data, such as census materials and surveys;
- Adopt free access or marginal cost pricing policies for the dissemination of research-useful data produced by government departments and agencies;
- Analyze, assess, and monitor policies, programs, and management practices related to data access and sharing polices within their national research and research funding organizations.

The widespread national, international and cross-disciplinary sharing of research data is no longer a technological impossibility. Technology itself, however, will not fulfill the promise of e-science. Information and communication technologies provide the physical infrastructure. It is up to national governments, international agencies, research institutions, and scientists themselves to ensure that the institutional, financial and economic, legal, and cultural and behavioural aspects of data sharing are taken into account.

1. Preface

At its March 2001 meeting, the OECD Committee on Scientific and Technology Policy (CSTP) accepted a proposal from The Netherlands to establish a working group on issues of *access to research information*. The plans of the working group were presented at the October 2001 CSTP meeting. Subsequently, the Committee narrowed the scope of activities to *access to and sharing of research data* produced from public funding.[1] Participation in the group was broadened to include Australia, Canada, Denmark, Finland, Germany, Japan, Poland, the Netherlands, the United Kingdom, and the United States. The CSTP asked the working group to:

- Report on current practices concerning access to and sharing of research data and their underlying principles on the basis of case studies;
- Report on the effects of selected current data sharing practices on the quality of research and the progress of science;
- Suggest principles for making policy on data sharing within the relevant national and international policies and regulatory frameworks.

The report's core principle is that ***publicly funded research data should be openly available to the maximum extent possible***. Adoption of this principle will promote good stewardship of public knowledge, strong value chains of innovation, and maximize benefits from international cooperation (see Box 1). The report's findings and recommendations are addressed to: CSTP members as representatives from the governments of OECD member countries that carry responsibilities for national and international science policy and the functioning of research funding agencies; research institutes; and professional and scholarly associations. The objective is to contribute to a better understanding of the importance of research data access and sharing, and to offer suggestions on how the new digital challenges should be met.

Box 1: This core principle guides many public scientific institutions and scientists. However, it remains unevenly implemented. Most recently, it was adopted by the United Kingdom's Medical Research Council. After a workshop hosted by the European Science Foundation, the MRC drafted the following statement: MRC promotes the creation of a diverse range of datasets, many of which are rich in informational content, unique and cannot be readily replicated. Sharing allows scientists to extend the value of these datasets through new, high quality, ethical research and exploitation. It also reduces unnecessary duplication of data collection. Building preservation systematically into routine data management is part of good research practice: it strengthens quality, enables replication and audit, and provides a sound basis for data sharing.[2]

Building on a number of case studies and a great deal of other research, the report focuses on issues related to the access and sharing of publicly funded research data, in digital form, across all disciplines in the natural, health, and social sciences. Attention is paid to the international aspects of access and sharing relevant to scientific cooperation among OECD member states. Three significant topical areas fell outside the charge of this working group, however, and will require separate follow-up: issues particular to developing countries; issues related to data produced by a mixture of public and private funding; and the issue of national security restrictions in light of recent global events since 11 September 2001.[3]

2. Introduction

2.1. The changing information technology context for scientific research and innovation

Information and communication technologies (ICTs) are rapidly transforming research and the broader society: witness the growth in the number of Internet hosts per person, in the percentage of computers per household,[4] and in the continued rate of growth of chip, storage, and network technology capacity.[5] Concurrently, there has been an explosion in the amount of data produced across all types of scientific endeavour.[6] Continuing ICT advances, such as the development of grid computing, large-capacity optical transmission networks, wireless networks of sensors and devices, and complex imaging systems, promise to push these transformations farther and faster. ICT-dependent research, such as geographic information systems, data visualisation systems, and realistic modelling, are adding tremendously to our ability to study and understand the world in which we live. These developments provide researchers in OECD countries, and increasingly in developing countries, with the opportunity not only to be more efficient, more effective and better connected, but also to dramatically expand the scope and nature of their investigations.[7] Together they create the possibility of an "e-science infrastructure."[8] The growing activities in data collection, storage, processing, distribution, and preservation are, however, only loosely connected. They require systematic planning to realize the full potential of the emerging e-science infrastructure.

2.2. The benefits of data access and sharing in public research

Within this new technological context, more widespread and efficient access to and sharing of research data will have substantial benefits for public scientific research (see Box 2). Open access to, and sharing of, data reinforces open scientific inquiry, encourages diversity of analysis and opinion, promotes new research, makes possible the testing of new or alternative hypotheses and methods of analysis, supports studies on data collection methods and measurement, facilitates the education of new researchers, enables the exploration of topics not envisioned by the initial investigators, and permits the creation of new data sets when data from multiple sources are combined.

BOX 2: ACCESS to international data has helped produce a better understanding of public health issues and worldwide disease prevention and control. For instance, research on cholera outbreaks and their relationship to numerous environmental factors relied upon data drawn from epidemiology, NASA remote sensing, marine biology, microbiology, genomic data, and social science data. This research—an example of 'biocomplexity' studies supported by the U.S. National Science Foundation—would have been impossible without access to numerous databases. The effect of this interdisciplinary and international research project is an increased scientific and sociological understanding of cholera outbreaks and their prevention.[9]

Sharing and open access to publicly funded research data not only helps to maximize the research potential of new digital technologies and networks, but provides greater returns from the public investment in research.[10]

Improving and expanding the open availability of public research data will help generate wealth through the downstream commercialisation of outputs, provide decision-makers with the necessary facts to address complex, often trans-national problems, and offer individuals the opportunity to better understand the social and physical world in which we all live (see Box 3).

As a key link in the value chain of investments in research, open access to factual data plays an increasingly important role in all these areas.

BOX 3: A recent analysis demonstrated the economic benefits of providing open access to government meteorological data without any restrictions on re-use. [11] The "value adding" meteorological information industry in the United States has revenues in excess of $500M annually. The public meteorological data also support a rapidly growing weather risk management industry that underwrites financial risk management instruments valued at approximately $8B. In contrast, the private-sector value adding industry for meteorological information in the European Union is very small, largely attributable to the highly restrictive data policies of most national governmental meteorological services. What are harder to measure, but certainly occur, are the countless lost opportunity costs for researchers, students, and various other potential public users who find the high costs of the public data to be too great to use.

2.3. Roles and responsibilities of governments

If researchers throughout the world are to take full advantage of ICTs to improve and expand access to, and sharing of, research data, existing technological, institutional and managerial, financial and budgetary, legal and policy, and cultural and behavioural aspects must be addressed comprehensively and in an integrated way. To date, these aspects have often been treated on an *ad hoc*, project-specific basis. Given that OECD countries spend tens of billions of dollars each year collecting data that can be used for research, and for other social and economic benefits, ensuring that these data are easily accessible so that they can be used as often and as widely as possible, is a matter of sound stewardship of public resources (see Box 4).

BOX 4: Poor stewardship and lost opportunity for data access is exemplified by the case of Statistics Canada, which attempted to recover costs for its data management by charging data users. The effect of this form of management of these public data was a dramatic decrease in their use. In a study of the case, it was found that "Cost recovery was supposed to introduce a market type discipline on the demand for and supply of goods and services provided by the government. Since in economic terms Statistics Canada's outputs are public goods, the type of discipline envisioned by this policy is impossible to attain. Instead we have users who complain, refuse to pay and generally attempt to find alternative sources for their information needs. This policy fails the improved management of resources test."[12]

Scientists, research institutions, and research funding agencies around the world are increasingly engaging in large-scale, data-intensive projects. Such projects require data-management infrastructure, data-exchange protocols and policy frameworks, and a broad professional understanding that more extensive availability and use of the data is both necessary and desirable. Over the past decade, numerous studies, disciplines, research programs, and agencies have begun to address the complexities and benefits of open data access and sharing arrangements.[13] As scientists become better connected with each other, particularly through the Internet, and as research focuses on issues of global importance, such as climate change, human health and biodiversity, there is growing need to systematically address data access and sharing issues beyond national jurisdictions and thereby create greater value from international co-operation. The goal should be to ensure that both researchers and the broader public receive the optimum return on public investments, and to build on the value chain of investments in research and research data.[14]

3. Core Principle and Premises

The findings and recommendations that follow are based on the central principle that:

Publicly funded research data should be openly available to the maximum extent possible.

As a general principle, publicly funded research data should be as open as possible and available at the lowest possible access cost, subject only to legitimate restriction and considerations. Restrictions may be necessary for reasons of national security, for the protection of privacy of citizens, or the confidentiality of trade secrets. Access to research may be limited by the respect for private intellectual property rights. Finally, there may be reasons for granting temporary exclusive access to those who collected the data. But the guiding principle should be openness.

In order to derive the maximum benefit from public investments in research data, access, use, management and preservation must be an integral part of the research process. Conversely, data should not be considered an expendable by-product of research. In many cases, data have value beyond the project and anticipated use for which they were originally collected. The reuse of publicly funded data for research and other types of applications should be promoted and not restricted.

The accessing and sharing of data is not merely a technical matter, but also a complex social process in which researchers have to balance different pressures and interests. Purely regulatory approaches to data sharing are not likely to be successful without consideration of these factors. Various approaches to data access and sharing are therefore necessary, including the establishment of regulations and incentives, and the dissemination of best practices.[15]

The following three premises complement and support the core principle of this report:

3.1. Data from publicly funded research are a public good produced in the public interest

Both the data from publicly funded research and research itself have strong public good characteristics that support their open availability to the public, and especially to other researchers.[16]

3.2. Factual data are central to the scientific research process

The production, open dissemination, and unfettered use of factual data are essential attributes of, and inputs to, modern systems of scientific research and technological innovation. Recognizing the role of digital data as fundamental to the value chain of science, technology and innovation will enable an optimum return on public investments.

3.3. Data access and sharing issues are international in scope

To more fully exploit the possibilities of global digital networks, and to capture their benefits for the global community, policy issues concerning access to and sharing of publicly funded scientific research data must be addressed, not only at the institutional and national levels, but also at the international level.

4. Data Access Operating Principles and Attributes

Data access and sharing requires effective organization and management. The necessary components that make for this organization and management may be characterized as "data access regimes." In their ideal form, these regimes enable all participants in the scientific research process to freely and efficiently access and share data. Adequate data access regimes require dispersed, as well as centralised, responsibilities across different management domains that include the technological, institutional and managerial, financial and budgetary, legal and policy, and cultural and behavioural.

No single approach to developing an effective data access regime is possible; however, a list of operating principles for and attributes of effective data access regimes and resources can be offered. This list of attributes and operating principles is based on a broad set of experiences, and supported by the case studies conducted for this report. Key attributes are listed below, and illustrated with an example from the case studies. [17]

4.1. More explicit access regimes

There is a universal need for the formalisation of institutional rules and data management policies. This formalisation follows from the growing complexity and scale of scientific research and the increasing expenditure on research data. At the moment, it is not clear who is authorised to distribute data across the globe. To reach the necessary transparency in the tasks and responsibilities of those involved, terms of access and use of data that rest on tacit agreements will have to be made explicit and formalised. A systematic and institutionalized approach is needed to help address operating characteristics of data access, and to take advantage of the opportunities arising from publicly funded research.

4.2. Operating Principles

4.2.1. Openness

Open availability of publicly funded research data to the maximum extent possible is the core principle of this report.

4.2.2. Transparency of access and active dissemination.

Open data access requires actively disseminating where the data can be found, what the context and structure of the data collection is (metadata), how long the resource will be accessible, and what protocols and standards are employed. In short, this principle refers to the systematic visibility and traceability of data resources.

