THE PUBLIC DOMAIN OF DIGITAL RESEARCH DATA
THE PUBLIC DOMAIN OF DIGITAL RESEARCH DATA

Paul Wouters and Peter Schröder
Series Editors
Contents

Peter Schröder
Digital research data as the floating capital of the Global Science System 7

Paul Wouters and Colin Reddy
Big science data policies 13

Kathleen Casey
Issues of electronic data access in biodiversity 41

Anne Beaulieu
Research woes and new data flows 65
Digital research data are developing into a third stream of scientific capital.
Digital research data as floating capital of the Global Science System

Peter Schröder
Co-ordinator Information Policy, Ministry of Education, Culture and Science, The Netherlands, p.schroeder@minocw.nl
Vice-chair of the OECD/CSTP Follow Up Group on access to publicly funded research data (Website: dataaccess.ucsd.edu)

Yesterday’s scientists studied nature
Today’s scientists study digital data

FROM INDIVIDUAL SCIENTISTS TO SCIENTIFIC INSTITUTIONS

Science once was about scientists creating knowledge by researching nature. As time passed science grew in size and complexity and became a more co-operative and cumulative affair. Science developed a Global Science System involving larger teams of scientists (human resources) using larger instrumental facilities (capital goods) to study aspects of nature (data) in order to disseminate knowledge (product).

Scientific knowledge differs from other forms of knowledge in its higher level of precision, more systematic structure and broader scope. As the information to be processed is more precise and systematically structured, the impact of the use of Information and Communication Technologies (ICT) on productivity will be greater. No real wonder that the use of ICT has dramatically increased the productivity and quality of science.

A SMALLER WORLD

The use of ICT also has spectacularly broadened the scale and scope of science. In this way ICT have enlarged the potential for international co-operation and scientific globalisation enormously. No need to be far sighted to see the current and future benefits of ICT for international communication amongst the extensive networks of human resources, international remote access to the expensive instrumental capital goods of science and international dissemination of the published knowledge. Science always has been an international endeavour, global use of ICT has much enhanced this character.
BROADENING THE SCALE AND SCOPE OF DATA USE

But there is more. Science embracing ICT puts a substantial premium on aspects of research and parts of research trajectories that can be digitised. The resulting developments have enhanced the importance of digital research data and made them into a far more autonomous factor in scientific research. Research data can be defined as the factual records (numerical scores, textual records, images and sounds) used as sources for scientific research. In the now almost forgotten non-digital context, use of most research data was limited largely to specific projects and their researchers. Outside this circle data-access first and foremost fulfilled an indispensable part in the (peer review) validation of research results. The US National Institutes of Health states: ‘Final research data comprise the recorded factual material commonly accepted in the scientific community as necessary to validate research findings’.

DATA AS A THIRD STREAM OF CAPITAL

In the non-digital environment, research data could be classified as partly human resources and partly knowledge. For one part ‘personalised’ between the ears of the researcher, inaccessible to the non-initiated, for the other part available in publications of research results. Digitisation characterised by automated collection and processing of standard records, made the data category more independent from specific research and researchers. Once collected, the same set of standardised digital research data now could be used by a variety of researchers from different institutes, disciplines and nations, at the times and places convenient to the individual users for an unlimited period of time. In a growing number of cases the collection and exploitation of data has been outsourced to specialist institutes. Alongside human capital and the instrumental capital goods, digital research data are developing into a third stream of scientific capital.
DIGITAL VALUE INCREASES BY ITS USE

Once collected and safely deposited, the promise of research data is that in principle they can be put to use all over the global science system. And this can be done in principle at negligible additional cost, once the prospective users have the necessary ICT infrastructure available. In this way, sharing of digital research data opens up substantial new vistas for international scientific cooperation.

Starting with the data will simplify co-operative arrangements because the value of the data in principle will only increase by making them available for multiple uses.

Data in principle are wear-proof and the more they are used by more researchers, the more results they will produce. After all the ultimate goal of the parties investing in the collection and storage of research data is to get as much knowledge (end product) as possible out of their investments in data resources. In the context of publicly financed science, the investor does not have to worry even if total strangers are using ‘his’ data: the results will be publicly available. The more data value you spread around, the more knowledge value you will get in return.

BARRIERS TO ADDITIONAL WORKFORCE AT NO EXTRA COST

To more than one listener/reader this may sound like some overgrown dotcom wisdom. The principle may be correct, but it will not come as a surprise that there can exist a certain discrepancy between the current research practice and the principles that will characterise the future international digital research environment.

Instead of considering data sharing as getting free additional help in getting the intended scientific work done, initial investors in digital data are often suspicious of unfair competition and free riding. Researchers looking for existing data sources from colleagues are not always welcomed heartily. Researchers, officials and managers mention financial, legal, and organisational barriers, cultural and ethical problems that complicate the full realisation the potential of digital data resources.
AWARENESS RAISING, AGENDA SETTING AND IMPLEMENTATION

From a distance it may seem simple to take general regulatory measures to insure an optimum benefit from public investments in digital research data: open up access, the sooner the better. To succeed however, requires a much closer look at the research shop floor. Broadening access to and sharing of digital research data can bring about substantial changes for the people involved in the specific research practice. A careful management of change to raise the awareness of the importance of data sharing, to discuss the issues thoroughly and to implement measures in an orderly manner is needed to underpin the ”simple” policy solutions aimed at. That’s where the insight from the specific case studies presented here will contribute to general science policy and research management on a more general level.

TIGHTLY ORGANISED MULTINATIONALS

Astronomical costs forced the successful international co-operation in the large scale scientific multinational European Organisation for Nuclear Research (CERN). Co-operative research practice at CERN has been the key factor in making standard data sets available for international research. Economies of scale in molecular biology were the prime reason for setting up the European Molecular Biology Laboratory (EMBL). As data management and data dissemination became of primary importance in for instance genomic research, the semi autonomous European Bioinformatics Institute (EBI) was established. In these cases international Data Access Regimes have followed the international research practices.

Now digitisation is becoming such a strong force in science this will also work the other way around: new governmental science policies on data access and new data policies of research funding organisations will stimulate the development of new international research practices.

In both ways the development international Data Access Regimes will open up exciting new dimensions of international scientific co-operation.
EMERGING GLOBAL NETWORKS

The case of the functional Magnetic Resonance Imaging research shows the somewhat hesitating start of an emerging scientific endeavour that (mark my words!) possibly will end up as a virtual global laboratory accommodating a ‘Human Brain Project’ that will surpass the Human Genome Project in scale and scope. Additional policy support certainly will help the larger project to take off. The Global Biodiversity Information Facility which services a rather dispersed community of scientists from research and museums, has its origins in the Global Science Forum of the intergovernmental OECD. It would be hard to imagine how this federative arrangement could have succeeded without such explicit central government support.

PULLING AND PUSHING TO REALISE DATA ACCESS

The establishment of successful Data Access Regimes for research data from public funding relies on an appropriate mix of incentives in policy making and research management conducive to additional advisory support, optimum funding conditions and investment in durable infra-structural facilities. The cases presented here illustrate the real life policy demand for differential treatment of ‘sticks’ and ‘carrots’, administrated ‘top-down’ as well as ‘bottom-up’.
More and more fields in science, social science and the humanities are becoming "data rich".
Big science data policies\footnote{Part of this work was funded by the Dutch Ministry of Education, Culture, and Sciences (OC&W). The authors wish to thank Dieter Schlatter, Janet Thornton, and Graham Cameron for their cooperation in this research and their comments on earlier versions. We appreciated the thoughtful advice from our colleague Martina Merz (EPFL) during our visit to CERN. We benefited greatly from comments on earlier drafts by Peter Schröder, Kathleen Casey, Anne Beaulieu and the other members of Nerdi. We also thank the participants in the 4th Global Research Village Conference as well as the experts of the OECD Working Group for their ideas and suggestions.}

Paul Wouters and Colin Reddy

Networked Research and Digital Information (Nerdi)
NIWI-KNAW, Joan Muyskenweg 25, PO Box 95110, 1090 HC Amsterdam,
The Netherlands
Email: paul.wouters@niwi.knaw.nl www: http://www.niwi.knaw.nl/nerdi

INTRODUCTION

Research question

The increased use of information and communication technologies in research and the development of novel measurement, simulation and imaging technologies has created a true data flood in science (Scherer & Harhoff, 2000; National Research Council, 1997; National Institutes of Health, 1999). More and more fields in science, social science and the humanities are becoming ‘data rich’. This is especially true for the experimental and observational sciences and for engineering research (Wouters & Schröder, 2000).

The increasing role of huge data sets in scientific research has important implications for the way the research is conducted, for the way it should be organized and funded, and for the training of new researchers. For example, the application of digital observation tools in astronomy has led to a deluge of data across all wavelengths (Sincell, 1999). Data production in these fields doubles every year (Szalay & Gray, 2001). The astronomy community is building a Virtual Observatory, an organization of the data into a coherent whole that can be accessed by anyone from anywhere on the planet. This is spurred by the promise of vastly increased knowledge discovery by the emergence of querying databases as a new research mode. This development can be seen across a wide variety of scientific specialties. Global data networks have become decisive in the further development of
seismology (Romanowicz & Giardini, 2001). The quality of data is no longer seen as the limiting factor in this field, but rather the spatial resolution, centralized archiving, and the continuity in time of the archives. In molecular biology, the deciphering of the human genome would not have been possible without providing immediate access to the new data generated in the decoding race (Marshall, 2001). The increased availability of this data, speeds up the cycle of knowledge production in a wide variety of fields. This varies from the approach to diseases (Peltonen & McKusick, 2001), plant biology (Bennetzen, 2002), and proteomics (Fields, 2001) to new research fronts in food and nutritional sciences (Desiere, German, Watzke, Pfeifer, & Saguy, 2001). The coupling of data which had hitherto been separate introduces new lines of thinking and discovery in interdisciplinary domains, e.g. biodiversity (Bisby, 2000), water scarcity (Brown, 2002a), new ways of making online topographic maps (Brown, 2002b) and the scientific use of monitoring systems created in the framework of the Test Ban Treaty (Stone, 2002). The new imaging technologies have led to a surge of new research in the neurosciences based on the availability of large-scale complex datasets (NET NEWS: Neuroscientists Link Up to Study Brain Diseases, 2001). Indeed a key element in the future of the neurosciences is the integration of the neurosciences with information science (Koslow & Huerta, 1997; Roland, 2000; Wong & Koslow, 2001).

These turbulent developments pose formidable challenges to the present practices of researchers, to the infrastructure of science, and to the way science is being regulated and funded. In the field of genomics, there has been a fierce debate about the way data should be made available, the relationship between data deposition and scientific publication, and the balance between public and private science (Marshall, 2002a; Marshall, 2001; Marshall, 2002b; Roberts, 2002). This debate has now resulted in rules that stipulate that data should be made available freely (Marshall, 2003). Rules are not enough, however. Practices have to be changed as well. In some sciences nothing less than a paradigm shift is at stake, e.g. in food science (Desiere et al., 2001). The way scientists are being evaluated and their careers are shaped is also at stake. For example, researchers who have spent years on building new databases such as the Sloan Digital Sky Survey in astronomy, have effectively put their scientific careers on hold even though these databases are critical for the future development of the field (Finkbeiner, 2001). The confluence of information sciences with the specific specialties of the user, has led to new demand of technical specialists in fields like bioinformatics and neuroinformatics (Global Science Forum’s Neuroinformatics Working Group, 2002; Roos, 2001). In these areas, the increasing interdependence of datasets is creating new challenges in statistics and forecasting (Sullivan, Timmermann, & White, 2002) on the basis of the available scientific data.
Increasingly, the availability of data is a crucial factor in the further development of these sciences. In the neurosciences, the estimate has been made that most of the data collected is effectively unavailable to the scientific community (Koslow, 2002). Scientific collaboration is more and more depending on effective access to and sharing of research data (Teasley & Wolinsky, 2001). This drives programmes to create new global scientific infrastructures, such as the Grid (Chien, Foster, & Goddette, 2002), e-science networks (Hey & Trefethen, 2002) and the Semantic Web (Hendler, 2003).

The creators of these networks are faced by the challenge of making the connection between the potential of new digital networks and networked tools on the one hand, and present scientific practices in a wide variety of fields on the other hand. It is therefore relevant to learn from the experience of the small number of scientific fields that have been dealing with large amounts of data for decades. Maybe the foremost of these is the field of high energy physics, in which researchers try to probe the essence of physical matter by smashing beams of particles near the speed of light into other particles. These collisions create enormous amounts of data which need automated processing before they can be interpreted. Without ICTs and advanced, high-performance computing this type of research would simply be unthinkable. A more recent development is the transformation of molecular biology from a data poor experimental science into an information science (Lenoir, 1998). Mainly as a consequence of instrumental breakthroughs in the manipulation of DNA, proteins, and RNA, automated sequencing became possible at the end of the 20th century (Kay, 2000). The working environment of the molecular biologist changed radically as a consequence of these technological innovations. The emergence of new standards in equipment and laboratory materials, and an increased role of commercial products in the laboratory also influenced daily life in the laboratory. Experiments that formerly had taken months, could now be conducted in a few days.

It seems timely, therefore, to study the way in which researchers have been dealing with large amounts of data, both as the input into and outcome of research, in these two fields. Such is the rationale behind this study. Our question is:

*What policies and practices have been developed for the use of data, both in-house and external data to facilitate access to research data and data sharing in the fields of high energy physics and molecular biology?*
Choice of the two case studies and research method

To understand current issues concerning access to digital research data, it is necessary to take the institutional organisation of modern scientific research into account. Rather than trying to organise a survey of individual researchers in these fields, it seems, therefore, more appropriate to study the way that the research institutions have organised their practices and policies. We have thus decided to conduct a case study of two central research institutions in the fields of high energy physics and molecular biology: CERN, the European central high energy physics laboratory in Geneva, Switzerland; and the European Bio-informatics Institute (EBI), at Hinxton, Cambridge, UK. Both are key institutions in their fields and often set new standards in the way that data are being produced, handled and processed as well as in the development of software and hardware to do this. At times, these initiatives have had wide ranging implications, such as the development of the hyperlink structures and internet protocols at CERN (Berners-Lee, 1999).

We conducted site interviews with key officials responsible for the institutions and their data policies. The interviews were complemented by the study of relevant documents and policy regulations obtained from the institutes themselves. We also conducted a content analysis of the Web sites of CERN and the EBI focussing on data policies.

BACKGROUND OF THE TWO INSTITUTIONS

Background of the EBI

The European Bioinformatics Institute is a non-profit academic organisation that emerged out of the European Molecular Biology Laboratory (EMBL). The EMBL was created as a multinational top research institute in Heidelberg, Germany, with sister institutes in Hamburg (Germany), Grenoble (France), and Monterotando (Italy). It functions as an international network of research institutes funded by contributions from fifteen countries and dedicated to research in molecular biology. Its mission is to perform outstanding research and thereby strengthen the discipline of molecular biology in Europe. This field is characterised by fierce competition between research groups at universities, research institutes and private companies. The EMBL does not have a monopoly on research in this field.
The EBI is the data service centre of the EMBL. Its roots lie in the EMBL Nucleotide Sequence Data Library, which was established in 1980 at the EMBL laboratories in Heidelberg, Germany and was the world’s first nucleotide sequence database. The original goal was to establish a central computer database of DNA sequences, rather than have scientists submit sequences to journals. What began as a modest task of abstracting information from literature soon became a major database activity with direct electronic submissions of data and the need for highly skilled informatics staff. The task grew in scale with the start of the genome projects, and grew in visibility as the data became relevant to research in the commercial sector.

It soon became apparent that the EMBL Nucleotide Sequence Data Library needed better financial security to ensure its long-term viability and to cope with the sheer scale of the task. There was also a need for research and development to provide services, to collaborate with global partners to support the project, and to provide assistance to industry. In short, the management of exploding amounts of research data in molecular biology became too big for the EMBL itself. Therefore, the EMBL Council voted in 1992 to establish the European Bioinformatics Institute and to locate it at the Wellcome Trust Genome Campus in Hinxton, the United Kingdom, where it would be in close proximity to the major sequencing efforts at the Sanger Centre and HGMP Resource Centre. In September 1995, the EBI occupied its current location on the Wellcome Trust Genome Campus.

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 accessible freely to all facets 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.

---

2 http://www.ebi.ac.uk/Information/About_EBI/about_esi.html
Background of CERN

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’ (De Solla Price, 1986). 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. Experiments at CERN usually involve a large part of the research community. When an experiment is split up and built by a large number of institutes, spread across Europe and other continents, efficient co-ordination is essential. One example is the ATLAS experiment, which is one of the big experiments under construction for the Large Hadron Collider (LHC). A mass of particle detectors will be packed into an assembly that weighs 7000 tonnes in all. The experiment is being constructed by 1700 collaborators in 144 institutes in thirty-three countries.

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.’ CERN is currently engaged in the realization of a new accelerator, the Large Hadron Collider (LHC), where high-intensity proton beams will collide head-on at extremely high levels of energy. The extreme conditions will according to the researchers involved give new research possibilities ‘to test predicted but as yet unobserved phenomena and to search for the unknown’.

The case for the existence of CERN is succinctly put in their own words: ‘Even if some of the larger European countries could afford to construct and run a laboratory as big as CERN, the human resources to run it could not be found within a single nation. For smaller countries especially, the problem in science is to sustain research activities in as many fields as possible. For economic reasons, countries have to maintain the level of insight needed to understand and exploit new ideas and discoveries leading to new applications. Unfortunately, smaller countries spend a smaller frac-
tion of their gross national product on fundamental research than larger countries, and the only solution to maintaining a high level of research in many fields is by international collaboration. Particle physics is relatively expensive but, thanks to CERN, most European countries have the chance to take an active part in the world’s most advanced experiments, without consuming an unaccept-
able fraction of the national science budget."

**PRESENT DATA SHARING REGULATIONS AND PRACTICES**

Present data sharing regulations and practices at the EBI

**Founding statement** 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. Article II paragraph 1 of the Agreement states: ‘The results of the experiments and theoretical work of the Laboratory shall be published or otherwise made generally available’. Article II paragraph 4 goes on: ‘Co-operation shall include in particular the promotion of contacts between, and the interchange of, scientists and the dissemination of information. So far as is consistent with its aims the Laboratory shall also seek to co-operate, to the fullest extent, with research institutions by collaborating with and advising them. The Laboratory should avoid duplication of work carried out in the said institutions’. ‘The 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 accessible freely to all facets of the scientific community in ways that promote scientific progress’.

**Consequences of being publicly funded** The EBI management puts its publicly funded status central, and the responsibility that it places on the organisation. Graham Cameron, associate director of the EBI: ‘It’s core to our principles that we’re a publicly funded organisation to preserve and make available the public record of public science’. This principle lies at the heart of the institute’s operations. The EBI manages a number of databases constructed from submissions of data intended to be in the public domain. Cameron: ‘For DNA sequence data or structure data, we are effectively

---

5 [http://public.web.cern.ch/Public/CERNOP/League2.html](http://public.web.cern.ch/Public/CERNOP/League2.html)

6 Interview with Graham Cameron and Janet Thornton, 30 April 2002, Hinxton, Cambridge, UK.
organisers and custodians of the data. So scientists throughout the world send us the data. We impose order on those data and redistribute them.’

Open access policy  As a consequence, the EBI has a policy of open access to the data. Cameron: ‘We impose no restrictions whatsoever on who can use our resources’. The EBI makes available a range of databases. How these are made available varies by topic area. ‘It differs from database to database in the sense that some of them are more like clearing house things, while some are very high flow rate databases. Some of them are complicated but low flow rate databases. Therefore, we use almost any method you can think of for making data available.’ The EBI deals almost exclusively with public data that are not covered under any IP encumbrance. These databases are available for a complete download. The EBI supports this by making Web tools available. Cameron: ‘We offer a range of web services layered on top of the databases, which are tools to enable you to search those databases. Depending on the nature of the database we offer a whole range of methods of keeping remote copies of databases in sync with the main database.’

For example, the EBI receives support from the Wellcome Trust Foundation for two projects:

– Ensembl, which aims to provide the ‘reference view of genome sequence data’ as a freely available resource for scientists and the public. A total of ca. 12 million Euro has been awarded for this project, over five years, equally shared between EBI and the Sanger Institute.

– European Macromolecular Structure Database, which has five goals including to integrate data from 3D-EM with those from X-Ray and NMR techniques. A total of 4.5 million Euro has been awarded over 4 years.

Data publication policy  The scientists submitting the data are made aware that their submissions will form part of the publicly available databases that the EBI manages. For example, section 6 of the EMBL nucleotide sequence database submission form, states: ‘Your data can be made public immediately, or they can be kept confidential until a release date which you provide. Confidential data are ALWAYS made available to the public after publication’. Thus, the main rule 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 however limited, after which publication follows.
Recently, there was an important debate within the scientific community in molecular biology as to whether scientists submitting data to repositories, such as the EBI, should have the right to impose restrictions on the publication and use of the data. Cameron: ‘We were about to adopt a policy which allowed submitters to impose certain kinds of restriction-usage on data. This was something which we had never wanted to do. However, a long debate involving our international collaborators and (particularly US) funding agencies had led us to conclude that this was a necessary evil. If we insisted on no restrictions, some crucial data would not be submitted, and the public collection would be incomplete. We had a meeting of the International Advisory Committee for Nucleotide Sequence Databases. They gave us a very strong recommendation that we should NOT accept such restrictions, and, on that basis, we never implemented the policy; – it was reversed. Pleasingly, faced with no alternative, the main submitters about whom we had been concerned gave us their data without restriction.’

The EBI policy is now clear. Cameron: ‘We accept no restrictions on the use of publicly visible data. However we 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)’.

**Adding value through data manipulation**  
The EBI adds value to the data by manipulating and analysing it, and compiling new databases from the results. Cameron: ‘We have a protein sequence database in collaboration with the Swiss Institute of Bioinformatics. When you have proteins, you would like to know what they do in organisms. The way you find that out is by rather complex computer analysis in which you compare your data with data in other databases. So we get protein sequences either from DNA translations or from direct submissions. But we do a lot of analysis of the data. Although the data comes from the community, we create a lot of the value in this database by these analyses. In some sense we own it in a way that we don’t own some other databases’. There are more than 20 mainstream database productions available from EBI. 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.
**Standards**  
The standards in use are the results of negotiations with equivalent bodies in the US and Japan. In a number of instances, the standards have been developed by the EBI, notably if the resources were built at EBI. In other cases the standards have been the result of discussions within the relevant communities and the EBI. The EBI has developed a routine set of meetings with Japanese and US partners to keep developing the standards in the subject of DNA sequence data. This has been going on for the last twenty years. In other areas the EBI conforms to standards that have been created elsewhere. Cameron: ‘I think we are, in the bio-molecular area, actually a pretty major player in standardisation in the world in a way even more than anyone in the US.’ The EBI has received significant European funding to work on standardisation in collaboration with biotechnology and pharmaceutical industry.

**Tools**  
The tools which the EBI applies and offers are partly developed in house, and partly produced by the global bioinformatics community. The EBI web site gives access to a wide variety of them which enable different types of analysis of the 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 and integrated them into a more or less seamless system which facilitates their use. Cameron: ‘I would not claim that we are doing things that no one else does, but rather that we have created a system that enables one to systematically present and analyse the data.’