4.2.3. Assignment and assumption of formal responsibility

Formal responsibility for tasks associated with data access must be assumed by the appropriate participants in the global science system. The various individuals and institutions involved in the chain of data-related activities all have specific manifest and latent duties and obligations. These are founded in formal legal and professional normative standards and in the regulations of various agencies. Responsibility must also be assumed for various rights in the data supply, such as authorship, producer credits, ownership, financial arrangements, licensing terms, and, where appropriate, restrictions on use.

4.2.4. Professionalism
Codes of conduct, and related normative standards, of professional scientists and their communities can help to promote good practice and simplify the regulatory aspect of access regimes.

4.2.5. Interoperability
Technical and software standards and protocols are required to ensure the access and usability of data. These should be clear to the user and adopted by as many data management organizations as possible.

4.2.6. Quality
Quality refers to the proper description of uncertainties surrounding the production of the data (e.g., the techniques employed in their collection and archiving, and the measuring instruments and their calibration), the ability to ensure that the cited source and value are ***authentic***, that the data retain ***integrity*** (complete and absent from introduced errors), and that they are ***secure*** against loss, destruction, modification, and unauthorized access.

4.2.7 Operational Efficiency
Open access to data increases the efficiency of research by avoiding unnecessary duplication of data collection and permitting the creation of new data sets by combining data from multiple sources. Coupled with open access, comprehensive documentation of data sets and how to access them provides a more efficient use of resources.

4.2.8. Flexibility
In general, scientific communities will approach data management requirements more consistently within their discipline internationally, than they will across other disciplines on a national level. Data access regimes need to be sufficiently flexible to take account of this variation.

4.2.9. Property
Institutional intellectual property rights as well as the individual rights of researchers are considerations of property interests. Unlike the private sector, public research operates on a principle of collective property interests, which are promoted by the open access and sharing of data resources.

4.2.10. Legality
Legal restrictions may limit access to and use of data.[18] Restrictions will apply primarily to 'secondary' data sets compiled for purposes other than scientific research. In some cases, the sensitive parts of data sets can be left out without rendering them useless. Specific types of legal restrictions include: national security, privacy and the protection of trade secrets.

4.2.11. Accountability
Accountability involves measuring the cost, benefit, and performance of data access and sharing regimes and taking appropriate actions in response to the results.

4.3 Building a Data Access Regime: the Global Biodiversity Information Facility (GBIF)

The Global Biodiversity Information Facility (GBIF), which began under the auspices of the OECD Megascience Forum, has sought to implement these principles as a means to achieve the larger goal of providing worldwide access to biodiversity data. GBIF's goal is to make "the world's scientific biodiversity data freely available to all **[openness]**."[19] The fundamental motivation for GBIF is to enable access to a vast amount of biodiversity data housed in databases distributed in numerous countries and institutions. By bringing all these data into one interoperable network, and producing a registry of biodiversity information resources, GBIF will produce systematic visibility and traceability of data resources **[transparency]**.

Formal responsibilities of different participants involved in the task of building GBIF's organisation and legal relationships have been put forth in GBIF's Memorandum of Understanding. GBIF's Secretariat is responsible for carrying out work programmes that are approved by the Governing Board, which consists of representatives of GBIF's Participants. This structure enables GBIF to have a legal identity as an international body, manage financial contributions and work programmes, while drawing upon efforts and resources from Participants. In his reflection on the establishment of GBIF submitted to the OECD, Eric James attests: "The way in which these legal requirements are met may be the most important factor determining the structure of the organisation that is created."[20] The establishment of GBIF's activities occurred in and through contact with existing scientific and political bodies to maintain and establish professional codes, gain consensus about scientific outcomes, and negotiate with government representatives about GBIF's larger social and economic roles **[professionalism]**. The review will evaluate GBIF's progress toward data availability and interoperability, its responsiveness to user needs, and the professionalism of the Secretariat.

Participants will provide stable gateways, or "nodes," to databases that contain primary or meta-level biodiversity data. These nodes must provide documentation and metadata about the data in the databases, vouch for data **quality**, ensure data authenticity and security. GBIF will help develop standards for database **interoperability** through one of its 4 work programmes, Data Access and Database Interoperability (DADI). GBIF aims to develop an interoperable network of distributed databases by coordinating and leveraging existing national and international programs and projects, which allows for **operational efficiency** and more cost-effective basis for making biodiversity data freely and easily available to a heterogeneous user community.

The databases and the data accessed through GBIF are in most cases owned and developed by other organisations and thus will not entail any assertion of IPRs by GBIF itself **[property]**. GBIF aims to provide best practices on how to deal with IPRs, particularly since it will be drawing from databases hosted by different institutions and countries with different legal frameworks, with a view to promoting open access and sharing to the maximum extent possible.[21] GBIF also asserts in its MOU that biodiversity data will be properly used and acknowledged by its participants **[legality].** Further, its efforts do not conflict with the Clearing House Mechanism, and they abide by the Global Taxonomic Initiative of the Convention on Biological Diversity concerning the proper and equitable use of biodiversity data and the resources to which they refer.

During the establishment of GBIF, the OECD provided the forum to assess the level of support for this new scientific collaboration, to bring together related proposals and to develop detailed plans that could then be taken up by interested countries. GBIF will have a third-year review of the effectiveness of its MOU, its scientific efforts and the "transparency of its dealings with politically sensitive issues"[22] **[accountability]**.

5. Data Access Management: Five Domains

Efficient data access can only take place with the proper administration and organization of different management domains within data access regimes. These domains include technological, institutional and managerial, financial and budgetary, legal and policy, and cultural and behavioural considerations (see Figure 1). These domains provide a framework for locating and analyzing where improvements to data access and sharing can be made.

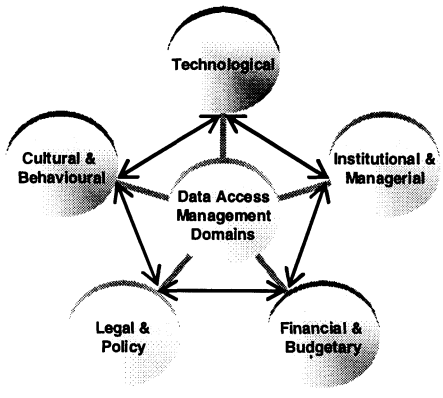

Figure 1. Components of a Data Access Regime

The five domains differ in character across the traditions and practices of specific scientific disciplines, e.g., astrophysics, biology. Thus, data access regimes may vary in significant ways. There is no single model for how data access should take place. The implementation of the core principle of open availability, however, requires a systematic approach that recognizes the necessity of implementing improvements across the interdependent management domains. This approach also requires the involvement of actors from various levels: governments, funding agencies, and research institutions and professional and scholarly societies, as well as individual scientists themselves.

5.1 ***Technological domain****: Broad access to research data, and their optimum exploitation, requires appropriately designed technological infrastructure, broad international agreement on interoperability, and effective data quality controls.*

A technical infrastructure that supports user needs is necessary to derive maximum benefits from data access and sharing. This infrastructure must be robust enough for long term use and, when appropriate, for diverse uses. It also must be flexible enough to respond to the continuous and rapid changes in scientific research and technology. While there are many technical issues to be resolved to take full advantage of past, current and future investments in ICT infrastructure, the main barriers to effective data access and sharing are no longer technical, but are institutional and managerial, financial and budgetary, legal and policy, and cultural and behavioural.

Technical operating principles for data access regimes include ***interoperability*** (of protocol and software to ensure the access and usability and multiple use of the data); and ***quality*** (including technical components of ***authenticity***, ***integrity***, and ***security***) of data.

Data Preparation and Metadata: ICPSR

In 1995, the Inter-University Consortium for Political and Social Research (ICPSR) initiated the development of the Data Documentation Initiative (DDI), an international criterion and methodology for the content, presentation, transport, and preservation of metadata about datasets in the social and behavioural sciences. DDI, which is in XML format, helps enhance users' ability to acquire and use data while it assists producers' in packaging and disseminating them. After a period of beta-testing with participating international organisations, DDI is now in use by a number of organisations, including Networked Social Science Tools and Resources (NESSTAR), Health Canada, and ICPSR. ICPSR continues to assist data producers in preparing their data through its "Guide to Social Science Data Preparation and Archiving," a guide with broad appeal for individuals and organisations searching for easy and effective ways to technically manage and prepare data so that they can be easily and effectively placed into network environments.[23]

5.2 ***Institutional and managerial domain:*** *While the core open access principle applies to all science communities, the diversity of the scientific enterprise suggests that a variety of institutional models and tailored data management approaches are most effective in meeting the needs of researchers.*

Because scientific data have many different characteristics and uses, there is no monolithic institutional and management approach that can be applied universally.[24] Key characteristics of data production and use include whether the data are (1) government-generated or generated at a research institution using public funds; (2) useful only within the discipline or across many disciplines; (3) useful over the very long term or only within short-term horizons; (4) have public-policy implications; or, (5) have significant broader economic and social value, among other factors.

Institutional and managerial operating principles for data access regimes include ***transparency*** (systematic visibility of the data source); ***responsibility*** (explicit formal institutional rules on data management); and ***accountability*** (rendering public account for the performance of data access regimes).

Negotiated collaborations: CERN

The European Organisation for Nuclear Research, CERN, is one of the world's largest scientific laboratories, presently financed by twenty European countries. CERN overtly subscribes to the core principle and premises outlined in this report, but leaves it to individual 'collaborations' of scientists to devise experiment-specific regulations to ensure compliance. Negotiations between different collaborations are necessary to enable data sharing, including agreement on definitions and standards. The type of data produced and the method of processing used will play a large part in deciding upon the most effective management model to adopt. This flexibility of management approach is a key factor in the data sharing environment at CERN.

5.3 ***Financial and budgetary domain:*** *Scientific data infrastructure requires continued, and dedicated, budgetary planning and appropriate financial support. The use of research data cannot be maximized if access, management and preservation costs are an add-on or after-thought in research projects.*

In many areas of public research, there are indications of discrepancies between the funding of the specific research itself and the related data-management requirements (which do not necessarily benefit the individual scientist, but which are necessary for data reuse). Generally, research organizations fund the former well, but pay scant attention to the latter. In the digital environment, scientific data sets must be viewed as a key element of the broader research infrastructure and as an investment in the future capacity to innovate and solve pressing problems. Adequate support is essential for data-management functions, such as the development of sufficient explanatory documentation for each data set (i.e., metadata), conversion of old formats onto new media, adaptation to new standards, and long-term preservation, archiving, and maintenance.

Budgetary operating principles for data access regimes include ***operational efficiency*** (maximizing the return on investment by promoting re-use of data, and providing proper documentation, specialists, and effective data management facilities).

Funding schemes "on a rolling basis:" the European Bioinformatics Institute (EBI)

The official mission of the EBI is to ensure that the growing body of information from molecular biology and genomic research is placed in the public domain and is freely accessible to the scientific community in ways that promote scientific progress. Like other scientific bodies, the EBI has a major problem in the funding for its building, maintaining and making available databases and information services even though they represent only a small fraction of the total research costs. The key issue is that funding for data sharing infrastructures needs to be constructed "on a rolling" or on-going basis to maintain effective data management. These funding requirements are very different from the funding schedules of research, which are usually project oriented. These differences in budgeting constitute the main threat to the EBI's commitment to maintaining the public availability of its data.