**Collaboration with industry**  
Despite the strong public ethos the EBI has, this does not prevent the institute from collaborating with industry. The institute has developed an industry programme which claims to be ‘consistent with the public domain policy of the EBI’7. The EBI management is aware, however, that this cooperation may lead to conflicts as a result of their open access policy. Cameron: ‘There is indeed pressure to a conflict in the sense that the commercial players would love to have proprietary information and us to do proprietary research for them. Because we view ourselves as a public organisation we don’t do that. The companies help us setting priorities but we are not performing proprietary research for them. In fact most of the resources we obtain via our industry programme are spent on doing trainings for industry. We are also an important centre for training in bioinformatics.’

---

7 See http://www.ebi.ac.uk/industry/index.html
Commercial exploitation and intellectual property rights

There is no commercial exploitation of the data that the EBI produces, mainly because the data it deals in has been supplied by scientists on the basis that it is going to be made publicly available. Cameron: ‘It’s not that we’re against commercial exploitation. The information is used by hundreds of companies throughout the world. But we view ourselves as a public information resource.’ This does not mean, however, that commercialisation is not an issue at all. There have been several instances where software developed by the EBI has been spun out and commercialised. With regards to the basic information, however, the EBI wishes to stress the non-commercial character of the data processing and archiving. The EBI argues for this with reference to the need to maintain relationships of trust with the scientists. Cameron: ‘In a sense our whole activity depends on the mutual goodwill between the scientists and us. If the scientists thought that we were getting rich off the back of their information then the exchange of data would break down and they would be asking: where’s my cut of the money.’

Software life cycle

The software that researchers need to be able 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 plainly unthinkable to try to read the data without advanced software tools. The EBI has worked together with providers of software tools. Cameron: ‘We’ve worked together with global collaborators to expose the data in a well organised and defined form. And the software for processing those data has been developed by third parties.’ This means that researchers get the data from EBI but the necessary software usually from a third party, either a commercial for-profit company or an academic, non-profit group. Commercial packages can be seen as a specific stage in a life cycle. Often advanced software for research is created within an academic or research context. After a period of beta testing and refinement, the burden to maintain the software and the necessary organization become too much for the research group that developed the basic idea and its implementation. This is the moment that the software is usually outsourced. As a consequence, the software starts to live a life of its own, distancing itself from its original research context. New issues in academic research will then possibly be the seeds for new software tools and the cycle may start anew. Cameron: ‘We have been around long enough that we have seen enough of those tools go through that life cycle. It’s not that the commercial sector is actually good at sustainable high quality scientific software, although this is the present norm for the software that works with this information.’
Whether this type of cyclical evolution will be the case in the future, where radically new ways of designing software may be expected, remains to be seen. However, Janet Thornton, the Director of the EBI, points to an additional dilemma: ‘most of the stuff that we deal with are other people’s data. It’s not our data. This is a big difference. The IP that we would generate would really 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. I think the problem for us is that we have to make sure that our services remain in the public domain if we expect to get public money’.

**Ethical issues**

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 vigorous commercial activities for ethical reasons. EBI director Thornton: ‘I think it’s also dangerous, but if you do go in a commercial route, when it becomes competitive and there is a buy-out situation and X takes over the world and then everybody is dependent on X. I certainly wouldn’t be happy with having the human genome owned by a private company’. Concerns about the possibility to use biological data to produce biological or chemical weapons have become more pressing due to the international developments in the wake of September 11, 2001.

**Present data sharing practices and regulations at CERN**

**Founding statement**

CERN’s commitment to making research data publicly available is laid down in its founding convention. Article II paragraph 1 of the convention states: ‘The Organization shall provide for collaboration among European states in nuclear research of a pure scientific and fundamental character, and in research essentially related thereto. The Organization shall have no concern with work for military requirements and the results of its experimental and theoretical work shall be published or otherwise made generally available.’

**Role of experimental collaborations**

Experiments at CERN are run by teams of ‘collaborations’. These collaborations contain a number of scientists from different countries. Some collaborations, particularly LHC ones, have more than one thousand members. Data sharing policy within a collaboration is unproblematic, as 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 else in an unprocessed form. They are not publicly available.

**Types and levels of data** According to Dieter Schlatter, Division Leader of the Experimental Physics Division at CERN, an important issue is the level at which the data is being made available as well as the type of data that is involved. Schlatter: ‘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, the data itself does not make sense except for the expert within that particular experiment. Therefore, the data is usually password protected at this stage. This has partly to do with the size of the data, but partly also with competition among researchers. Most researchers will not give their competitors direct access to this raw data. The raw data resulting from the experiments is therefore extracted on the basis of the specific research question and cleaned up. This results in a Data Summary table which is what most physicists use.’ This data summary table is made accessible worldwide in a password protected way, because it is still specific to the collaboration involved. It will be clear that the mass of data that is being shared among researchers within a collaboration is huge and increasing. This is the main rationale behind present experiments and design of more advanced computing networks like the GRID. At the next level, the data are further filtered to facilitate storage on CD, DVD or hard disk. ‘In this way, the larger community of physicists has access to data from collaborations in which they do not participate’, according to Schlatter. 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.

**External use of data** The CERN management is somewhat sceptical with respect to the relevance of making their data available to the public at large. Schlatter: ‘Some attempts have been made to say this is public data, why not let the general public participate in it. There is clearly nothing against this from the general point of view. The question is more: is this useful in the sense that you increase knowledge transfer or would you rather increase confusion’. As a consequence, most experiments have not made their data available to the general public.

---

8 Interview with Dieter Schlatter, 3 May 2002, CERN, Geneva, Switzerland.
Outside of the scientific community the CERN data are not distributed. The exception are data sets for educational purposes but these have been tailored and filtered. ‘We took out all the confusing stuff. It’s mostly used in universities for education and this data is available on the web. It is organized by physicists and makes it possible for people to repeat a measurement using certain data sets. It’s original data but put in such a way that you can get some meaningful result. You can’t participate in the research in the sense that you will discover something.’

**Standards and databases**

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 develop in relation to technological developments. Schlatter: ‘For example in the past everything was done using Fortran as the basic language and now everything is done using C++. This already affects the standard’. The type of database to use is another standards issue. In the past, relational databases were the main format, later object-oriented databases became en vogue. Presently, relational databases may have become more important again. As a consequence, both relational and object-oriented databases are in use. Schlatter: ‘You need both relational databases and object oriented databases for higher level data in which you also describe the context’. Strikingly enough, there is not one standard that can be implemented to transform data sets for different uses. Instead, there many different technical standards. According to Schlatter several attempts at reaching agreement on one universal standard have been undertaken but so far this turned out to be difficult. ‘There is not one universal standard. Of course, within a collaboration a single standard is used otherwise the cooperation would become too difficult. But researchers who want to read the data of other experiments would need the programmes from that experiment. They cannot use their own software to read the data. For the LHC experiments we have been trying to create one standard. My prediction is that we will have more than one standard, we will probably end up with two or three standards for the data. Presently, working groups of physicists are attempting to reach common definitions, which is difficult because every physicist thinks he knows best. But the tighter resources for research are also pushing us to shared and common data standards and formats.’

**Data sharing between collaborations**

The main drive for creating more unified and common standards is not so much data sharing as making the maintenance of the data easier. Schlatter: ‘If you have two standards, you need two software programmes to access and maintain the data. Sharing is not the main concern here because in practice sharing will mostly take place within the experi-
ments and within the collaborations. Data sharing between collaborations will always be 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. Schlatter: ‘During LEP, one experiment had a hint of a discovery and released certain data to other collaborations so they could examine it. It took weeks of discussions before everybody agreed on the meaning of certain concepts like “energy flow”. This was done, but with a small and well-defined sub set of data that has been shared. Pre selected at some level and so on and I think this will happen again. But you could never do this with the total data set and nobody wants to do it because it would mean that he has to repeat the whole work again. It is plain useless.’

This demonstrates one important factor to be borne in mind when considering the CERN practices for data sharing: the lack of potential usefulness in sharing data between collaborations. As collaborations have different standards and working practices, the possibility to meaningfully transfer raw data sets is limited if not very difficult. The majority of data exchanged is therefore the processed data tables.

**Collaboration outside CERN to publish data** Much of CERN’s processed data is shared within the scientific community in collaboration with our laboratories in the form of the *Particle Review Book* produced by the particle research group at Berkeley. This group records the properties of particles, as well as cross section measurements. The Particle Review Book is published every other year, both in print and electronically. Schlatter: ‘It is sort of the bible because it has not only the data but also standard interpretations and this is very, very popular. There are far more copies sold than there are particle physicists. The summarised version is distributed in forty thousand volumes while there are only around 8,000 particle physicists. The Particle Review Book is becoming more and more a text book. This is really special because it summarises the complete knowledge in the form of data and interpretation from particle physics and it goes also a little bit into astrophysics. I have not seen anything like this in other fields. This endeavour is funded by the US Department of Energy and half by CERN. It is a very good way of distributing knowledge.’
Funding schemes

The key issue in funding schemes relates to the infra-structural character of databases and data sharing environments and tools. This creates a tension between on the one hand the funding schemes of research (usually project oriented) and funding schemes for infrastructure (of which data requirements form an increasingly important part). This tension is a manifestation of the 'essential tension' (Kuhn, 1977) that has by definition been built into research funding: the tension between stability and flexibility. On the one hand, research funding must be able to respond to new, unexpected developments. Research needs, on the other hand, sustained funding to reach breakthroughs. This means that research funding must be able to play two different games at the same time: support short-term projects and a pool of researchers that can respond to new developments, and sustain long-term infrastructures and research institutions with the necessary critical mass and endurance.

At CERN, this tension has been solved by the large scope of the institution. Basically, CERN can be seen as a huge infrastructure that houses a huge, temporary experimental projects (see for a comparison between the institutes the Conclusions section). The EBI however does have a major problem. The key issue is that funding for data sharing infra-structures needs to be done 'on a rolling basis', according to EBI director Janet Thornton, whereas presently, funding for databases is often acquired on a research project basis.

This contradiction between the need for funding on a rolling basis and the practice of project funding is the main threat for EBI's commitment to the public availability of its data. ‘This area of science has advanced to where there is a kind of high throughput science that delivers a lot of its value as as electronic information resources’, according to Cameron. ‘We have grown up as part of the scientific infrastructure that is designed to satisfy, to be the vehicle for, that part of the electronic record of science, and it is fair to say that whilst we have received generous support from all sorts of clever directions, the life sciences haven’t faced up to the fact that there is now this new infra-structural requirement. Researchers are quite aware of the fact that if you want to do science you have to pay for a library which goes on for ever and ever, and they haven’t quite grasped the fact that we’re now in an era of science where there are comparable electronic resources. It’s an uphill battle.’
The EBI is part of the European Molecular Biology Laboratory, which is funded by the member states of the European Molecular Biology Organisation. Because the EBI budget is already relatively large in comparison to the total budget of the EMBL, the EBI management estimates that it can only increase its financial possibilities on the basis of an increase in the total budget of the EMBL. Presently, the institute is using its research functions to get funding for combined research and infra-structural projects, for example from the European Framework Programmes. Cameron: ‘Framework programmes are designed to create funding for specific research projects, not for ongoing parts of infrastructure. Yet, we have managed to dress up what we do and pick out the bits that could be cast as individual research projects to win some of that European Union funding, but it’s unsatisfactory. The issue will have to be faced up to, and we’re hoping that we can create the political will to face up to it before we have a complete disaster, rather than after. If we have a complete disaster and the EBI goes under, then I’m sure the political will will be created.’

**Information management and integration**

With the increase in the amount of data and the variety in data types, information management tasks in research are becoming more complex. At CERN this complexity seems to be partly solved by decentralisation: many issues are decided at the level where the data are being produced: the level of large-scale experiments and collaborations (see above). At the EBI, the variety of ways of data submission and the steep increase in the amount of data creates the challenge of data integration management in a somewhat different shape. Cameron: ‘The task of information management gets more and more daunting, and I suspect that we’re experiencing a set of problems that actually are ubiquitous in information management and are always threatening to drag you under. One of the biggest challenges of all is to traverse the information space, accessing all the information you want without taking shuddering leaps through hyperspace to completely different information resources.’