5.4 ***Legal and policy domain****: National laws and international agreements directly affect data access and sharing practices, despite the fact that they are often adopted without due consideration of the impact on the sharing of publicly funded research data.*

Intellectual property laws, information policies, institutional guidelines, and contracts at the national and international levels often impose terms and conditions on data access and sharing practices. Laws and policies governing data access and sharing practices may vary among different countries, resulting in barriers to scientific cooperation and progress. Based on a recent Web survey, most of the national research organization managers who responded expected that data sharing will become a major policy issue in the next five years. This situation requires greater attention by the science policy community at all levels. In particular, restrictions on re-use of public data by the research community must be eliminated or minimised as much as possible. Research grant provisions and licensing templates for promoting open access and unrestricted re-use of public research data already exist, but have not yet been broadly adopted.

Legal and policy operating principles for data access regimes include ***property*** (balance intellectual property rights of investigator and institution versus public good); and ***legality*** (lawful data management, respecting national security, privacy and trade secrets).

Policy interconnections: functional MRI and the Institutional Review Boards

The functional Magnetic Resonance Imaging Data Center's (fMRIDC) principal endeavour is to promote data sharing in brain mapping. The Western tradition of informed consent in bio-medicine operates according to the principle that the 'most specific consent is the best consent.' When data are to be gathered for submission to databases, the specificity of consent may run counter to the goals of meta-analysis or re-analysis by third parties, to investigate issues different from those for which the data was originally gathered. The creation of infrastructures for data sharing, therefore, has to conform to the rules of regulatory bodies, such as institutional review boards (IRBs), whose approval must be obtained to share data. As such, these bodies function as gatekeepers to the circulation of data. International coordination may also be necessary. Researchers submitting or requesting data across national boundaries may find it especially difficult to act in accordance with the various ethical guidelines that exist in different countries. The fMRIDC has been hesitant to accept data from non-US settings because of concerns regarding IRB compliance.

5.5 ***Cultural and behavioural domain:*** *Appropriate reward structures are a necessary component for promoting data access and sharing practices. These apply to both those who produce and those who manage research data.*

Although formal policy frameworks and regulations are necessary to make research data publicly available, they need to be supplemented by appropriate community-based norms and incentives for researchers to share and provide access to their data and for appropriate recognition of their data-related work. In many cases, there is a general lack of reward structures and mechanisms to promote open access to, and sharing of, data from public research.

Cultural and behavioural operating principles for data access regimes include ***quality*** (trust that data are what they purport to be); ***professionalism*** (build on codes of conduct and ethics of the scientific community); ***flexibility*** (there is no single model on how data access must be provided.)

Incentives: the Protein Data Bank

To publish in scientific journals, U.S. scientists involved in the field of crystallography must deposit their data in the Protein Data Bank (PDB) and acquire an accession number. "By requiring everyone to submit data, the community is assured of having the most up to date information possible. Now, increasingly, under our regime, a lot of [data] depositors have come to realize that the practice that we use has some advantages for them in that we check things and we find errors and inconsistencies. That actually improves the quality of the product they produce."[25]

6. Possible Action Areas

Our findings from the case studies and from other research indicate a number of action areas by the different actors involved in making possible open access to, and sharing of, publicly funded research data. In this section we recommend possible action areas for the OECD and national governments.

OECD

As an international organization with credibility and stature in the science policy arena, the OECD, through the CSTP, can play a crucial role in promoting access to, and sharing of, data from publicly funded research. Central to this role is the gathering and sharing of information on data related activities and policies. At the international level, only a small handful of organizations have undertaken to do this, usually in the context of a specific discipline or research program. The recent, and vast, expansion of research data assets and the trend towards issue-based, interdisciplinary research, however, suggests that all countries and all fields of science stand to benefit from greater attention and an organized and coordinated approach to effective policy actions

1. **The OECD should put the issues of data access and sharing on the agenda of the next Ministerial meeting.** ICT advances have created the ability to transform science. New tools allow researchers to find data in seconds that would have taken months just a few years ago. Effective data access and sharing requires a comprehensive policy approach for implementation by public research institutions. Monitoring progress and

devoting attention to the public research data issues and activities would assist decision-makers and research support agencies in developing appropriate policies and allocating resources.

Areas in Conjunction with Relevant Member Country Research Organizations

2. **The OECD should consider conducting or coordinating a study to survey national laws and policies that affect data access and sharing practices.** This relatively simple undertaking could determine what policies exist, how accessible they are, and result in listing of the web sites where these policies are posted. This study would be of considerable benefit to science policy-makers, research administrators, and information resource managers in all countries, both within OECD and beyond. The study could look at the feasibility of developing a central and easily accessible repository of national laws and policies that affect data access and sharing practices. Such a compilation does not currently exist, and could be useful to facilitate international research collaborations.
3. **The OECD should consider conducting or coordinating a study to compile model licensing agreements and templates for access to and sharing of publicly funded data.** Depending on the context, numerous factors need to be considered in data access and sharing arrangements. Nevertheless, many contractual models already exist that have been developed by research funding organisations, research program managers, university administrators, librarians, and others. The OECD, as a global organization, is ideally suited to span national domains where examples do exist, and thereby bring an international perspective. The study could compile and review existing agreements and models to find exemplary approaches. Having readily available models on hand would be of considerable benefit to researchers, universities, and research institutions, as well as data centers and archives, and could facilitate international research collaboration

Areas for Further Examination

4. **Governments within the OECD should expand their policy framework of research data access and sharing to include data produced from a mixture of public and private funds.** Collaborative public/private research projects, and the resulting data, have their own unique set of characteristics and issues. As more national governments promote public-private partnerships in research, these issues will be of increasing importance to both public researchers and the companies that are involved. A further examination of the state of data sharing and access in these types of research arrangements needs to be made to develop sound science policy guidance.
5. **The OECD should consider examinations of research data access and sharing to include issues of interacting with developing countries.** The increase of participation in the research enterprise benefits the global science system and innovation. Providing developing countries with access to data from publicly funded research increases their participation in science. Further, as United Nations Education, Scientific and Cultural Organization (UNESCO), the International Council of Scientific Unions (ICSU), private foundations, and other organizations have emphasized, access to scientific knowledge by developing countries is vital to the progress of the entire world. This access is particularly important in the context of global issues such as population health, environmental change, and food production. Of course, open access to data from publicly funded research in developed countries can provide a valuable resource for economic

development, education, and scientific capacity building. Many efforts are already underway to improve access for researchers in developing countries (e.g. providing free or below costs access to data and scientific information) as well as establishing optimal data regimes for developing countries to share their data (e.g. addressing issues of data repatriation). A systematic examination of barriers and best practices would provide both a picture of the current situation and a set of guidelines for further action.

6. **The OECD should promote further research, including a comprehensive economic analysis of existing data access regimes, at both the national and research projectd or program levels.** To date, no one has yet undertaken a comprehensive, economic analysis of different data access regimes. Several key issues have not been closely examined, including the relative costs of providing data openly, the impact of cost recovery on the use of those data, and the positive externalities and network effects from providing open access to publicly funded research data. The OECD should consider conducting this type of analysis or encouraging member country research organizations to fund such studies.

National Governments

Although the OECD, UNESCO, ICSU, and other international bodies can play a role in improving the current situation regarding research data access and sharing, it is at the national level that many important decisions and actions must be taken. National governments provide the resources for making data accessible, establish the overall policies for data management, regulate matters such as confidentiality and privacy, and determine restrictions based on national security. Most importantly, it is national governments that are responsible for the major research support and funding organizations, and it is here that many of the managerial aspects of data sharing need to be addressed.

The national governments of OECD countries should consider:

1. **Adopting, and effectively implementing, the principle that data produced from publicly funded research should be openly available to the maximum extent possible.** The public investments made in research data collection can only be maximized if the data are preserved, managed, and made accessible. This requires coordinated attention by governments at all levels, and adequate policy and financial support. The starting point for these actions, however, is the affirmation that data collected using public funds should be openly accessible to all.
2. **Encouraging their research funding agencies and major data producing departments to work together to find ways to enhance access to statistical data, such as census materials and surveys.** Many countries have taken steps to facilitate access to census and survey materials by developing catalogues, user-friendly repositories, off-site research facilities, training programs, and regulatory frameworks for providing appropriately guarded access to confidential information. Such steps have proven enormously effective in maximizing the use of national surveys and producing insights into the functions of economies and societies.
3. **Adopting free access, or marginal cost pricing, policies for the dissemination of research-useful data produced by government departments and agencies.** The use of information collected through public funding should be freely accessible for research

purposes. This maximizes the use of such information for public policy and public knowledge development.

4. **Analyzing, assessing and monitoring policies, programs, and management practices related to data access and sharing policies within their national research and research funding organizations.** This information would be useful to national governments so that they may assess the implementation of the previous three considerations. The resources, support programs, policies, and regulations related to research data sharing are, in large part, developed and implemented by research funding organizations. The operations of these organizations play a crucial role in determining the degree to which data are made accessible and shared between researchers. Many organizations, such as NSF and NIH in the United States, Social Sciences and Humanities Research Council in Canada, and the European Science Foundation are now developing, or have developed, policies, regulations and support programs that promote data sharing. Issues such as establishing protocols for the collection and release of confidential information, developing technical infrastructure, agreeing on metadata standards, requiring data preservation strategies within individual research projects, and including data management costs as eligible expenditures in grant applications have been dealt with by one or more of these agencies. It would benefit the global scientific community if decision-makers within national governments had a clear understanding of where their respective agencies stood in relation to those in other countries.

7. Conclusion

Improving access to and sharing of publicly funded research data is an issue that touches on all aspects of the research enterprise and the development of knowledge, and involves all participants in the conduct of research. For the individual researcher, the sharing of data, particularly prior to publication[26], can be burdensome, time consuming, and unrewarding if the necessary measures are not taken to provide funding, facilities, and a social context that emphasises its value to the research community and to society.

Advances in ICTs, the internationalisation of science, and the trend toward issue-based research hold great potential for the advancement of knowledge and for the benefit of all people. This potential will not be fully realized unless all of the major elements of data access regimes identified in this report are properly developed. To do so will take considerable discussion, understanding, and commitment on the part of all those involved in research, particularly at the policy level.

Agreement among OECD governments on a set of general principles to shape specific data access regimes, as well as adoption of the recommendations set forth above, would be enabling for scientists, empowering for citizens, and provide an important contribution to fulfill the promises of e-science.

---

[1] In this report, we define "access to data" as the act of making the data available for use by others; by "sharing" we mean a researcher allowing one or more other individuals to use data, typically with the implicit, if not explicit assumption that it is on a reciprocal basis. The sharing of data involves providing specific access, whereas the act of providing access by itself does not necessarily involve any sharing arrangement. Data sharing focuses on data exchanges between individual researchers rather than institutions, while access may be provided at any level. Sharing also reflects the cooperative norms of public science as practiced within many disciplines by many

researchers in OECD countries. We define data as in the U.S. National Institutes of Health definition of final research data: "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings".

[2] See http://www.mrc.ac.uk/index/strategy/strategy-science_strategy/strategy-strategic_implementation/strategy-data_sharing/strategy-data_sharing_policy-link

[3] CODATA, the interdisciplinary Committee on Data for Science and Technology of ICSU, is currently examining barriers to data access and sharing that are particular to developing countries. CODATA, however, does not normally examine issues related to social science and humanities research. Related to issues of national security, see "NAS Censors Report on Agricultural Threats," *Science* 20, p. 1973-1975, on the several scenarios that were left out of a public report of the U.S. National Academy of Sciences.

[4] See NSF 2002 Science and Engineering Indicators, http://www.nsf.gov/sbe/srs/seind02/start.htm

[5] Gary Stix (2001), "Triumph of Light," *Scientific American*, January, available at http://www.sciam.com/2001/0101issue/0101stix.html

[6] Examples range from genetic sequence and protein structure data in bioinformatics, to various types of brain imagery in neuroscience, to sky surveys and virtual observatories in astronomy, and geospatial data such as Global Spatial Data Infrastructure.

[7] Examples include combining data from multiple data sources to gain a greater statistical power to resolve hypotheses (see the Biomedical Informatics Research Network, http://www.nbirn.net); and obtaining real-time global measurement on environmental observations.

[8] John Taylor, Director General of (UK) Research Councils (UK), www.research-councils.ac.uk/escience/. "E-Science will refer to the large scale science that will increasingly be carried out through distributed global collaborations enabled by the Internet. Typically, a feature of such collaborative scientific enterprises is that they will require access to very large data collections, very large scale computing resources and high performance visualisation back to the individual user scientist. . . . Besides information stored in Web pages, scientists will need easy access to remote facilities, to computer – either as dedicated Teraflop computers or cheap collections of PCs – and to information stored in dedicated databases. The Grid is architecture to bring all these issues together." See also Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue Ribbon Report on Cyberinfrastructure, http://www.cise.nsf.gov/evnt/reports/atkins_annc_020303.htm .

[9] Rita Colwell (2002), "A Global Thirst for Safe Water: The Case of Cholera," Abel Wolman Lecture at the National Academy of Sciences, January 25, 2002, available at http://www7.nationalacademies.org/wstb/2002_Wolman_Lecture.pdf. Other examples of the impact of access to and sharing of international data in the control and elimination of worldwide diseases include the World Health Organisation's network of collaboration centres. In the worldwide programme of epidemiological surveillance of influenza, these receive epidemiological information on outbreaks of influenza from national institutions throughout the world. They also receive new strains of the virus for characterization and give advice as to their possible use in vaccine preparation. The centres then distribute the necessary reagents, antigens and anti-sera to national laboratories, and high-yielding recombinant viruses for to vaccine producers. See http://whqlily.who.int/general_infos.asp.

[10] For more benefits of data sharing, see National Academy Press (1985), *Sharing Research Data*, available at http://books.nap.edu/catalog/2033.html

[11] Peter Weiss (forthcoming 2003) presentation in Proceedings of the Symposium on the Role of Scientific and Technical Data in the Public Domain, National Academies Press. See also, European Union Green Paper (1998), "Public Sector Information: A Key Resource for Europe," COM 585, and PIRA International, "Commercial Exploitation of Europe's Public Sector Information, Final Report for the European Commission (2000),"Directorate General for the Information Society," which provide similar comparisons of such policies in other information sectors.

[12] Ronald C. McMahon (1996), "Cost Recovery and Statistics Canada," in Government Information in Canada, Volume 2, number 4 (spring 1996), retrieved from http://www.usask.ca/library/gic/v2n4/mcmahon/mcmahon.html, February 2003

[13] Studies include: National Research Council (1997), *Bits of Power: Issues in Global Access to Scientific Data*, National Academy Press, Washington, D.C.; National Research Council (1999); and *A Question of Balance: Private Rights and The Public Interest in Scientific and Technical Databases*, National Academy Press, Washington, D.C.; Stephen Hilgartner (1996), "Access to Data and Intellectual Property: Scientific Exchange in Genome Research" in *Intellectual Property Rights and the Dissemination of Research Tools in Molecular Biology: Summary of a Workshop held at the National Academy of Science, February 15-16, 1996;* and National Research Council (1995),

On the Full and Open Exchange of Scientific Data, National Academy Press, Washington, D.C.; and National Research Council (2002), *Community Standards for Sharing Publication-Related Data and Materials,* National Academy Press, Washington, D.C. The European Bioinformatics Institute, the Global Change Program, the Global Biodiversity Information Facility, the European Social Survey, the International Union of Crystallography, the international Ocean Drilling Program; The European Organization for Nuclear Research, otherwise known as CERN, provide good examples of research programmes with effective data policies. Funding agency statements include: NSF at http://www.nsf.gov/sbe/ses/common/archive.htm) and "NIH Draft Statement on Sharing Research Data" at http://grants2.nih.gov/grants/policy/data_sharing/. Sites that discuss how to develop a data policy include Smithsonian Environmental Research Center at http://www.serc.si.edu/datamgmnt/policy1.htmand the Ecological Sciences Network at www.esnet.edu. The policy of the Long Term Ecological Research (LTER) network is at http://www.lternet.edu/data/netpolicy.html.

[14] For more on the Role of Governments in the Digital Age, see Stiglitz, Orzag and Orzag at http://www.ccianet.org/govt_comp.php3. In particular note the following three:
Principle 1: Providing public data and information is a proper governmental role
Principle 2: Improving the efficiency with which governmental services are provided is a proper governmental role
Principle 3: The support of basic research is a proper governmental role

[15] As one researcher put it, "Incentives for data sharing need to be offered that offset the investigators' loss of control over their databases. Usually, this is some form of added scientific value. By sharing data, an investigator may gain access to more data or other tools. Ultimately, there has to be a procedural framework that makes sharing sensible, efficient, and value-added. If all those pieces are in place, fewer external or coercive forces are needed to convince researchers to share." From minutes from an NIMH meeting, see Paul Wouters, Data Sharing Policies, 10 June 2002. Networked Research and Digital Information, NIWI-KNAW on http://dataaccess.ucsd.edu

[16] In economics, a good is considered a "public good" if it is "non-rivalrous" and "non-excludable." The former means that the marginal costs of providing the good to an additional person are zero. The latter means that once the good is produced, the producer cannot exclude others from benefiting from it. See, Inge Kaul, Isabelle Grunberg, and Marc Stern (1999), "Defining Global Public Goods," in *Global Public Goods: International Cooperation in the 21st Century*, eds. Both publicly funded basic research and the data produced from it and disseminated on digital networks are non-rivalrous. They are not purely excludable, however, although their excludability, especially for other researchers, is neither economically efficient nor desirable as a matter of public policy, absent countervailing and superseding reasons to the contrary.

[17]These operating principles evolved from the document produced by Hans Franken, Access to Publicly Financed Research, Conference Conclusion. Global Research Village III Amsterdam 2000. For other principles on data access and sharing see http://www.codata.org/data_access/principles.html . Examples of successful guidelines based on a systematic set of principles are the OECD Guidelines on the Protection of Privacy and Trans-border Flows of Personal Data (1980) and the Principles and Guidelines for the Sharing of Biomedical Research Resources (1999) from the US National Institutes of Health (NIH) and the OECD Guidelines for Security of Information Systems and Networks (2002).

[18] Examples include *National security*: Data sets from some oceanographic or geological surveys may be (partly) classified and not accessible; *Privacy*: Data from human subjects are vulnerable to breaches of confidentiality and privacy and therefore should only be obtained by fair and lawful means, with knowledge or consent of the data subjects; and *Trade secrets*: Data potentially relevant to prospective patenting or commercial opportunities may contain (partly) confidential information.

[19] See www.gbif.net

[20] Eric James, "Establishing International Scientific Collaborations: Lessons Learned from the Global Biodiversity Information Facility," submitted to Sixth Meeting of the OECD Global Science Forum, available at http://www.oecd.org/pdf/M00027000/M00027203.pdf

[21] Research on issues of IPR, particularly for natural history museums, is being conducted by European Natural History Specimen Information Network, see. http://www.nhm.ac.uk/science/rco/enhsin/details.html and "Beset with pitfalls-specimens and databases, intellectual property and copyright," Simon J. Owens and Alyson Prior, from the 2000 meeting of the Taxonomic Databases Working Group, November, 2000; Senckenberg Museum, Frankfurt, available at http://www.tdwg.org/tdwg2000/ipr.htm.

[22] Eric James, "Establishing International Scientific Collaborations: Lessons Learned from the Global Biodiversity Information Facility," submitted to Sixth Meeting of the OECD Global Science Forum, Section 10.

[23] For more information on ICPSR and DDI, http://www.icpsr.umich.edu/DDI/ORG/index.html. For ICPSR's Guide, see www.icpsr.umich.edu/ACCESS/dpm.html. For information on the importance and development of DDI,

see "Providing Global Access to Distributed Data through Metadata Standardisation -- The Parallel Stories of NESSATAR and DDI", submitted by the Norwegian Social Science Data Services to the Conference of European Statisticians, UN/ECE Work Session on Statistical Metadata, Geneva, Switzerland, 22-24 September 1999, at http://www.nesstar.org/papers/GlobalAccess.html.

[24] For example, mathematics presents a special feature in that published material never becomes obsolete, so that the data needed by the working mathematician is ideally the full collection of published papers, past and present. With the development of internet access, this is not an impossible objective. New papers are almost always produced in electronic form, and therefore could be stored and accessed. The amount of past literature to be scanned and digitized is estimated to be around 50 million pages. Under the umbrella of the International Mathematical Union, an attempt is made to coordinate national efforts to insure permanent accessibility at a reasonable cost for the users to both new and digitized papers. Without this, research will be limited to rich parts of the world, where libraries can be heavily funded. See http://www.mathematik.uni-bielefeld.de/~rehmann/DML/ and http://www.library.cornell.edu/dmlib/

[25] Berman, Helen. Director, Protein Data Bank. Personal communication.

[26] Rapid Data Release Policy: "Ever since the 1996 Bermuda Principles provided guidelines on the rapid release of data from large-scale sequencing projects, access to the pre-publication sequence data that has been made freely available in public nucleotide sequence databases has accelerated biomedical research. However, in 2002, it became clear that new strategies and other advances in large-scale DNA sequencing necessitated a re-examination and updating of the data release policies originally developed to implement the Bermuda Principles for pre-publication sequence data. At its February 10-11, 2003 meeting, the National Adivsory Council for Human Genome Research (NACHGR), the main advisory group to the National Human Genome Research Institute (NHGRI) on genetics and genomic research, discussed the subject of pre-publication release of large-scale sequencing data. NACHGR approved a draft policy that would reaffirm and extend the rapid data release policies developed to implement the 1996 Bermuda Principles, and recommended that NHGRI publicize the draft policy statement for the purpose of obtaining comment from the scientific community." (http://www.genome.gov/page.cfm?pageID=10506376). For a reaffirmation and exteions of the NHGRI rapid data release policy, see http://www.genome.gov/page.cfm?pageID=10506537. For community discussion see Sacrifice for the greater good? Nature 421, 875 (2003), and Draft guidels ease restrictions on use of genome sequence data, Nature 421, 877-878 (2003).

ANNEX 1.
THE FOUR CASE STUDIES: CERN, EBI, fMRIDC, GBIF/Biodiversity

Preface

One product of the OECD Follow Up Group on Issues of Access to Publicly Funded Research Data is a set of four case studies. These studies have been used to guide the Group's discussion and inform the Group of various practices. Provided in this Annex is an overview of the case studies, short summaries of each one, and a set of research questions motivated by the case studies. A longer version of these case studies has been published as *Promise and Practice in Data Sharing*[27] by the Netherlands Institute of Scientific Information Services and is available through its Website.