A related issue is the amount of data and the variety of data types that will have to be dealt with in the genomics research. According to Cameron, researchers tend to expect that most genomics data will be like reference genome data. ‘Actually I think it’s going to be profoundly different. There will be all sorts of experiments collecting little bits of information. So far, in the genome world we were not forced to think about how you cut your coat according to your cloth because we knew the shape
and size of the coat. It was the reference human genome. In fact there will be all sorts of information demands on us and we will actually have to take far more hard-nosed decisions in future about what’s worth doing and what’s not. So it’s going to be a different world. Huge quantities of information; which of it is actually worth archiving and how do you integrate it all? It’s going to be a different kind of task in some ways. This goes far beyond transferring genomics data from one medium to the other.. It is a very different game trying to decide on what sort of information architecture we will need in the future.’

**Criteria and standards for archiving**

At CERN, one of the major problems in information management is what to archive and in which standard. As Dieter Schlatter says: ‘There are a lot of discussions about the best policy in data archiving, particularly with the LEP data. It took an enormous amount of data on the $z$ resonance, and later in the high energies, and this will not be repeated very quickly, maybe never. So we discussed very seriously how to archive this data.’ In this debate different sections of the scientific community had different wishes and perspectives. ‘The theorists said maybe in ten years we have a new idea concerning some QCD calculation and we would like to check it with that data. So therefore we would like to still have access to the data. For this kind of thing you may just need only specific subsets of the data, you do not need to archive all of it. On the other hand, the collaborations want to archive it for their own purpose. When collaborations come to an end, at the universities there are still people doing PhDs or who may have a new idea and look into the data. For this, one does need more comprehensive data archiving, including the data that describe the context of the experimental results. This is necessary because also people who are not member of the particular collaboration need to be able to read it.’ 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. Current ideas at CERN to solve this problem include putting data, along with the appropriate software and a virtual computer operating system, onto DVDs, so that access could be achieved through a regular PC. Nevertheless, Schlatter doubts whether this type of archiving would last longer than twenty years.
In the preceding sections we have described the basic principles and organisation of access to and sharing of research data at two key institutions in ‘big science’: the EBI and CERN. We have also identified a number of key issues which will have to be confronted if one wishes to develop more advanced practices and policies with respect to the access to and sharing of research data. In this section, we will reflect upon these issues and sketch implications for research policies.

**Public funding obligations**

Both organisations cite that as they are publicly funded bodies, they have a formal responsibility to make what they do as freely available as possible. Indeed, both have commitments to this end in their founding documents, albeit not specifically focused on research data.

In the case of EBI (which is mainly a data centre with some supporting research and training responsibilities), this is the drive behind their open access policy. They see themselves very much as custodians of data, with a responsibility to maintain a public record of science. Thornton: ‘We have to make sure our services remain in the public domain if we expect to get public money.’ Hence schools and individual members of the public at large have the same access rights as the scientific community. 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 in a comprehensive way. The EBI takes into account that the present policy drive to secure more resources from non-public funding could compromise their activities in the public sphere. Cameron: ‘We don’t make our income by selling our services. If we did we would be very much richer than we are now. We serve science across the board from the most basic science up to commercialisation in world giant pharmaceutical companies. It is inescapable that if we were forced down the commercial path we would customise our services to the people with money. If you’ve got some graduate student in Leicester University calling you up and you’ve got Glaxo Smith Kline calling you up, you know who you’re going to respond to if you’re trying to respond to commercial pressures. We’re an electronic public library and we serve everyone. We have constantly to defend the right to a budget to do that, but we believe strongly enough in it that we’ll keep doing so.’
The research institute CERN takes a different approach. It 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 one experimental collaboration would not be useful to other collaborations. Summary tables are produced of the most useful, heavily processed data. CERN also collaborate with a group in Berkeley to contribute data to the Particle Physics Review Book which is published every two years. In addition, CERN makes some heavily processed data available on line to universities for educational purposes.

Thus, both institutions emphasise the basic principle that publicly funded research information should be publicly available. They differ, however, in the way this principle has been implemented. This can be related to crucial differences in:

1. the institutional structure of the field and the role of the institutions EBI and CERN;
2. the structure and role of the research data in the creation of scientific knowledge.

Institutional roles and the function of research data

As international database centre, the EBI receives submissions of data from all over the world; they deal with a very diffuse large number of individual institutions and individual scientists. EBI is to a large part a public repository of data. Much of what they hold, they do not own. They are responsible for recording submissions and manipulating and processing that submitted data into formats agreed within their scientific community. They add value to the data in this sense.

CERN’s ‘collaborations’, on the other hand produce vast quantities of data themselves. They decide the best way to handle and process the data rather than submitting it onwards. The way the data are collected and processed determines how easily the data can be made available, since groups outside the data producing collaborations would need to understand both the context in which the data was produced as well as having access to the collaborations algorithms for the processing software. There is no philosophical bar to the sharing of data according to the CERN management; for CERN it is more a question of which part of the massive amounts of raw data produced should be shared, and in which format.
Thus, the two institutions play quite different roles in their respective scientific communities. The EBI sees itself primarily as a custodian of research data (comparable to a classic data archive), as adding value to the data by error correction and putting the data in context (apart from its separate roles as research and training institute). The data are not produced at EBI itself but primarily in a distributed global network of research centres. Basically, the EBI is a central node in a cyclical flow of data through the international bio-molecular and bioinformatics research network. Because of the value adding character of depositing research data in the EBI databases, researchers do have incentives to make their data publicly available. There is also a more general public interest involved, since depositing the data in principle allows for more long-term and sustained error-correction procedures. This does not mean, however, that no possible conflicts of interests may occur here. Researchers try to prevent others to scoop them with their own data. Hence, they will have the tendency to delay making their data publicly available. This means that without formal regulations that hold for every researcher the tendency to withhold data from public circulation may prevail even if the data themselves would gain value by being shared. However, the EBI tries to resist such pressures, wherever possible, through discussions aimed at ensuring common policies with its equivalents in the US and Japan.

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, collaboration agreements, and making tacit knowledge as explicit as possible (which can never be complete though). 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 themselves are useless for other purposes; only the extraction of meaning from them based on specified research questions by the relevant experts makes sense. Hence, making ‘the data’ publicly available does in the eyes of the physicist not relate to the raw data but to data that have been contextualised, processed and refined. In this process most of the raw data are discarded. Thus, the difference with the bioinformatics data is that the particle data cannot travel so easily outside of their experimental context. They need interpretation as, for example, happens in the Particle Review Book. To the public at large, data are primarily made available in the form of educational materials and theories about the material universe. An important difference with the data in the
field of bioinformatics is that re-use of data is not very important for the development of the field. It seldom occurs that high energy physicists want to re-interpret old data. If it happens, they are usually the same researchers that were involved in the experiments. If new hypotheses need to be tested, the way forward is the organisation of a new experiment. This means that the production of new data seems to be far more important for the creation of knowledge than the re-use of already produced data.

Commercial exploitation of research data

The use of the data for commercial purposes at either institution is problematic, for different reasons. EBI does have strong commercial links, particularly with pharmaceutical industry, but this must not be seen to influence their core activities. Cameron: ‘There are instances where pieces of software that we have developed have spun out and been commercialised, but the basic information is pretty much in the public domain. Indeed, in a sense the whole activity depends on this goodwill between us and the scientists. If the scientists thought that we were getting rich off the back of their information then the whole exchange thing would break down.’ Hence the submitted data remains in the public domain, but the software EBI develops to process it may be spun out commercially. But even that could lead to problems. Thornton: ‘The IP that we would generate would really 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. I think the problem for us is that we have to make sure that our services remain in the public domain if we expect to get public money.’

At CERN, it is also true that some of the software developed in house to analyse the data may find a commercial home, but the data itself has a much more limited use outside its home scientific community. The software itself is also shared between the scientific community outside of CERN, namely Fermilab and Slac, who also contribute software to CERN. So inside the particle physics scientific community, sharing is the norm. Outside, there seems to be little audience for much of CERN’s data.

IPR issues are not important with respect to research data as such, given the fact that data do not fall under intellectual property rights regimes. Moreover, until now the new European Directive on the Legal Protection of Databases has not restricted access to the databases of CERN or the EBI. IPR issues do play a role, however, in those cases where embedded data are involved and also at the
principles may be especially important transformation of the science system.

interface between research data and software tools needed to be able to read those data (Wouters, 2002).

Funding

Due to these different data configurations in which the EBI and CERN are immersed, a number of policy issues are important in varying degrees. In a data producing research centre like CERN, funding is first and foremost related to the preparation of new experimental instruments and accelerators. Making data available is only a separate issue with respect to data archiving. This is different in the case of the EBI. The problems the EBI encounters with the present dominant funding schema in science is probably illustrative of many fields in which the amount of data has exploded in the last couple of decades (e.g. astronomy, brain imaging, neurosciences, environmental science, space science). The fundamental contradiction is that research is increasingly funded on a per project base in which research projects generally do not last longer than 4 to 5 years. However, making the data available, and keep them accessible, demands permanent infra-structures. This calls for a funding scheme on a more permanent (‘rolling’) basis and also needs different assessment criteria. The process of the accountability of researchers needs, in other words, also to be modified. Thus, the access to and sharing of research data will in many fields, though not in all, ask for funding schemes that focus on permanent funding of new institutional arrangements in which data depositories, processing centres, and archives play central roles.

To sum up, the management of large scale data sets in research calls for a long-term, sustained type of funding. At CERN, this is managed by virtue of data being an integral part of the regular research activities. EBI is an example of an international data depository cum service centre that houses contributions from many different sources dispersed from all over the world without a basis in an international treaty like CERN. Diffuse multinational project funding is not conducive to the maintenance as long as the project budgets do not include data overhead cost for data archiving and additional services.
The essential tension

This is not only a matter of funding schemes. It is the consequence of the ‘essential tension’ between data management infrastructure and (mostly project oriented) research we discussed earlier in this report. Scientific research that depends on the handling of increasing amounts of digital data requires explicit formal rules for data management (covering the generation, processing, distribution, analysis, and selective and sustainable archiving of data). This requirement should be a crucial element of science policy at the international, the national as well as the institutional level. CERN is an example of a research institution with a very well integrated data management based on international science policy. It has integrated complex data management seamlessly in its daily operations. The establishment of EBI as a separate data service division of EMBL is another way of recognising the growing autonomous role of data in life science research. While data generated at CERN are used almost exclusively by CERN affiliated researchers and their institutions, EBI serves a research community that goes way beyond EMBL. While data from CERN have no direct commercial value, the EBI databases are also used by industry. The tensions between data management and research within EMBO are also partly due to the novelty of large scale data sets in the life sciences: the level of integration of data management and research management in the life sciences is considerably less than in high energy physics.

Re-use of data

The extent to which data must be made available for re-use varies by discipline. The role of data and data types also varies by scientific discipline. There are especially important differences in this ‘data structure’ between observational sciences (e.g. astronomy), experimental science (e.g. physics), historical research and humanities, and research on human subjects (e.g. medical research). With respect to this case study, high energy physics has a different ‘data structure’ than the life sciences. The particular data structure of a scientific field has to be taken into account in data management policies. In high energy physics, data are rarely, if ever, re-used and processed data usually are not used outside of the field. This explains the fact that the data produced at CERN are in fact not publicly available. In the life sciences, on the contrary, re-use of data (including error-correction) can be an important source of new knowledge for a long time. There is moreover also an economic market for processed data on genes and proteins. To sum up, when implementing the general
principle of public access to research data in specific situations, many different aspects of varying data demand must be taken into account.