Introduction

The formulation of data sharing policy principles has been central to this report; such principles are central to the development of new data access guidelines. Policies are only useful, however, to the extent that they inform and modify practice. Thus, in developing our guidelines, we drew on four case studies of data sharing practices both to gain ideas of best practices and to learn what difficulties there might be along the road to implementation. This annex presents in more detail the most important findings of four case studies that have been conducted in the framework of writing the report. This selection is valuable in that the cases come from a variety of disciplinary and national/international settings, and address emerging (fMRI, GBIF) and long-standing practices of data-sharing (CERN, EBI). While the particular cases each have their own unique features, the elements that make a difference to data-sharing also become visible when these cases are read in relation to each other. The Annex concludes with a discussion of key areas for further research in the development of effective and efficient data sharing policies.

The cases are an indication of the type of benefits that can be expected from increased data sharing among researchers, and between researchers and society at large. They provide a vivid picture of the challenges – sometimes quite formidable - that must be overcome to increase access to, and sharing of, research data. For example, in each case we uncovered the entwinement of data sharing practices with a range of other scientific activities (publication mechanisms, peer review process and so forth). At the same time, the four case studies point to certain lacunae in our knowledge. There is still a lot about the actual behaviour of researchers that we do not know in enough detail to give firm empirical support to research policies. This is addressed in this annex by formulating points for follow-up research.

It is clear that interaction between data production (frequently, but not exclusively, a research activity) and data management are key in characterising the kind of sharing that takes place. EBI draws a clear distinction between data production and data sharing. These are, therefore, activities that have separate settings and which are accepted by researchers. In the case of CERN, the management of data is part of the production of data. It is seen as a specific activity, one which is subsumed under data production. This explains why there is no sharing, and no sense that sharing is necessary. The case of fMRIDC is somewhere between these two. Traditionally, data has been closely tied to its context of production in this field. The encouragement to share data often entails separating data from their experimental context of

production. This means, however, that new practices of data management must arise, and that researchers must agree on drawing a line between the context of production, in which they maintain control of data, and the circulation of data, in a public setting. For biodiversity databases, the development of data management practices (including standards, annotations and formats) is also in progress.

In each case, materials were gathered via Internet from the websites of the data-infrastructures and their host institutions, from the published literature and through interviews. A common framework was used as a basis for the interviews. For the EMBL, CERN and GBIF, site visits were also conducted.

The authors wish to thank the interviewees and respondents for their involvement in this research.

Case study of Data Policies at CERN

Dr Paul Wouters and Colin Reddy
Networked Research and Digital Information (Nerdi)
NIWI-KNAW
PO Box 95110
1090 HC Amsterdam
The Netherlands
email: paul.wouters@niwi.knaw.nl
www: www.niwi.knaw.nl/nerdi

Introduction

This case study explores the data policies in one of the key institutions in physics. The European Organization for Nuclear Research, CERN, is one of the world's largest scientific laboratories. It was founded in 1954 and is located at the Swiss-French border near Geneva. CERN is presently financed by twenty European countries and has developed collaboration with laboratories in the US. CERN is the paradigmatic example of "big science." More than 7000 scientists, from laboratories and universities all over the globe, work there to study the constituents of matter and the nature of fundamental forces. CERN's official mission is "to create new knowledge on subjects ranging from anti-hydrogen to neutrinos, to the proton's inner structure, to the generation of mass and dark matter."

Benefits

CERN's commitment to making research data publicly available is laid down in its founding convention. Experiments at CERN are run by teams in the form of 'collaborations'. These collaborations contain a number of scientists from different countries. Some collaborations, such as the one on the Large Hadron Collider, have more than one thousand members. Data sharing policy within a particular collaboration is unproblematic, since members of the collaboration are allowed access to both the data and the programs developed to process them. Access is usually controlled by the use of encryption and the use of passwords. The raw experimental data produced may be of little use to anyone outside the collaborations in an unprocessed form. These data are not publicly available.

At the highest level of data extraction are databases with the refined data and the Particle Review Book produced by the particle data research group at the University of California at Berkeley. This is publicly available. The Particle Review Book is published every other year, both in print and electronically. The Particle Review Book is becoming more and more a text book in this field.

Relations to other activities

CERN is a research institute that produces and processes data on a massive scale. This already is such a complex endeavour that running the experiments has necessitated a large organisation in which procedures have been put down in formal protocols and collaboration agreements. This has resulted in making tacit knowledge as explicit as possible. This is the reason that data sharing, as such, is not an issue separate from the production of data. On the contrary, the data policies at CERN form a natural part of the experimental goals and procedures. This has important implications for the role research data play in high energy physics research. The experiments produce such huge quantities of raw data that immediate automated processing is a necessity. The raw data itself is useless for other purposes; only the extraction of meaning from it based on specified research questions by the relevant experts makes sense. Hence, in the eyes of the physicist making "the data" publicly available does not relate to the raw data but to data that has been contextualised, processed and refined. In this process, most of the raw data is discarded.

Therefore, outside of the scientific community, the CERN data is not distributed. The exceptions are data sets for educational purposes, but these have been tailored and filtered. For other scientists to benefit from the data obtained experimentally, the data have to be processed using a variety of different technical and data format standards. These standards are being developed in relation to technological developments. There is not a single universal standard. Several attempts at reaching agreement on one universal standard have been undertaken but so far this turned out to be difficult. The main drive for creating more unified and common standards is not so much data sharing as making the maintenance of the data easier.

Ownership

The data is effectively owned by the CERN collaborations that have produced it. Data sharing between collaborations is hampered by the competition between the researchers. This means, among other things, that negotiations between different collaborations are necessary to enable data sharing. In this process it is not uncommon that reaching agreement on definitions and standards is a prerequisite for successful collaborations.

Restrictions and obstacles

CERN considers that the vast majority of the data it handles would be of little use outside its own specialist part of the scientific community. Indeed, much of the data produced by a single experimental collaboration would not be useful to other collaborations. As collaborations have different standards and working practices, the possibility to meaningfully transfer raw data sets is limited. The majority of data exchanged is therefore the processed data tables.

An important issue is the level at which the data is being made available, as well as the type of data that is involved. As said, much of the raw data produced is useless without heavy

processing, which requires the use of specially written programs and algorithms. The sheer size of the raw data already necessitates processing. Moreover, because of the special format of the data, it does not make sense to anyone except the expert within that particular experiment. Therefore, the data is usually password protected at this stage. This protection has partly to do with the size of the data, but partly also with competition amongst researchers. Most researchers will not give their competitors direct access to this raw data. The raw data resulting from the experiments are therefore extracted on the basis of the specific research question and cleaned up. This cleaning process results in a Data Summary table. This data summary table is made accessible world-wide in a password protected way, because it is still specific to the collaboration involved.

At CERN, major problems in information management are what to archive and in which standard. Different sections of the scientific community have different wishes and perspectives on this, due to the scientific research questions that are central to them. A related issue is that software tends to become obsolete within a few years. As software is vital to access the archived research data, a solution is needed for archiving software together with the data.

Privacy and legal issues

Due to the nature and role of the CERN data, there are no specific privacy or legal problems related to data sharing.

Case study of Data Policies at the EBI

Dr Paul Wouters and Colin Reddy
Networked Research and Digital Information (Nerdi)
NIWI-KNAW
PO Box 95110
1090 HC Amsterdam
The Netherlands
email: paul.wouters@niwi.knaw.nl
www: www.niwi.knaw.nl/nerdi

Introduction

This case study explores the data policies in one of the key institutions in molecular biology. The European Bioinformatics Institute is a non-profit academic organisation that emerged out of the European Molecular Biology Laboratory (EMBL). The official mission of the EBI is to ensure that the growing body of information from molecular biology and genome research is placed in the public domain and is freely accessible to all members of the scientific community in ways that promote scientific progress. The EBI serves researchers in molecular biology, genetics, medicine and agriculture from academia, and the agricultural, biotechnology, chemical and pharmaceutical industries. The EBI does this by building, maintaining and making available databases and information services relevant to molecular biology, as well as carrying out research in bio-informatics and computational molecular biology.

Benefits

The main users of EBI data are scientific researchers in academia and industry. The EBI, as the data service centre of the EMBL, is covered by the commitment to making scientific information publicly available as laid out in the EMBL founding agreement. The EBI manages a number of databases constructed from submissions of data intended to be placed in the public domain. As a consequence, the EBI has a policy of open access to the data. The EBI makes a range of databases available. How this is achieved varies by topic area. The individual databases are available for a complete download and are supported by the EBI making Web tools available. Scientists submitting data are made aware that their submissions will form part of the publicly available databases that the EBI manages.

The EBI not only makes deposited data available for re-use but also adds value to the data by manipulating and analysing it, and compiling new databases from the results. Each of these has submissions and free search services ranging from sequence scanning to full text retrieval, as well as secure on-line submission and analysis of user owned data. These databases fall under different categories: nucleotide databases, structure databases, protein databases, and (access to) literature databases. Much effort is made to ensure that the EBI's products match the requirements of the various users of their databases, entailing the construction of many different databases and avenues of access to present different aspects of the same information as comprehensively as possible.

Relations to other activities

The data are usually submitted to EBI in conjunction with the preparation of scientific publications. Scientific journals often require that large data sets are deposited in an approved publicly available database, with the access number and coordinates included in the paper. In those areas of research where large data sets are being produced, the traditional balance between publishing a paper and publishing the data is upset. In fields like genomics and proteomics, it has become impossible to read all the publications on a certain gene or protein. Instead, researchers will use the most up to date version of the datasets. This underlines the strategic importance of databases in knowledge production in these fields.

The standards in use result from negotiations with equivalent bodies in the US and Japan. In a number of instances, the standards have been developed by the EBI, particularly where resources originate from EBI. In other cases the standards have been the result of discussions within the relevant communities and the EU.

The tools which are applied and offered by the EBI are partly developed in house, and partly produced by the global bioinformatics community. The EBI web site gives access to a wide variety, which enable different types of analysis of data to be produced. The different categories of tools are: general search tools, dedicated DNA, DNA/protein or protein searching tools, submission and annotation of sequences, structural analysis, functional analysis and sequence analysis. Basically, the EBI has amassed a palette of widely available tools, integrating them into a more or less seamless system which facilitates their use.

The software that researchers need to access the data seems to undergo a regular kind of life cycle. This may have important implications for data access and sharing because it is

impossible to try to read the data without advanced software tools provided. The EBI has, therefore, worked together with providers of software tools.

Despite the EBI's strong public ethos, the institute still has strong links and collaborations with industry. An example of this is its industry programme which the EBI maintains is "consistent with the public domain policy of the EBI".

Ownership

The EBI deals almost exclusively with public data that are not covered under any IP encumbrance. There is no commercial exploitation of the data that the EBI produces. This is mainly because the data it deals in has been supplied by scientists on the basis that it is going to be made publicly available. This does not mean, however, that commercialisation is not an issue at all. The software EBI develops to process it may be spun out commercially. However, even that could lead to problems. The IP that the EBI would generate would be from software, but since that software has been developed to allow people to look at their own data it would be very difficult to really try to commercialise it. With regards to the basic information, the EBI wishes to stress the non-commercial character of the data processing and archiving process. The EBI makes this argument with reference to the need to maintain relationships of trust with the scientists.

It also has to be borne in mind that, as the management sees it, the EBI works in a sensitive area of molecular biology which prevents it from engaging in vigorous commercial activities for ethical reasons.