**Incentives for making research data available**

Formal regulations to make research data publicly available, however important they are (Wouters, 2002) will need to be supplemented by proper incentives for researchers to make their data available. At CERN this is not a separate issue because the data do not travel outside of the context of their production. EBI as data depository provides incentives to researchers on the basis of the data structure of genomics research. Researchers have an interest in depositing their data at EBI because the data will become more valuable if they are part of a larger set of data. The error correction procedures and the data processing at the EBI further enhance this type of incentive. Nevertheless, the experience of the EBI management does indicate that researchers still need to be convinced of the value of data depositing. Apparently, sharing data does not come as natural to scientists as the scientific ethos (Merton, 1973) assumes.

**The new role of digital data repositories**

Traditionally, libraries, collections and paper archives were crucial centres of scientific information, data and objects. The ‘informational turn’ in science (Nerdi (Networked Research and Digital Information), 2002) 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 (bio-informatics). The present international research infrastructure is in the process of re-alignment in response to these developments. With respect to this case study, the uncertain future of the EBI results from this as yet incomplete institutional reorganisation. Since the newly emerging data infrastructure will be international of scope, science policy principles may be especially important in the structuring of this transformation of the science system.
<table>
<thead>
<tr>
<th>Reference</th>
<th>Author(s)</th>
<th>Title</th>
<th>Journal</th>
<th>Volume</th>
<th>Issue</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NET NEWS: Neuroscientists Link Up to Study Brain Diseases.</td>
<td>(2001)</td>
<td>Science</td>
<td>294</td>
<td>(5543),</td>
<td>751d</td>
</tr>
<tr>
<td>2</td>
<td>Bennetzen, J.</td>
<td>THE RICE GENOME: Opening the Door to Comparative Plant Biology.</td>
<td>Science</td>
<td>296</td>
<td>(5565),</td>
<td>60-63</td>
</tr>
<tr>
<td>4</td>
<td>Bisby, F. A.</td>
<td>The Quiet Revolution: Biodiversity Informatics and the Internet.</td>
<td>Science</td>
<td>289</td>
<td>(5488),</td>
<td>2309-2312</td>
</tr>
<tr>
<td>5</td>
<td>Brown, K.</td>
<td>Environmental data: Water Scarcity: Forecasting the Future With Spotty Data.</td>
<td>Science</td>
<td>297</td>
<td>(5583),</td>
<td>926-927</td>
</tr>
<tr>
<td>6</td>
<td>Brown, K.</td>
<td>DIGITAL INFORMATION SYSTEMS: Mapping the Future.</td>
<td>Science</td>
<td>298</td>
<td>(5600),</td>
<td>1874-1875</td>
</tr>
<tr>
<td>7</td>
<td>Chien, A., Foster, I., &amp; Goddette, D.</td>
<td>Grid technologies empowering drug discovery.</td>
<td>Drug Discovery Today</td>
<td>7(20),</td>
<td></td>
<td>5176-5180</td>
</tr>
<tr>
<td>8</td>
<td>De Solla Price, D. J.</td>
<td>Little science, big science... and beyond.</td>
<td>New York: Columbia University Press</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Fields, S.</td>
<td>PROTEOMICS: Proteomics in Genomeland.</td>
<td>Science</td>
<td>291</td>
<td>(5507),</td>
<td>1221-1224</td>
</tr>
<tr>
<td>11</td>
<td>Finkbeiner, A.</td>
<td>'Invisible' Astronomers Give Their All to the Sloan.</td>
<td>Science</td>
<td>292</td>
<td>(5521),</td>
<td>1472-1475</td>
</tr>
<tr>
<td>14</td>
<td>Hey, T., &amp; Trefethen, A. E.</td>
<td>The UK e-Science Core Programme and the Grid.</td>
<td>Future Generation Computer Systems</td>
<td>18(8),</td>
<td></td>
<td>1017-1031</td>
</tr>
</tbody>
</table>


1 INTRODUCTION

1.1 Background to the study

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.

The focus of this report is to delineate the benefits and aims of making biodiversity data electronically available to the public. It will also outline primary concerns about developing the infrastructure required to make useful access to incredible amounts of data distributed across institutions around the globe. In this report, I will draw from research on issues of access within the biodiversity community in general and GBIF in particular. This research entailed interviews of scientists, museum and herbaria directors, program directors, and software developers, participant observation at a GBIF Database Access and Database Interoperability meeting, as well as extensive literature reviews of materials on GBIF, biodiversity informatics and the economics of biological resources, articles available on-line through a multitude of biodiversity organisations, and other academic articles on biodiversity.

---

1 The author would like to thank Geoffrey Bowker for his guidance in the conduct of this study, as well as all the interviewees.

2 For a list of institutions with publicly accessible databases and networks, as well as information on legal and policy initiatives that concern biodiversity and biodiversity data, see http://www.biosis.org/zrdocs/zoolinfo/biodiv.htm
1.2 Framework of analysis

In this report, infrastructure is recognized not only as a technology for moving information, but as a multi-layered, complex, socio-technical network that requires constant upkeep, support from policy and legal arrangements, and appropriate funding and career structures. The case of GBIF demonstrates a central principle of the social informatics perspective; namely, as Bowker and Star assert, ‘In order to develop robust long term information infrastructures, we must combine technical developments and organizational innovation.’ GBIF shows such combinations and their relevance to promoting data access. For example, GBIF is a clear case where the scientific aims of working across data sets from many different data sources will be unrealized if both information technologies and the policies of different nations are not made interoperable. An appreciation of these interconnections will then help us focus on a set of interdependencies so that we can draw out what needs to be considered within different policy aims.

Another perspective brought to the issue of data access is how infrastructure inevitably undergoes change but must be robust enough to maintain open access. The concerns articulated here point to how such infrastructure should be thought through as a means to facilitate good stewardship of public data, and the problems that might curtail such good stewardship. In an interview, Meredith Lane of GBIF clearly points to importance of the design process of infrastructure:

A properly implemented infrastructure goes invisible. The problem is you don’t want it to become invisible before you thought through some of the issues because what happens is if you have a properly invisible infrastructure then people forget to support it. And they will only recognize it is there if it breaks. The infrastructure needs to be invisible, but the development needs to be kept up in the limelight.

Lane considers a network as infrastructure, but also such communicative mechanisms in the biological sciences such as a catalog of species names. Like a properly connected set of databases, such a catalog enables scientists to communicate with one another.

---

4 Author interview, Meredith Lane, January 2, 2002.
The importance of infrastructure in this case arises in the words of scientists themselves who have committed to the belief that data about biodiversity should be considered and appropriately brought into publicly accessible networked environments because information about biodiversity: (1) has typically been acquired through public funds and is thus public; (2) is required by both scientists and non-scientists alike for research and decision-making on how the resources of the planet are used and; as such, (3) is necessary for assuring that scientific research on biological resources benefits society at large. In opposition to these values are issues pertaining to unclear ownership of data and the heavy use of intellectual property rights, the risks of data accessibility for sustainable biodiversity use and benefit-sharing, costs, and other concerns covered later in this report.

1.3 Report structure

This report will address basic issues of access to biodiversity information, including benefits, and then proceed through some of the institutional, financial, cultural and professional, and technological issues that arise in developing data access regimes. It will then present some of the basic features of a few institutions that have sought to create data access regimes, including GBIF. Finally, the report will conclude with some follow-up questions.

2 ACCESS TO BIODIVERSITY INFORMATION: BENEFITS TO SCIENCE AND TO SOCIETY

Biodiversity is ‘the sum of all the different kinds of organisms inhabiting a region,’ including the region of the ‘entire earth’ or some sub-region. Data about biodiversity refers to organisms in a variety of ways, from specimen names to molecular information. These data are housed in a variety of institutions, such as natural history museums, botanical gardens, and universities research centers. A relatively small number of countries have historically housed specimens data on the world’s biodiversity gathered through field expeditions, although efforts to correct this imbalance by ‘data

---

Andrew P. Dobson (1998), Conservation and Biodiversity, New York: Scientific American Library, p. 10
A tremendous amount of biodiversity data are not yet in digital form, although digitizing biodiversity information has been an aim of many institutions. This data comes in multiple forms: information about specimens, photographs, field notes, and actual collected specimens. By and large these objects are not easily accessible. The task to provide access through digital environments is tremendous: consider how many pieces of data point to an individual organism, how many organisms have been identified and collected, and the thousands of institutions across the globe that house such specimens and data. There is between 2 and 3 billions of these objects distributed across natural history museums. As one of the developers of GBIF states, the content is global, therefore the effort to make this content digitally accessible must be globally organized. This effort has been motivated by recognition of the range of benefits that come with making biodiversity data freely and digitally available to all.

### 2.1 Benefits of access to biodiversity data

Access to biodiversity information through a digital environment brings benefits to many different kinds of users, including scientists and government officials. These benefits include some of the following:

1. Helps heterogeneous users acquire data more efficiently;
2. Benefits institutions by providing greater visibility and use of their collections, including use that leads to the correction of errors in their data;
3. Improves the ability of scientists to explore new research areas, including research that requires querying data from different institutions, sometimes simultaneously;
4. Alters the nature of the data by making it ‘dynamic,’ allowing it to be brought into larger data sets of comparable and different data. The new opportunities to work with digital data promotes development of software tools that will allow further processing and use of the data, such as geo-referencing software tools;

---

6 Data repatriation is the process of giving data to a country from which the specimen to which that data refers has been collected. While some countries have proposed the return of actual specimens, and thus the data about them, the repatriation of data is significantly less expensive and less risky to the specimen collections, which can get damaged in the move.
Provides 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;

Helps to further the repatriation of data to countries which have specimen collections and information housed outside their borders;

Makes research more efficient and cost effective through the pooling and sharing of data resources.

In general, scientists believe that by making data accessible to the public, their research would contribute to the knowledge that can promote the sustainable use of biodiversity in the wake of threats to biodiversity. Indeed, Mexico’s Conabio includes this position in their mission statement:

> Currently the conservation of biological diversity and the sustainable use of its components is a priority, given the environmental crisis of the planet in recent decades. Knowledge of biodiversity becomes urgent in view of the rapid process of the loss of ecosystems, species and genes, as well as a broad spectrum of environmental services and products derived from plants and animals pending discovery or study.

**Instantaneous access**  For scientists, the ability to access data on biodiversity reduces the amount of work it takes to find information. For instance, if a scientist were 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, the museums receive more requests to send out physical specimens housed in their collection to the scientists who found information about those specimens through their on-line database. Take the following as an example of the change in access practices:

> Museums that already have public access to their specimen data have seen increased use of their collections and a dramatic decrease in staff time needed to fill information requests. For example, the Museum of Vertebrate Zoology web site fulfilled 41,937 specimen queries, representing 19,001,503 specimen records delivered, in the first year of providing public access to its data. This is up from 95 requests representing 160,471 specimen records delivered manually by its staff in the preceding year!7

7 'Impact of the Project' section of the MaNIS NSF proposal, see http://dlp.cs.berkeley.edu/manis/ProjectDescription.html
Institutional support  Natural history museums and other institutions that house biological specimens that make their data available to the public achieve greater visibility and the potential for increased use of their collections. Clearly, making accessible databases changes the nature of the work of the scientist as well as the curators at the museum, an issue discussed further below. At the institutional level, museum directors hope that circulating their data through accessible databases hooked into an interoperable network will allow them to ‘Have their data used, to have their institutional acronyms floating around the world, saying look they do have important stuff.’ In other words, by making their data accessible to persons outside of their institution, they might access the ‘lifeblood’ brought by the visibility of their data. With data visibility, institutions may prove worthy of more financial support.

More efficient and cost-effective scientific research  Ease of access to biodiversity data is reported to be a long-standing need for scientists that was made realizable with the inception of the Internet; according to Stan Blum, of the California Academy of Sciences, the dream of having the ability to query data sets from the scientist’s desktop was discussed as early as the 1970s. The Internet was thus recognized as an opportunity for scientists to make the acquisition of data for research easier. In addition to reducing the time it takes to find information about biodiversity, scientists have spoken of the kind of research that on-line queries makes available. By querying multiple datasets across natural history museum collections, scientists can compare data sets, and process the data through available and developing software tools to produce information about ecological systems, distributions of specimens, and so forth. As the software develops alongside the development of database access, scientists believe different types of questions can be asked of the data.

Improving data quality  Some scientists fear sharing imperfect data because of potentially negative effects on their reputation or possible exposure to legal liability. It has been found that incorrect data undergoes more rapid correction the more that data are searched and used by scientists. Both the host institution and scientists benefit from the clean-up that accompanies the use of data. Generally speaking, scientists have not objected to the discovery of errors because they would rather have the data than not. Explicit provisions within databases that users should direct the

8 Author interview, John Wiercozek, Museum of Vertebrate Zoology at the University of California, Berkeley, May 15, 2002.
9 Author interview, Stan Blum, May 7, 2002.
Curator’s attention to any errors help motivate corrections to data. Another issue of quality entails changes in taxonomy. Data may be outdated in relation to changes in taxonomy and are thus rendered less useful. Groups such as Taxonomic Digital Working Group (TDWG) are addressing issues of taxonomy as it pertains to creating networked information systems.

**Non-scientific access** The opening up of data through on-line facilities also serves non-scientific communities. The broad range of users that are drawn to data facilities shows the usefulness of biodiversity information and motivates the development of software tools that allow for heterogeneous uses and users alike. Conabio, for example, has found that its main users are government officials. Government agencies, non-government organisations, land use developers, and so forth, can turn to publicly accessible data networks for information about species in areas that are potential sites of development or for conservation projects. Here, proper use of the data becomes an important issue, an issue discussed below. Since it ‘extremely impractical’ to get sufficient amounts of data by going to separate institutions one-by-one, or when one must rely upon experts to acquire those data, the benefit of an accessible network of databases is equally valuable for non-scientists and scientists.

**Data repatriation** Access to biodiversity data also benefits nations that have data about their biodiversity housed in other countries. The repatriation of data to these nations is recognized as being of utmost importance by the Convention on Biological Diversity (see the most recent reports on the CBD meetings in The Hague, 7-19 April 2002 at www.biodiv.org). While some nations have made tremendous efforts to repatriate their data through the development of data networks and other facilities, much financial and technical support is required to develop the infrastructure necessary to place digitized data into a database, to access data available on-line in other countries, or to connect databases in other countries as a node in their own networks.

---

10 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
3 LEVERAGING DATA ACCESS

Although there are many lessons to be learned already from GBIF, the fact that it is not fully operational means that deriving full benefits from data access remains one of its unrealized goals. Leveraging electronic data access facilities entails integrating the technology, procuring funding, moving data from the ‘museum cabinet’ to a digital format, and training and employing support staff.

3.1 Costs of making data electronically accessible

Figuring costs  The costs for making biodiversity data electronically available to the public are generally unavailable. Scientists and museum directors have stated that it is difficult to divide the entire collection and research process according to dollars spent, and to figure the costs of moving data into digital form and network environments. However, individuals at institutions that have the appropriate hardware and software insist that the cost of uploading data into the database is fractional relative to the cost of collecting and cataloguing those data. For example, at the California Academy of Sciences, Stan Blum spoke of a recent field trip to collect specimens in Myanmar. The cost of this trip was roughly US $5000 for six weeks of work. The field researchers brought back 1000 specimens. It took about 2 or 3 days for a single staff person to catalog and digitize all the data that was collected from that trip, including field notes. The cost of the work to make the data available, consisting primarily of staff time and technology upkeep, was fractional compared to the field trip, but was able to be accomplished because the Academy already had necessary resources.11

One of the reasons it is difficult to figure the cost of moving data to an on-line database is that the data itself does not remain the same as it becomes digital. When data become part of an interoperable network of databases, it becomes part of larger data sets and drawn upon by a larger community of users. These data sets contribute to research that is different in kind from research that relies solely upon non-digitized biodiversity data. It is therefore difficult to quantify the cost of moving data into databases since it changes in quality and can be brought into geo-referencing systems, and other software programs that help create multi-layered depictions of ecological systems and other tools within the field of bio-informatics.

11 Author interview, Stan Blum, May 7, 2002.
The costs of Conabio  Conabio has invested MEX $9 million since 1993 in the data collections that it has brought together, the development of its network of databases housed outside of Conabio, software, quality control tools, and other technologies to create a system that has data on 6 million specimens. Given its short existence, relative to some hundred year old museum, Conabio would be one place to determine costs most clearly as its main efforts have been directed toward bringing together biodiversity data and making it freely available to the public. Still, the director of Conabio conveyed his sense that the calculation of the cost is difficult because Conabio does incite and draw upon research in the field and is thus part of the larger enterprise of biodiversity research. He provides the following explanation about costs:

In terms of the whole process of obtaining information, it’s a minute fraction. If you think of an average question that we answer here with an average of 3 databases, those 3 databases may come from 20 institutions. And each one of those institutions may have dedicated 50 years sometimes, 20 years, of PhD time, and the per diems to go to the field and obtain information and curate it and prepare information – all of that is a huge expense. ... For about 1 and 10 dollars per specimen you make that information accessible to a lot of people. And it’s not just that this information is accessible but it’s the fact that you’re tapping many sources of data simultaneously.

Data backlogs  While institutions such as the California Academy of Science has an effective mechanism to move physical objects about specimens to its on-line database, it has to tackle the less efficient work of digitizing its data backlogs. Institutions may have to deal with data that may have been collected well over a century ago. These data may still be useful, but are more cumbersome to deal with because they may not be catalogued, do not reflect changes in taxonomy, or contain the proper information required for them to be electronically queried. Thus these data require more time and effort to make them database-ready. Additionally, it generally has been the case that cataloguing has focused more on vertebrates – a product of, in part, differences in funding and research. Thus those data that have historically been easier to access remain easier to make digitally available. It is commonplace for a database to include vertebrates before less catalogued species such as insects.
Software and hardware development

A primary cost for making data electronically available lies in software and hardware costs. Some software has been developed for the purpose of making online databases usable by the public, but much of this software is under some form of redevelopment. Hardware costs depend on the size of the database and the extent to which the database is made part of a larger network. The Museum of Vertebrate Zoology improved its hardware and software through an NSF grant amounting to US $271,849. However, it is difficult to determine whether this amount covered the cost of the entire enterprise of moving 32 different specimen databases to a fully operational on-line system because this work entailed maintenance and some changes to the system as problems became apparent after the life of the grant. Ultimately, the MVZ has become a heavily used resource with these developments, as discussed below.

3.2 Funding for data access

Since biodiversity is both an issue of research and impacts our understanding of the earth’s resources and its economic use, any threat to access to data about biodiversity impacts our understanding of our use of the planet’s resources. Few, if any, uses of our environment is free from considerations about biodiversity impacts and, as such, the commitment to keeping data about biodiversity within the public’s reach is critical to maintaining sound uses of valuable and limited resources. Moreover, as our global economy expands and our use of the earth’s resources crosses national boundaries, we must keep data available on biodiversity to maintain responsibility and accountability for the sharing and equitable use of resources that impact biodiversity.

Funding levels

Different levels of funding provide for different kinds of electronically accessible facilities. Institutions may be able through small grants to create a standalone database that is useful for staff at the institution. Second, funding is sometimes provided to develop a particular type collection – say, a herpetology database collection. To develop a system that may be used effectively by the public, or a database designed for an interoperable network, is exponentially more expensive, often more than what government agencies are able or willing to provide. Funding must be sufficient to cover the costs of not simply creating the database, but an interoperable database that can become part of a larger infrastructure. This cost entails not only the technology, but skilled staff and time. Often, museums do not have these resources available, and must ultimately weigh these costs against the general costs of operating the institution. When governments reduce funding for
these public institutions, the opportunity to produce such infrastructural support for publicly accessible databases is relatively slim. Unfortunately, some governments have had to reduce such funding, despite a general interests and intent in making data publicly accessible and increasing the electronic resources of public institutions.

Many scientists hope that the disaggregate specimen, institutional and national databases will be drawn together, as it would benefit research to have access to multiple collections simultaneously. GBIF is clearly an important and timely response to these hopes, and many developments for the development of national and regional nodes is underway both by GBIF and other scientific and standards setting bodies. However, the issue of funding levels still requires attention in many places where GBIF’s nodes are to be developed so that there is no longer a tension between funding for institutions and the development of interoperable networks.

**Funding cycles** Funding also comes episodically and in short-term periods so that networks and databases may not receive consistent support, or enough time necessary to develop the organisation alongside the infrastructure. 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 subject to yearly cut-backs. The result would be detrimental to the facility and the public who relies 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, its 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. While yearly reviews are often inevitable, consideration of the public nature of biodiversity data, and its usefulness for understanding larger issues of economic resources and sustainability must be recognized appropriately.

---

3.3 Career and reward structures

Skills for database work  The relatively new enterprise of making data available through on-line databases does not entirely match up with career and reward structures. Biologists are not trained in the skills necessary to make their data electronically accessible, nor do they have expertise in software development. While software programmers may have such skills, there needs to be some training appropriate to the needs of the biological sciences. Such skills can be acquired in many different ways. In one instance, a software developer, who was trained in physics, spoke of how he gained his understanding of the needs of the museum he represented because he joined his girlfriend who works in the field of biology on her field expeditions. Clearly, this form of personalized training is not available for all software developers. However, the requirement for these skills has become more apparent, and the development of institutions such as Conabio has led to more training opportunities. Indeed, Conabio has seen an increase in PhD students training in the field who have contributed to and worked with Conabio’s databases.

Rewards for database work  Another problem is that scientists do not receive any 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 goal of managing and sharing data is undermined. Tenure reviews do not consider work on depositing data into a database. Thus Meredith Lane of GBIF spoke in her interview about how more data would be made publicly available if the tenure review system would change.13 She herself has a 6000 entry database stored away in a drawer with little time to make it publicly available. According to Lane, the process of making one’s data available must be easy and recognized work. Currently, there is no institutional support.

3.4 Technology development

The technological requirements for creating an interoperable network for biodiversity data remain complex and challenging. Technological projects include producing the necessary protocols through which different databases can communicate and producing standards for data. As Meredith Lane asserted, sharing cannot happen without the proper infrastructure and standardized technologies.

13 Author interview, Meredith Lane, January 2, 2002.
Biodiversity data and standards, schemas, and protocols  

There are a number of approaches to developing standards. Discussions on standards can be exhausting and unproductive when there is a search for the perfect standard or for the more effective data schema that can anticipate all possible variations in the data. In some cases, a single individual may independently decide upon a working standard once it becomes clear that the impasse cannot be resolved. The aim is to create standards with some open-endedness. Some scientists argue that the minor differences between institutions will be surmountable by the general agreement that basic data querying is necessary and possible. Some look toward the development of software tools that can process data in different ways, and that the primary steps involves getting basic data available on-line.\textsuperscript{14}

Scientists interviewed were of the general feeling that the most urgently needed data to access is basic data, including data on the species name, collector's name, date of collection, and location of the collection. The accompanying aim is to develop a basic federated data schema. These schemas specify a number of fields, such as genus name, species name, and so forth. The sub-fields within biodiversity – herpetology, mammology, etc. – may require slightly different data schemas. In some cases, standards developed by scientific associations can be relied upon. Such is the case for the standard adopted by the Mammal Networked Information System (MaNIS),\textsuperscript{15} which adopted the categories developed by the American Society of Mammalogists. Darwin Core is an example of a more generic standard that may be used by different disciplines. Currently, Darwin Core 2 is under development.