Restrictions and obstacles

The main premise of the EBI is that data will be made public. Researchers may ask for a limited period of confidentiality as an exception to this rule, usually to enable them to analyse the data and publish their results before their competitors. This period is limited, after which publication follows. The EBI accepts no restrictions on the use of publicly visible data. However it can keep data confidential for a period specified by the submitter, with the proviso that data discussed in a publication will be released as soon as the publication appears (even if the original confidentiality period requested has not expired).

The EBI has a major problem in its funding basis. The key issue is that funding for data sharing infra-structures needs to be constructed "on a rolling basis", in contrast to the project based funding dominant in science. It relates to the infra-structural nature of databases as well as data sharing environments and tools. This creates a tension between, on the one hand, the funding schemes of research (usually project oriented) and, on the other, funding schemes for infrastructure (of which data requirements form an increasingly important part). This contradiction between the need for funding on a rolling basis and the practice of project funding is the main threat to EBI's commitment to maintaining the public availability of its data.

The EBI sees itself as custodian of data, with a responsibility to maintain a public record of science. The EBI takes into account that the present policy drive to secure more resources from non-public funding could compromise its activities in the public sphere.

With the increase in the amount of data and the variety of data types, information management tasks in research are becoming more complex. At the EBI, the variety of ways in which data is submitted, along with the steep increase in the amount of data, is a continuously mounting challenge. A related issue is the amount of data and the variety of data types that will have to be dealt with in genomics and proteomics research.

Privacy and legal issues

Most legal and ethical issues deal with ownership issues and limitations to access (see above). There are no specific privacy issues in the life cycle of the data after their submission to the EBI, although they may well have arisen in the research that produced the data. Ethical issues do rise again in the stage of re-use. The EBI databases are publicly available, and some uses of this data will be ethically dubious. For example, recent concerns have been raised about the possible use of EBI data to produce biological weapons.

Case Study of Data-sharing at the fMRI Data Center, Dartmouth College, USA

Dr Anne Beaulieu
Networked Research and Digital Information (Nerdi)
NIWI-KNAW
Joan Muyskenweg 25
Postbus 95110
1090 HC Amsterdam
The Netherlands
Tel: (31)(20) 462-8739
Fax: (31) (20) 665-8013
anne.beaulieu@niwi.knaw.nl
http://www.niwi.knaw.nl/nerdi/

Introduction

This case study discusses one of the principal endeavours to promote data sharing in brain mapping, the functional Magnetic Resonance Imaging Data Center (fMRIDC). Launched in 2000, this database is spearheaded by Dr. Michael Gazzaniga, Director of the Center for Cognitive Neuroscience, Dartmouth College, US. The fMRIDC receives funding from NSF/NIH, the Keck Foundation, and Informix (IBM) and Sun Microsystems. Part of the funding of the fMRIDC is provided by the NSF/NIH, under the aegis of neuroinformatics/the Human Brain Project, which is an important funding and coordinating mechanism for neuroscience databases in the US.

Benefits

The goals of the Center are to provide "a publicly accessible repository of peer-reviewed fMRI studies and their underlying data."[28] The fact that the database is publicly accessible and that it archives raw data differentiates the database from other projects in the imaging community, which circulate processed results or include commercial elements. Two important notions guide this data repository: the scientific benefits that can be derived from sharing data and the importance of the public accessibility of research data.

Whether and which benefits will follow from this data-sharing initiative remain to be proven for what is a new initiative in a fairly young field. Objections to this data-sharing initiative have been voiced by a significant number of researchers in the field of functional imaging. The underlying concerns focus on the possibility of separating data from experimental contexts. Because no clear distinctions exist in this field between data management and research, the benefits of the circulation of data are not yet visible to this community.

Relations to other activities

The new flow of data developed in relation to this data repository interacts with existing practices. The main quality control mechanism in the fMRIDC is the traditional peer-review system associated with journal publication. The data deposited in the database is therefore only that derived from peer-reviewed papers, accepted for publication. Up to now, all contributions have come from a single source, the *Journal of Cognitive Neuroscience.* Authors who wish to publish in this journal must submit their data to the fMRIDC. Data from articles in *Journal of Neuroscience* and *Cerebral Cortex* have also been submitted but were still being processed as of January 2003.

This interaction, meant to serve as both incentive and as quality control mechanism, and therefore, affects traditional publishing practices. An earlier effort in the functional imaging community to build a database of published results depended entirely on the voluntary contribution of researchers, but was not very successful in attracting submissions.[29] Three issues arise from this coupling. First, it has been perceived as an indirect, negative incentive ('stick' rather than 'carrot') to share data by researchers at the launch of the database, and a highly visible controversy ensued. The result was that many researchers and editors of other journals distanced themselves publicly from the initiative. One other journal has since endorsed the initiative, though it encourages and does not require submission of data by its authors.

Second, an analysis of the controversy about this coupling makes clear that data sharing, and the circulation of data more generally, is not considered to be a clearly separate activity from pursuing and communicating about research. In other words, because data management has traditionally been integrated in the pursuit of experiments, the functional imaging community lacks conventions and protocols for manipulating data separate from the context of specific experiments. Objections to data sharing, such as 'data is not meaningful to anyone else' or 'data cannot be abstracted from the specific experiment' are related to the fact that data management and experimentation are not separate in this context. Effectively, the fMRIDC proposes a new stream for data, one that contrasts with the research activities with which researchers are familiar. While this new stream offers new possibilities, it also affects existing practices of research, also in a 'downstream' manner. This part of the debate, therefore, illustrates that by making submission of data a compulsory activity in relation to publishing, a range of factors relating to the structure of the field come into play, besides the willingness of individual scientists to participate. In contrast to other fields, such as high energy physics, molecular biology, and some areas of biodiversity, functional imaging does not have a clearly articulated relation between research and data management as separate sets of activities.

Finally, a third aspect of this interaction is the unintended effect that data submission requirements may have on the peer-review system. The fMRIDC relies on the association with publication as a trust-building mechanism, to give potential users of the database some assurance

of the quality of data. Two important consequences should be queried in relation to the reliance on peer-review. First, it may put added strain on the peer-review system if data are also to be examined in detail by reviewers. Second, reviewing data may be different from what is involved in evaluating a publication. More work and a different kind of work may thus be demanded of reviewers. Whether reviewers will, and perhaps even should, be up to this task in the view of current data floods is an important question for journal editors, developers and funders of data-sharing infrastructures.

There may be other tensions between publishing and data-sharing, and it is not clear that the interaction is as symbiotic as one might imagine. The cross-references between journal and database are minimal. Articles in the *Journal of Cognitive Neuroscience* include an 'accession number' in the acknowledgement section, which identifies the data in the fMRIDC. Some electronic versions of the journal provide hyperlink via this accession number to the fMRIDC website. Otherwise, there is no visible interaction between the publisher (MIT Press) and the database, neither mutual hyperlinking nor submission requirements being mentioned in the instructions to authors. This may indicate divergent interests between the journal, which aims to provide unique materials to its subscribers, and the goal of the fMIRDC to make data freely and widely available.

Ownership

In the original announcement of the fMRIDC, the directive to submit all data regarding an experiment at the time of publication of an article raised some concerns. While the debate was framed in *Nature* as one of ownership of data, it could also be framed in terms of organisation of control and use of data. In further discussions with researchers in the course of fieldwork, the issue of ownership was rather less prominent, and the consequences of a new circulation of data were fore grounded. Objections mainly concerned the circulation of information in relation to the organisation of research. The argument went as follows: Submitting data to the database might have different consequences for researchers working in different settings. In larger labs, where one has many post-docs, one can have the analysis of an experiment done 'in parallel', so that all papers are then submitted more or less simultaneously. In smaller centers, the analysis proceeds a bit more slowly, resulting in a more linear submission pattern. The danger of being 'scooped' would be, therefore, much greater for smaller labs. The issue here cannot be addressed solely with the principle that publicly funded data should be publicly available. By extending the notion of ownership to include the issues of control of data, a better characterization of the problems of data-sharing will emerge. This finding indicates how a more sensitive problem definition can lead to better recommendations for data sharing policy.

Restrictions and obstacles

Conventions need to be developed

The importance of field-specific conventions about how knowledge and data can packaged is important for understanding access to data. These conventions are not yet stabilised in functional imaging, as made visible by the various aspects of the controversy around the launch of a database. There is no consensus answer in the research community to questions like: How is data best described? How can the results of different analysis packages be compared? What is the best format for data?

The launch of the fMRIDC has stimulated discussions about conventions for data, and the importance of coordination of standards and formats has been placed on the agenda of various funding and professional bodies. Still, these conditions may not achieve a culture of data sharing in and of themselves. Remaining hurdles to data sharing fall under two broad categories: lack of clear incentives and constraints in data-sharing, and the need to consider interactions of the new data flow with other dynamics in the field.

Obligations to share

Funding organisations also have a potential role to play in shaping incentives and obligations to share data. A policy of the NIH, currently under development, may mean that data sharing becomes an increasingly explicit requirement for researchers. As such, the fMRIDC may provide researchers with an existing infrastructure to comply with data sharing requirements of funding agencies.

Support for new types of work and novel tools

By all accounts, developing good data management practices involves work and new skills for researchers. The amount of work involved in preparing and submitting data was mentioned as a disincentive in discussions with researchers around the fMRIDC. The relative burden placed on researchers, and the possibility for support for this kind of work may play a role in the degree and speed with which data sharing develops.

The value of submitting data to the fMRIDC has not yet been demonstrated to researchers in terms of clear research benefits.[30] As mentioned, the software tools needed to be able to handle, let alone 'exploit', data across studies are not yet well developed.[31] This has implications for 'added value' functions of data repositories, and for the motivation of researchers to re-use data. The incentives for developing software for the use of databases, however, may not be so obvious to software developers working in laboratories. In these settings, the agendas are set by the local needs for tools for analysing experiments. It may be necessary to change the current culture of free distribution in the functional imaging community, to one where commercial exploitation of software provides incentives. Another possibility is tailoring funding possibilities to support this kind of work.

Privacy and Legal Issues

Ethical Issues and Human Subject Data

The construction of databases and data sharing raises novel ethical issues, especially in relation to the open-endedness of use and circulation of data contained in databases. This open-endedness is desirable in the eyes of developers and users, but can be particularly problematic for bodies charged with regulation of research ethics. It is important to note that the Western tradition of informed consent in bio-medicine is shaped by the principle that the 'the most specific consent is the best consent'. When data is to be gathered for submission to databases, this specificity may run counter to the goals of meta-analysis or re-analysis by third parties, to investigate issues different from those for which the data was originally gathered. The creation of infrastructures for data sharing therefore interacts with regulatory bodies (such as institutional review boards or 'IRBSs'), whose approval must be obtained to share data. As such, these bodies affect the circulation of data.

More attention, on the policy level, to the interaction of regulatory bodies and data sharing initiatives can be very valuable. Especially deserving of attention is the coordination of various IRBs, since there is anecdotal evidence that various institutions' IRBs may respond differently to novel ways of working spurred by data-sharing initiatives. Furthermore, international coordination may also be a worthwhile course of action. Researchers submitting or requesting data across national boundaries may find it especially difficult to act in accordance with the various ethical guidelines that exist in different countries. The fMRIDC has also been hesitant to accept data from non-US settings because of concerns regarding IRB compliance.