Protocols are necessary for different databases to seamlessly work together. Unless all data were placed into a single type of database software—already an impossible scenario—protocols are necessary. One particular development in protocol technologies under development is called DiGIR, or distributed generic information retrieval, which is under consideration by GBIF.\textsuperscript{16} The aim of this protocol, among other things, is to offer a portal to many different databases that have data that

\textsuperscript{14} For an on-line site that presents current work on standards and biological collections data, see the Task Group on Biological Collections Data, which is a joint project of the Taxonomic Digital Working Group (TDWG) and CODATA, at http://www.bgbm.org/TDWG/CODATA/default.htm. See also its list of references about standards for biological collection data and other technology issues at http://www.bgbm.org/TDWG/acc/

\textsuperscript{15} MaNIS can be found at http://dlp.cs.berkeley.edu/manis/

\textsuperscript{16} See the website for DiGIR at http://digir.sourceforge.net/ For a presentation on DiGIR by Stan Blum at the 2002 TDWG meeting in Brazil, see http://www.cria.org.br/eventos/tdbi/tdwg/presentations/DiGIR\_overview\_TDWG\_2002.ppt
conform to a data schema so that users will be able to search and retrieve structured data that are in fact quite different and originate from different database software. By letting its data ‘migrate’ from its in-house database through a federation schema that creates ‘views’ of the data, the host institution can filter data from public use, such as data on the location of endangered species. Such filtering must be explained in metadata offered by the host institution. The Museum of Vertebrate Zoology’s on-line system, discussed below, is based on the principle of decoupling an in-house database from the publicly accessible information system. Many scientists and software developers spoke of their hopes that GBIF initiatives will incite the use of protocols that will enable the different sub-field collections (for example, herpetology and mammology) to be accessible together, as they are currently separated into different data access facilities (HerpNet, MaNIS). Discussions about the development and use of protocols provide another occasion to bring up the potential of GBIF as a mechanism to leverage the different initiatives into an interoperable whole.

**Incomplete data and standards** There are always instances where data standards will not fit with the data available. For example, data might be connected to a particular specimen that was deposited in an institution without information on its collection date. Or, there may be variation between deposited specimens in how the collection data are specified. A baseline must be decided upon to make as many data as query-able as possible; for example, making only the year necessary for queries. However, the more variation a standard can handle, the more data will be able to be drawn into a given query. Decisions on how to deal with incomplete data must be decided upon in the early stages of standards development.

**Technology and policy** Interviewees in this study conveyed a sense that technology development must be met with policy development that allows that technology to be useful to the public. There is a sense that technical problems will be solved while policy and data rights may not be solved in due time. Technical interoperability fails completely in the wake of policy and legal non-interoperability. When data are digitized and made part of a database in an interoperable network, it will be unable to circulate if an institution or nation-state has heavy restrictions on its use. No technical solution can overcome the barriers erected through legal and economic restrictions. In addition,

---

17 Author interview, Wouter Los, May 11, 2002; Lane also shares this concern that technical issues are easy compared to the legal and economic ones.
heavy restrictions that require enforcement mechanisms, such as signed user releases, will only make access less efficient and less useful because of the bureaucracy that they will require.  

### 4 EXAMPLES OF INSTITUTIONS ENGAGED IN LEVERAGING ELECTRONIC DATA ACCESS FACILITIES

#### 4.1 The Global Biodiversity Information Facility (GBIF)

**GBIF's main purpose**  GBIF aims to become an interoperable, distributed network of scientific biodiversity databases whereby different hosts ‘affiliate’ with GBIF in the shared goal to make ‘the world’s scientific biodiversity data freely available to all.’ 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.’ This work will amount to bridging other regional and national data networks to create an ‘interoperable whole.’

**Background to GBIF and its structure**  GBIF was initially proposed within the OECD’s Megascience Forum (now the Global Science Forum). This report cannot provide a history of OECD, or the details of its organizational structure. However, the way that GBIF’s developers conceived of the structure so that financial and staff requirements could be met, and a legal entity could be established without creating too much of a top-heavy bureaucracy, is important to consider. The creation of an internationally viable organization absent an international charter required a number of considerations, many of which are treated in independent reviewer Eric James’ report to the OECD on GBIF’s establishment. In this report, he wrote ‘The way in which these legal requirements are met may be the

---

18 This is a worry articulated by almost all interviewees, including Meredith Lane and Wouter Los of GBIF’s original Steering Committee. Simon Owens argues that restrictions on data access—by required signed agreements, clearing data rights, and so forth—will undermine the usefulness of access. See his article with Alyson Prior, ‘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.


20 Business Plan for the Global Biodiversity Information Facility, Section 1.1

most important factor determining the structure of the organisation that is created.’ The lessons learned from GBIF, including why it was not appropriate nor feasible to be handled by the OECD, are important for considering the development of international information services that draw upon distributed resources. In short, the importance of access to biodiversity information is global, and thus involves non-OECD countries. Second, the interest in creating GBIF’s infrastructure must be motivated by individual nations, or Participants. As such, GBIF’s structure is such that there is some top-level management of work programmes, control over funds and staff through an Executive Secretariat, hosted by the Denmark (the ‘Secretariat Host’) at the University of Copenhagen Zoological Museum.

**Participation in GBIF** Participants of GBIF (defined as a ‘country, economy, organisation or inter-governmental organisation, or entity designated by a country, economy, inter-governmental organisation or other organisation, that has signed [GBIF’s Memorandum of Understanding]') will share data and develop GBIF nodes. Participants financially contribute to GBIF through a fee. This fee is determined by the Participant’s rank within seven different levels of per capita GDP. Participants have the right to vote on GBIF activities. They do not need to be a member of the OECD as GBIF is not formally linked to the OECD but rather originated out of the OECD. Most of the activities that will contribute to the development of GBIF’s network will come at the national level. The GBIF International Secretariat will coordinate these activities. In some cases, nations will develop biodiversity data programs with the help of the Secretariat and seed money available through GBIF.

**Participant nodes** Participants are responsible for setting up one or more nodes. A node is ‘a stable computing gateway that allows real-time inter-operational search of multiple institutional, national, regional and/or subregional databases containing primary or meta-level biodiversity data (such as specimen records, catalogues, bibliographic, sequence, protein and ecosystem data) or a single, web-accessible computer containing one or more significant maintained biodiversity data-

---

22 The OECD offers its perspective on the issue in the following: ‘If opportunities for co-operation are identified, interested governments will decide whether they wish to participate in discussions leading to the negotiation and implementation of international projects. The responsibility for negotiating final agreement and administering the co-operation will reside with the participating governments.’ See An informal introductory note from the OECD secretariat to Mr. Eric James’ paper *Establishing International Scientific Collaborations: Lessons Learned from the Global Biodiversity Information Facility*, available at http://www.oecd.org/pdf/M00027000/M00027487.pdf February 2002

bases. The node must contain information about metadata, provide information about data content and quality and provide statements about its policy on matters such as Intellectual Property Rights.

**GBIF’s programmes** GBIF has a number of work programmes to help develop a global biodiversity network that is supported by the necessary organisational, technological, financial and budgetary, and professional features. These programmes include:

- Data Access and Database Interoperability (DADI)
- Digitisation of Natural History Collections (DIGIT)
- Electronic Catalogue of Names of Known Organisms (ECAT)
- Outreach and Capacity Building (OCB)
- Digital Biodiversity Literature Resources (DILIB)
- Species Bank (SB)

GBIF must further develop its relationships with the organisations that can help realize these work programmes through offering expertise and efforts to make data digitally available on an interoperable network. For instance, at present, a request for proposal has been sent out to help institutions make their taxonomic information digitally available through GBIF seed money. The request is one of many ways that GBIF will perform outreach and support outside organisations. It is through such outreach and support that GBIF will succeed, and that a network of this scale will be brought to the public. Presently, GBIF aims to have its initial version of the full GBIF information system available for testing by June 2003, and to bring the system to the general public by December of 2003. This information system will be located at www.gbif.net.

**GBIF as leveraging institution** For the purpose of this report, the case of GBIF provokes concerns about how an infrastructure requires an array of support mechanisms that will ultimately serve the purpose of creating open access. Moreover, the necessity for GBIF is not simply to create a ‘global

---

24 See http://beta.gbif.org/nodes
25 For more description about these programmes, see www.gbif.org and http://beta.gbif.org/GBIF_org/wp
26 See http://beta.gbif.org/Stories/STORY1043078210
27 For more information on current work at GBIF, see its 2003 Work Programme at www.gbif.org
infrastructure’ but to help support local and regional initiatives from which GBIF will gain its strength and content. Some of this strength will entail turning to other institutions that have already developed open access regimes for advice. On the other hand, regional networks require GBIF to take the next step to creating an ‘interoperable whole’ by coordinating and connecting what is already present or under development so that regional initiatives can grow outward. GBIF can help prompt the development of other national nodes that will draw together different institutions. GBIF is in this way an important leveraging mechanism.

4.2 Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (Conabio)

Historical background to the institution Mexico began the development of Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (National commission for the knowledge and use of biodiversity, or Conabio) in 1992 to coordinate and facilitate basic research on biodiversity, to hasten conservation projects and risk assessments, to improve land use projects, and to repatriate data for use by Mexican researchers and government officials. Conabio has received most of its funding through federal funds, but has also received private grants. Conabio benefited from having approximately 4 years to develop as an institution before it was expected to show results. Consequently, the long-time director of Conabio believes that it managed to create the internal organisation necessary to undertake the substantial work necessary to both facilitate in-house database development and a network to repatriate Mexican data.

Scientific advancement Jorge Soberon of Conabio discussed the concomitant incitement and support of researchers, education, and publications based upon new analytics made possible by Conabio. Conabio provides grants to universities, among other institutions, to support fieldwork and the compilation of biodiversity data; as such, more research has been taking place in the field, and Ph.D.’s are being trained. Publications have been written based upon the aggregation of data drawn from a multitude of institutions connected through Conabio (one might be drawing data from 20 different institutions without notice). Further, recently developed software tools have been recently written, giving rise to new forms of research and knowledge.
Statistics on use

600,000 webpage hits/month; 150,000/wk; 21,400/day.

One question per day from a government official or other person who needs to investigate impacts of a given economic development project on wildlife and ecological systems in Mexico. These questions can now be asked of the information that comes from the distributed institutions that house the data.

4.3 REMIB

Data repatriation under Conabio  Conabio’s repatriation of data has been enabled through REMIB (Red Mundial de Información sobre Biodiversidad or Mexican Network of Biodiversity Information). REMIB is a network of nodes that connects databases housed in non-Mexican institutions. The repatriation of data to this biodiversity ‘mega-diverse’ country is critical as, for example, approximately 80% of Mexican bird specimens are housed outside of Mexico. While repatriation of data used to require researchers to travel to outside collections with a laptop and a camera to retrieve data on specimens, the repatriation of data by connecting databases clearly improves the rate at which such repatriation can occur.28

Historical background to REMIB  REMIB currently contains 4,099,365 records made available through a network with 17 nodes. The development of REMIB began not long after the founding of Conabio as it was recognized that a lot of work was required to keep data current with research, including research that leads to changes in taxonomy. In the words of Jorge Soberon, ‘instead of getting the data we were tapping into the databases of those people who were working with those databases.’ The decision to connect other institution’s databases to a network by making that institution a ‘node’ in the network established by Conabio was at the time rather revolutionary and took some ‘coaxing and convincing’.

28 Information about REMIB in English can be found at http://www.conabio.gob.mx/remib_ingles/doctos/remib_ing.html
The production and features of REMIB nodes

Part of REMIB’s development involved developing the regulations that serves Conabio as well as an impetus for non-Mexican institutions to become a node. This process has entailed:

- Granting the database owner the right to determine conditions of access to the data. Conabio will respect the institutions’ intellectual property rights over its database. It will also accept some restrictions to access; including, restrictions on releasing data pertaining to the location of endangered or commercially lucrative species and data that a researcher is actively working on for publication.
- Providing the institution with information on how much their data are being searched by users of REMIB, as well as where those users are located (through IP addresses). These data are helpful for institutions in its reports and proposals to funding agencies.