Moreover, these regulatory bodies are also relevant to the larger context of data sharing, since they function as trust building mechanisms for the public. The dangers of breaches of privacy from brain scanning data have already been the subject of attention in the media. The alignment of practices of brain scan repositories with the requirements of ethics committees may therefore also alleviate some of the concerns of the public about privacy issues.

Case study of data policies at Global Biodiversity Information Facility and other biodiversity data facilities[32]

Kathleen Casey
Department of Communication
University of California, San Diego
9500 Gilman Dr.
La Jolla, CA USA 92037-0503
kcasey@ucsd.edu

Introduction

A range of initiatives has arisen to make data on biological specimens available through digital environments. From the international initiative, the Global Biodiversity Information Facility (GBIF), to national initiatives such as Comisión Nacional para el Conocimiento y Uso de la Bidiversidad (National commission for the knowledge and use of biodiversity, or Conabio), to individual natural history museums, such as the Museum of Vertebrate Zoology at the University of California, Berkeley, the aim to make data on biodiversity accessible has been welcomed by a broad spectrum of communities, including those representing scientific, governmental and commercial organisations.

Because GBIF is not yet fully operable as of this writing, the research on biodiversity data access practices has involved interviews with an array of scientists, museum directors, software developers and other participants from different countries who have been directly involved or affected by the process of making biodiversity data freely accessible. This research contributed to an understanding the general landscape of biodiversity data access practices that GBIF is aims to improve.

GBIF, which began under the auspices of the OECD, aims to become an interoperable, distributed network of scientific biodiversity databases whereby different hosts of databases 'affiliate' with GBIF in the shared goal of making "the world's scientific biodiversity data freely available to all."[33] More specifically, GBIF aims to "to design, implement, co-ordinate, and

promote the compilation, linking, standardization, digitisation, and global dissemination of the world's biodiversity data, with an appropriate framework for property rights and due attribution."[34] This work will amount to bridging other regional and local data networks to create an interoperable network. Some of GBIF's strength will entail turning to other institutions that have already developed open access regimes. On the other hand, regional networks require GBIF to take the a crucial next step to help create an interoperable network by coordinating and connecting what is already present or under development so that regional initiatives can grow outward and by motivating the development of other national nodes that will draw together different institutions. GBIF is in this way an important leveraging mechanism.

Benefits

Access to biodiversity information through a digital environment is expected to bring an array of benefits to a range of users, including scientists and government officials. These benefits include some of the following:

1. Helping heterogeneous classes of users acquire data more efficiently;
2. Providing greater visibility and use of natural history museums and other institutions' collections, including use that leads to correction of errors in data;
3. Improving the ability for scientists to explore new research areas, including research that requires querying data from several to many different institutions, often simultaneously;
4. Altering the nature of data by making them 'dynamic,' and bringing them into larger data sets of comparable and differing data. These new opportunities to work with digital data also promote the development of software tools that will allow further processing and use of the data, such as geo-referencing software tools;
5. Providing the means to acquire information needed for public policy formation, conservation projects, economic development, education, and other projects that entail the use or conservation of biological resources to which biological data refer;
6. Furthering the repatriation of data to countries whose biodiversity information is housed outside their borders;
7. Making research more efficient and cost effective through the pooling and sharing of data resources.

In general, most scientists think that by making their data accessible to the public, their research will contribute to knowledge that can promote sustainability in the wake of threats to biodiversity and the environment.

Relationship of data access to other scientific practices

Ease of access to biodiversity data is reported to have been a long-standing need for scientists. The answer to that need has been promoted by the inception of the Internet and multiple international efforts to develop interoperable facilities such as GBIF. For scientists, the ability to access data on biodiversity reduces the amount of work it takes to find information. For instance, if a scientist is looking for information on a given specimen, he or she will no longer track down information through phone calls or travel to museums. In turn, museums will receive

fewer electronic requests to send out physical specimens housed in their collections to scientists who find information about those specimens in museums' on-line databases because the need to examine every specimen is significantly reduced if the label data of all specimens are available electronically. Take the following as an example of the change in access practices:

> "In the first year after the inception of its publicly accessible database, the Museum of Vertebrate Zoology (MVZ) web site fulfilled 41,937 [electronic] specimen queries, representing 19,001,503 specimen records delivered. This is up from 95 requests representing 160,471 specimen records delivered manually by its staff in the preceding year." [35]

Digitized data and network availability are changing the nature of the work of data archives and museums. Instead of tracking down information for scientists who email or call, curators are spending their time (1) carefully sorting materials for scientists and (2) performing data entry and clean-up.

By querying multiple datasets across several to many natural history museum collections, scientists can compile and compare data for processing through available and developing software tools to produce information about ecological systems, placement of natural reserves, and many other areas of interest. However, for these goals to be fully realized, data standards and metadata standards must be developed so that the millions of records from the thousands of potential databases can be made fully searchable. These standards development projects are being forwarded by a number of national and international initiatives. And, of course, the data must be digitized.

The development of improved access to biodiversity data is most often accomplished through the support of researchers, educators, and publishers for new kinds of projects based upon data that are digitized specifically for these purposes. For example, Conabio provides grants to universities, among other institutions, to support fieldwork and compilation of biodiversity data to be placed on the Internet. As such, more research and training of doctoral students has been taking place in the field. Publications that rely upon the aggregation of data drawn from many institutions and recently developed software tools, give rise to new forms of research and knowledge, as well as information necessary for policy and economic development projects. It is important to note, however, that there are great stores of untapped data that are laying waste if funds are not provided to digitize them, even though their immediate usefulness to a particular project is not readily apparent.

Restrictions & obligations

A pertinent lesson from GBIF is the urgent need for the development of international agreements on data access policies that will allow the open and free flow of information about biodiversity on the Internet. The scientific and policy developments that such agreements could engender would contribute much to science and sustainability. Hindrances to information flow among all countries include:

> ***Technological***: Lack of hardware and connectivity
>
> ***Institutional and Managerial:*** Lack of personnel and training to continuously support the infrastructure required for data access

Legal and Policy: Lack of open policies

Budgetary and Financial: Lack of funding scenarios and investments that will make biodiversity information perpetually available worldwide

Cultural and Behavioural: Lack of appropriate career and reward structures for scientists who make the intellectual effort to place their data in scientifically credible databases

The lack of open policies is the greatest barrier to progress toward GBIF's goals. Budgetary and cultural hindrances—which directly affect the availability of hardware and network connectivity—are summarized below.

Budgetary and Financial Issues

Biodiversity information is required by scientists and decision makers throughout the world. For biodiversity data to be made available on a global scale, however, investments need to be worldwide to enable the relatively small number of "high data" institutions to "liberate" those data. As discussed in the main body of the Working Group's report, the sound stewardship of public resources entails making the relatively small investment of making scientific research data publicly available. In the case of biodiversity data, such stewardship clearly takes on a global character because information about biodiversity is required to deal with the sustainable use of our planet's resources. Simply said, the global issue of biodiversity requires global investments in making accessible scientific research data about biodiversity.

At the institutional level, funding schemes are not always appropriate to implementing and managing data access facilities. The first issue is the type of databases that are developed. Often, institutions may be able through small grants to digitize their data or create a standalone database that is useful for their local users. To make a system that may be used by the heterogeneous user communities from local and global access points is exponentially more expensive. Funding must be appropriate to the costs of not simply creating a database, but an interoperable database that can become part of a sustainable infrastructure. This cost entails technology and skilled staff and time. Museums may not have these resources available, and must ultimately weigh these costs against the general costs of operating the institution.

Another issue for funding for electronic data access is the scope of the funding, which affects the type of network infrastructures that can be developed. Grants often are directed at a single, or sometimes a few, institutions. This limited scope may result in the development of weak or unsustainable infrastructures; that is, infrastructures that are unable to draw in other databases. Or, funding will be provided to a certain kind of collection—say, a herpetology database collection—rather than across many kinds of collections. This limits the possibilities for research on complex phenomenon such as ecological relationships among organisms and other questions regarding inter-organism relationships. It can be even more difficult to draw databases into an interoperable network across national borders. GBIF's aim is largely to fulfill this demand, but the development of actual databases and networks will largely take place at the national level. Thus nations must fund these programs at a level appropriate to creating a national node.[36] Funding also comes in short-term, not always continuous, episodes so that infrastructural needs may not receive the consistent support they require. Nor will the institutions have the rest of their infrastructures maintained alongside their information infrastructures.[37]

Overall, the issue of making biodiversity data publicly available on a global scale requires novel approaches to investing in national and international network infrastructures as well as a commitment to funding digital data facilities in ways that overturn the provincial approach that is generally maintained today. This short-sighted approach not only affects the efficiency and cost-effectiveness of scientific research, but the decision-making process about how to use the planet's resources in an economically and environmentally sustainable fashion.

Career and reward structures

The relatively new enterprise of making data available through on-line databases does not entirely match up with time-tested academic career and reward structures. Biologists to date have not been trained in the skills necessary to make their data electronically useful, nor do they have expertise in software development. While software programmers may have such skills, they usually do not receive training in the needs of the biological sciences. However, the requirement for these skills has become more apparent, leading to the development of outreach and training programs, such as those at GBIF, and nascent programs at individual universities.[38]

Typically, biologists do not receive any financial rewards for taking the time to place their data into a database. When academic recognition is most directed toward publication, the direction of one's research toward the end of sharing data is undermined. Tenure reviews do not reflect the intellectual effort of depositing data into a scientifically credible database. It is vastly clear that vastly more data would be made publicly available if academic criteria for promotion and tenure were expanded to include the development of data stores that are dynamically and perpetually available. As reflected in the main body of this report, there is no real institutional support for data work.

Ownership

In general, scientists believe that publicly funded research on biodiversity that results in the collection and curation of biological specimens results in the public ownership of those data. However, there are cases where it is not clear who collected a given specimen and whether the information about that specimen can be made publicly available. In cases where an institution holds specimens taken illegally, the ownership of the specimen and related data are equally problematic. Institutions within nations that have ratified the Convention on Biological Diversity must follow the conditions of that agreement. These conditions include sharing the monetary or other benefits that may be derived from the specimen and/or its associated data with the nation or indigenous area from which the specimen was taken. While the ownership of data is distinguishable from the ownership of biological resources, there is potential overlap that concerns some museum directors. Issues of ownership need further clarification.

Data Repatriation

A large portion of biodiversity data is housed in a relatively small number of countries. The repatriation of data from these repositories to the nations who do not hold the data on their own biodiversity is recognized as critically important both by the Convention on Biological Diversity and by those nations who require the data for scientific and economic endeavours.[39] While some countries have made great efforts to repatriate data through the development of data networks and other facilities, there is still much financial and technical support is required to develop the infrastructure necessary to fulfill this goal.

Privacy & other Legal issue

Conditions of data use

Data within collections from publicly accessible collections can be restricted if they include: information pertaining to the location of endangered or otherwise sensitive species, or the names of people, such as indigenous hunters, who collected specimens deposited at the institution. Data may also be inaccessible if they are being actively researched or to be used in publication.

Institutions with publicly accessible databases generally assert that their data may be used for research purposes only. A more specific example of this condition, from REMIB's user agreement, is that users querying the data must not use their data in any way that will harm the "equilibrium of ecological systems" or other conservation programs. Other conditions for data use generally encountered include the following: (1) users querying the data must acknowledge the original source (and sometimes the database or network owners) of the data in reports, analyses or other publications that rely upon the data; (2) users must receive prior consent for any repackaging, reselling, or redistribution of the data from the institution that provided the data; (3) users must agree to not redistribute the data at all; and/or, (4) users must not fault the institution for inaccuracies or errors in the data.

Intellectual Property Rights

Intellectual Property has been labelled as one of the most unclear components to tackle in creating open access to biodiversity data. In part, this uncertainty stems from the awareness that museums must remain economically viable institutions, despite reductions in government support. Intellectual property rights (IPRs) are believed to be one way that museums can recover the costs of developing databases, for instance. IPRs are also recognized as something that scientists may pursue for personal benefit and protection, such as in cases where scientists are competing for publishable research on a given specimen. There is some sense that these scientists do not fully consider the implications that their own attempt to hold IP has for the museums that they rely upon to store their specimens and conduct research. Overall, the extent to which intellectual property rights might apply to biodiversity data is unclear to many scientists.[40]

CONCLUSION: A RESEARCH AGENDA FOR SUCCESSFUL DATA POLICY IMPLEMENTATION

The world of data sharing research can be broken into four parts:

- The creation of new knowledge and the practice of data sharing (in particular data re-use);
- The role and nature of the repositories to be built;
- Institutional facilitation of data sharing;
- The ethics and politics of data sharing.

Our four case studies have spoken to each one of these areas. In this conclusion, we highlight a research agenda for each. We draw attention to what we consider the most pressing research questions emerging given the current state of knowledge.

The practice of data sharing

It is surprising to many that despite the general assertion of the use of data sharing for science in the production of basic knowledge, economic innovation or global environmental management, there are almost no detailed studies of the actual practices of data reuse in science.

At present, three sources of information on this exist, besides the case studies in this report. These are policy reports by academic organizations, which are often partly based on experience of researchers themselves; anecdotal evidence from researchers and database developers; and more formal but highly indirect measures such as 'hits' to a database website or counts of shipments of datasets. Presently, there is not enough information to seriously evaluate whether present data-sharing policies are achieving their goals.

We need both metrics of data reuse and detailed ethnographic studies of data sharing practice to inform policy development. The conclusions in the four case studies indicate that this combination may indeed be particularly fruitful. Unexpected barriers to data sharing may arise from the interaction between practice and regulation that are difficult to identify without studying science as it is actually practiced, on the 'shop floor'. And we should not forget that data sharing does come as naturally to scientists as the ethos of science assumes. Identifying the barriers specific to different fields can be invaluable in developing effective policy.

Repositories

A commonplace from the computer scientist is that with data sharing we will be able to query multiple databases, in different formats and from different sources. Thus, at GBIF one will be able to draw on information from all over the world in order to decide about the health of a given species. However, there is little research on repository development. Our CERN case study pointed to widely divergent views on data standards, format, selection and mode of archiving at CERN.

Quality control mechanisms are central here, especially with the spectre of floods of information from multiple sources. There are a number of models currently in operation – from a formal peer review process, to the designation of gatekeepers, to building in community mechanisms for 'cleaning up' databases in use. We need more studies that investigate which mechanisms work best for given forms of scientific practice.

Repositories are sites of potential conflict between the public nature of the data and the private production of the software needed to read and analyze the data. This area is worthy of further study. A particular component of this study could look at the relation between open source software and proprietary code.

Institutional facilitation of data sharing

There are calls in science policy to rethink the basic institutions within the scientific system. The relationship between academia and publishing houses is at stake in the wake of the World Wide Web. New research fronts have been opened that are fundamentally interdisciplinary while most universities are still organized along disciplinary lines. Equally, data formats, standards and data sharing practices fall along disciplinary lines. This discontinuity may hamper the exchange of information and data between different fields.

Again, there is as yet no concerted research agenda to address the nature of the institutional changes that are occurring across a broad span of disciplines. A pool of shared knowledge about this issue will be immediately relevant for policy implementation, and at the same time will address some basic research issues in social science.

Traditionally, libraries, collections and paper archives have been crucial centres of scientific information, data and objects. The "informational turn" in science has undermined the traditional roles of these institutions (libraries, publishers as well as archives) by integrating these functions more closely with daily research practices. This process is redefining the tasks of librarians, publishers, archivists and researchers themselves. It has led to the emergence of new types of support staff (ICT experts, information scientists) and of new fields, such as bio-informatics.

A key issue that emerges from our case studies here is that of the funding cycle for research. The development of data sharing protocols is a part of this emergent global infrastructure. This infrastructure needs ongoing maintenance, skilled staff and time, and its governance needs careful attention. However, scientists working on data sharing initiatives are generally funded for only three to five year research cycles, and infrastructures are not funded at an appropriate level. We need to understand from the ground up how infrastructure can be grown.

The Ethics and Politics of Data Sharing

It is not universally accepted that data sharing is a good thing. In the biodiversity world, there has been talk of 'information imperialism' to be answered by the repatriation of biodiversity data to its country of origin, for example. Or again, why share information about an endangered species if a hunter might use that information to hunt it to extinction? This uncertainty leads into a discussion of the nature of the ideal polity for data sharing.

Equally, fundamental ethical issues are frequently posed in data sharing policies. Thus, sharing data about human subjects has been shown to be a point of serious concern for a number of parties, including researchers, IRBs and the public. Underlying these concerns are important changes in research, where not only research practices but also data management must be regulated. We need studies of the formal and informal effectiveness of such regulations.

Data sharing is central to global economic development and to global planetary management around such burning issues as the supply of fresh water, the preservation of biodiversity and the effects of global warming. Only with a set of policies informed by scrupulous quantitative and qualitative empirical research will we be prepared to face the challenge of creating equitable and effective policies.

Publication

Finally, a systematic and detailed analysis of the negotiations about publication delays and/or conflicts about priority on analysis of primary and secondary data seems necessary. An increase of this type of conflict may be expected in the most competitive branches of molecular biology. The relative importance of published articles versus data sets for the development of new biological knowledge should also be studied carefully. This balance is shifting in a number of biological areas, but probably not in all of them, and not always in the same way. Given the

scarce funding for research in most countries, it is crucial to have a better understanding of the most critical factors for knowledge creation in these critical areas.

Endnotes Annex 1

[27] Paul Wouters & Peter Schröder (eds.) (2003), *Promise and Practice in Data Sharing*, in: *The Public Domain of Digital Research Data* (Paul Wouters & Peter Schröder eds.), NIWI-KNAW, Amsterdam, the Netherlands, ISBN 90 6472 184 x

[28] http://www.fmridc.org/ (29 May 2002).

[29] Its format is currently being revised and the database re-launched.

[30] The fMRIDC is attempting to demonstrate this, by encouraging articles, which reuse data from the fMRIDC, to be submitted to the Journal of Cognitive Neuroscience.

[31] As of January 2003, tools of this kind are presented and promised on the website of the database.

[32] Research for this report entailed interviews of scientists, museum and herbaria directors, program directors, and software developers involved in the field of biodiversity. Institutions included in the research are: Comisión Nacional para el Conocimiento y Uso de la Bidiversidad (National commission for the knowledge and use of biodiversity, or Conabio), Global Biodiversity Information Facility (GBIF), the Illinois Natural History Survey, and the Museum of Vertebrate Zoology at the University of California, Berkeley (MVZ).

[33] GBIF Presentation at Fourth Global Science Forum, OECD, 25-26 January 2001 by Ebbe Nielsen.

[34] Business Plan for the Global Biodiversity Information Facility, Section 1.1

[35] See "Impact of the Project" section of the MaNIS NSF proposal, available at http://dlp.cs.berkeley.edu/manis/ProjectDescription.html)

[36] In line with the EU's "Subsidiary principle"

[37] For example, the Illinois Natural History Survey has managed to make some headway in this regard through soft money provided by the Illinois Department of Transportation, which had an interest in having easy access to information on biodiversity in Illinois. But, because this is soft money, the staff support necessary for these facilities are temporary and thus open to yearly cut-backs. The result would be detrimental to the facility and the public who relied upon it. Hard money cannot be directed towards the electronic access facility because it is not sufficient enough when weighed against the costs of running the institution as a whole. Conabio is an example where federal funds have been directed toward creating a national infrastructure; however, their federal funds are also determined annually, making for a yearly struggle for that institution in their efforts to maintain and develop a broad-based, publicly accessible network with updated data.

[38] Indeed, Conabio has seen an increase in Ph.D's training in the field who have contributed to and worked with Conabio's databases.

[39] One of the 5 operational objects of the Global Taxonomy Initiative of the CBD is to "Facilitate an improved and effective infrastructure/system for access to taxonomic information; with priority on ensuring that countries of origin gain access to information concerning elements of their biodiversity." See Annex I, Decisions adopted by the Conference of the Parties to the Convention on biological diversity at its sixth meeting, Sec. VI/8, Global Taxonomy Initiative, available at www.biodiv.org. See the reports on the CBD meetings in The Hague, 7-19 April 2002 at www.biodiv.org

[40] Research on the different policies held by institutions that house biological specimens is currently being conducted for the European Natural History Specimen Information Network (ENHSIN). Simon Owens of the Royal Botanic Gardens, Kew, conductor this research, states that ENHSIN is a "pilot project to look to see whether we could link up databases at a European level" and whether "people could have access to the data and use it for all sorts of things." So far Owens and his colleague, Alyson Prior, have found that "no museum or garden yet has a firm policy on IPR either for specimens or for databases (yet laws exist)." See Simon J. Owens and Alyson Prior, "Beset with pitfalls—specimens and databases, intellectual property and copyright" from the 2000 meeting of the Taxonomic Databases Working Group; 10-12 November, 2000; Senckenberg Museum, Frankfurt; Digitising biological collections, available at www.tdwg.org/tdwg2000/ipr.htm

ANNEX 2.
Participants of the Follow-up Group

I. Members

Peter Arzberger (Chair)
University of California, San Diego - USA

Peter Schroeder (Vice-Chair)
Ministry of Education, Culture and Science - The Netherlands

Geoffrey Bowker
University of California, San Diego - USA

Sigrun Eckelmann
Deutsche Forschungsgemeinschaft (DFG) - Germany

Tim Hubbard
Wellcome Trust Sanger Institute - UK

Koji Kamitani
MEXT (Ministry of Education, Culture, Sports, Science & Technology) - Japan

Leif Laaksonen
CSC - Scientific Computing Ltd - Finland

Doug McEachern
Australian Research Council - Australia

David Moorman
Social Sciences and Humanities Research Council - Canada

Masamitsu Negishi
National Institute of Informatics - Japan

Paul Uhlir
U.S. National Academy of Sciences/National Research Council - USA

Andrzej P. Wierzbicki
National Institute of Telecommunications - Poland

Jan Windmueller
Ministry of Science, Technology and Innovation - Denmark

II. Experts

Anne Beaulieu
Networked Research and Digital Information (Nerdi) - The Netherlands

Kathleen Casey
University of California, San Diego - USA

Colin Reddy
Networked Research and Digital Information (Nerdi) - The Netherlands

Paul Wouters
Networked Research and Digital Information (Nerdi) - The Netherlands

III. Observers

Jacky Bax
Ministry of Education, Culture and Science - The Netherlands

Kathleen Cass
CODATA Secretariat - Paris, France

Gudrun Maas
OECD, Directorate for Science, Technology and Industry
Science and Technology Policy Division

Tony Mayer
European Science Foundation - France

David Schindel
National Science Foundation - France

Hugo von Linstow
Global Biodiversity Information Facility- Denmark

IV. Administrative Coordinator

Teri Simas
University of California, San Diego - USA