To qualify as a node, an institution must agree to open access and shared policies, and have the necessary human resources, infrastructure, and collections, in addition to a curator and a node administrator. REMIB explicitly recognizes that an institution must match its policy on open access with institutional and financial support.29

4.4 The Museum of Vertebrate Zoology

Data access at MVZ

The Museum of Vertebrate Zoology at the University of California, Berkeley, has an open access policy for its data, with the occasional delay of data that researchers are actively working on for publication. To enable this access, the MVZ created an on-line system in 1997 through an NSF grant. This work entailed connecting 32 different specimen databases that were not interoperable. Second, the MVZ designed a separate database to handle the public’s on-line queries. The separation of the institution’s database from the publicly accessible database allows the different operations required for internal and external use to work at a more optimal speed. Every week, data gets uploaded from the museums’ internal database to the public one. At the time

29 Institutions that have become part of REMIB include New York Botanical Garden Herbarium, The University of Southern Mississippi, the California Academy of Sciences, San Diego Herbarium, Texas University at Austin Herbarium, and the Museum of Vertebrate Zoology at the University of California, Berkeley, discussed in this report.
of this report’s writing, the on-line database contained 198,630 mammals, 174,989 birds, 141,147 amphibians, and 92,371 reptiles, in addition to scanned images of field notebook pages, photographs of specimens and other miscellaneous photographs such as photographs of habitats. MVZ, like Conabio and REMIB, tracks queries of the data, including who (based on the IP address or domain that originated the query) and which part of the collection was being queried. In May of 2002, 6271 on-line queries were made, with an average of 202/day.30

**Institutional up-scaling** MVZ is a node of REMIB. From an institutional standpoint, MVZ had the necessary features – including a dedicated software developer, a policy on open access, and a supportive curatorial staff – for REMIB to send one of its experts to help MVZ connect to the REMIB network. Of course, MVZ also had data on Mexican biodiversity. The institutional up-scaling that MVZ underwent with the assistance of an NSF grant allowed REMIB to increase the reach of its own network as it simultaneously increased the amount of users drawing from the MVZ. In this way, the MVZ provides one good example of how an institution can gain visibility and the use of its collections by working with other institutions toward the ultimate aim of expanding access globally.

**Museum work and data improvement** The development of an on-line system has changed the work at the Museum. As mentioned above, the increase of access to MVZ’s data has led to a tremendous increase in the delivery of specimen records to researchers: from 160,471 to 19,001,503 specimen records delivered within the first year of the inception of the on-line system! Instead of tracking down information for scientists who email or call the museum for information about their collections, curators are spending their time (1) loaning materials to scientists and (2) performing data clean-up. In the words of the software developer for the MVZ, ‘people just love to go on-line and see how MVZ has done something wrong, and tell us about it – which is perfect. And this is a great fear that a lot of folks have, is we’re not going to put our data online until their perfect, we don’t want to be responsible. So, we went the other route and said “we’re going to put our data online, we’re not responsible, tell us if you find something”. And it [has] worked out just perfectly. We get lots of information that helps us to clean up the data.’31

31 Author interview, John Wiercozek, May 15, 2002.
A pertinent lesson from GBIF is the urgent need for data access policies that will promote the access to basic data. The scientific work that GBIF could enable will not be fully realized if the many differences in access policies are not systematically addressed and appropriately resolved. Where data sharing facilities are hindered by problems of technical interoperability, data sharing will not happen without smoothing out the many policy hindrances. Barriers to open access include: the existence of different policies both within individual nations and between nations, the withholding of basic data and only offering metadata, the commercialization of data, and aggressive use of intellectual property rights, among others. As Wouter Los of GBIF asserted, the very success of GBIF requires overcoming these barriers.32

5.1 Conditions for use of data

One of the most basic conditions for use of an institution’s database is that the data may be used for research only. This condition is important for many reasons, but directly addresses the fact that biodiversity data refer to actual resources and ecological systems that must be carefully used and equitably shared. Other common conditions for use entail the following:

The user querying the data must:

– acknowledge the original source (and sometimes the database or network owners) of the data in reports, analysis or other publications that rely upon the data;
– receive prior consent for any repackaging, reselling, or redistribution of the data from the institution that provided the data, or agree to not redistribute the data at all;
– refrain from faulting the institution for inaccuracies or errors in the data.

Another condition, included in REMIB’s agreement, is that the user querying the data must not use the data in any way that will harm the ‘equilibrium of ecological systems’ or other conservation programs.

32 Author interview, Wouter Los, May 11, 2002.
Common data restrictions

Restrictions on access to data include: restricting access to information pertaining to the location of endangered species or the names of people, such as indigenous hunters, who collected the specimens to which the data refer; and, withholding the release of data that are actively being researched or to be used in publication.

Enforcement of conditions of use

The most explicit forms of protection come in the form of passing through a webpage with an agreement on conditions of use that must be submitted by the user. REMIB also notes that national authorities will be notified about improper use and any failure to abide by the conditions of the agreement. Simon Owens of Kew Gardens spoke of the potential difficulty involved in monitoring users. This monitoring might involve user submissions of signed legal agreements or credit card numbers. Museum staff would be forced to handle the logistics of arranging and monitoring these agreements. The skills and time required to engage in these legal exchanges exceed those currently available at many institutions. Moreover, the fact of having to sign agreements or provide credit card statements might deter users from accessing the system, thereby undermining the project of making data available.

Unclear potential users and uses of data

In the case of restricting access to non-profit users only, it can be difficult to ensure that data are being accessed by those with the intention to use the data for non-profit research. Electronic access can differ somewhat to access to data by telephone or museums visits because it generally lacks the conversation between museum staff and user about the user’s intentions. Some users who would be operating against the interests of the institutions who make their data accessible include persons or organisations seeking commercial gain (for example, bio-prospectors) and persons who will redistribute the data to other mediums without due attribution or permission (for example, placing electronic images into a book).

5.2 Legal uncertainties about data access

Data ownership

In general, scientists believe that research on publicly funded research on biodiversity that results in the collection and curation of biological specimens results in the public ownership of that data. However, there might be cases where it is not clear where the specimen has

33 Author interview, Simon Owens, May 1, 2002.
come from and whether the information about that specimen can be made publicly available. In cases where an institution holds specimens taken from illegal field trips or without informed consent, the ownership of the specimen and the data is problematic. Institutions within nations that have signed the Convention on Biological Diversity must follow the conditions of that agreement pertaining to any ‘benefit-sharing.’ Some scientists think that there may be an extension of the conditions of benefit-sharing, which is an important mechanism to ensure that a country benefits from the uses of its biodiversity, to include the accession of data that point to such resources. While the ownership of data is distinguishable from the ownership of biological resources, there is potential overlap that concerns some museum directors.

**Intellectual Property Rights** Intellectual Property has been labeled 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, which is a great concern by those facing reductions in government support. Museums can use IPRs to help them recover 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 a scientist is competing for publishable research on a given specimen. Moreover, there is concern that individual scientists do not necessarily consider the implication of their attempts to gain IP for the museums which they rely upon to store their specimens and conduct research.

Overall, the extent to which IPRs might apply to biodiversity data is unclear to many scientists. For example, it is recognized that it is possible that there might be multiple rights holders to a single specimen or database. It is not clear how those rights might extend to electronic images of that data. 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). ENHSIN is a pilot project to determine whether a European network can be developed and maintained, and what kinds of uses such a network would provide. Many museums or botanical gardens lack policy on IPRs and are uncertain about who may have rights to some of their data (e.g. photographs). More legal advice and research is necessary for networks to be developed properly and to remove obstacles to access.
INTRODUCTION

This case study is part of a larger project addressing the notion of trust as a key element that sustains data sharing practices among researchers and policies aiming at the promotion of data sharing. Two structural elements of data sharing, regulation and technology, are specifically examined as to their effect on building trust and thereby establishing (or impeding) data sharing. Given the interests of the OECD Follow Up Group on access to publicly funded research data, the role of data sharing regulation is most prominently discussed here.

Method

The project is pursued using virtual ethnography (Hine, 2000). This method aims to draw on the strengths of traditional ethnography, while adapting the method to the particular setting of cyber-science, in which elements like ‘location’, ‘presence’ and ‘interaction’ are problematized in relation to electronic and digital media. Notions of flow, networks and connectivity are especially important and explored as to their social and technological realisation (e.g. the significance of ‘links’). Methodological innovation is therefore part of this project and discussed elsewhere in some detail.

---

1 Paul Wouters and Peter Schröder provided helpful feedback on earlier versions of this case study. The author would like to thank J. van Horn for his participation in an interview, discussions and feedback about the fMRIDC activities, as well as other experts on the OECD project and members of Nerdi for comments and discussions on data sharing. Part of this work was funded by the Dutch Ministry of Education, Culture and Sciences (OC&W).

2 The project is entitled ‘Trust Building and Data Sharing – an exploration of research practices, technologies and policies’, Nerdi, NIWI-KNAW, Amsterdam. See project section of website: http://www.niwi.knaw.nl/nerdi/

The material used in this report consists of digital and paper documents, interviews conducted by phone and face to face, and an ethnographic analysis of the information flow in the database and on the Internet.

**BACKGROUND OF THE FMRIDC**

Data sharing is the subject of an on-going debate in the field of brain mapping, a recent stream of research that emerged at the intersection of neuroscience and cognitive psychology. In a field where experiments are generally expensive, in which groups of subjects studied are small, and where potential clinical applications in the fields of mental illness are hoped for, it seems that data sharing should be a welcome development. This debate has been punctuated by an intense controversy. Recent events around the launch of the fMRIDC show that data sharing is neither simply desirable, nor a straightforward concept. Furthermore, it is not an issue that researchers feel can be addressed without their active involvement. The very act of data sharing, it has been argued, requires the active (meaning face to face, collaborative, or at least interpersonal) participation of researchers, in order for the data to remain meaningful.

This case study specifically discusses one of the principal endeavours to promote data sharing in brain mapping, the functional Magnetic Resonance Imaging Data Center (fMRIDC). The data to be shared is mainly data from scanning experiments. These datasets are especially large (where a single study can be compared to the total data of the Human Genome Project) and growing (Yencharis, 2002). This database is spearheaded by Dr. Michael Gazzaniga, Director of the Center for Cognitive Neuroscience, Dartmouth College, US. The database was announced in a memo in June 2000, and almost immediately found itself in the middle of a controversy. The fMRIDC, in its short existence, has thus raised a number of issues, which are crucial to developing data sharing practices. Some of these will be discussed below, and at greater length elsewhere.

The fMRIDC receives funding from NSF/NIH, the Keck Foundation, and SUN, though it had links to Informix (IBM) for its first two years of operation. The goals of the Center are to provide ‘a publicly accessible repository of peer-reviewed fMRI studies and their underlying data.’ Part of the funding

---

of the fMRIDC is provided by the NSF/NIH, under the aegis of neuroinformatics, which is an important umbrella organization for developing neuroscience databases in the US.

The fMRIDC started accepting submissions of data in July 2000. The first data set was made available for request on November 15, 2000. That first data set, as well as all early contributions to the database, was part of a special issue of the *Journal of Cognitive Neuroscience*. Most of the articles from that special issue (10/11) were available by December 2001. The Center had shipped 50 data sets within the first 6 months of its existence, and this number had risen to over 300 by March 2002 (Interview, van Horn, 2002). As of August 2002, it contains 26 datasets, and as of February 2003, 36 datasets, which can be requested via the website of the database.

In terms of usage of the database and the data, these practices are difficult to evaluate. At present (19 February 2003), there have been 598 requests for data sets since the beginning of operations. (The community of researchers can be estimated to be about 1000 active researchers, based on attendance at meetings of the Organisation for Mapping of the Human Brain.) The center has also organised a competition for papers that make use of the data in the fMRIDC. The first winning paper was published in August 2002 in the *Journal of Cognitive Neuroscience*, and it made use of data sets in the collection (Lloyd, 2002). A second competition has been announced for 2003. No other publication noting use of fMRIDC data could be traced.

The fRMIDC also aims to be an educational resource. Besides serving as a repository and providing data to users, the fMRIDC hosts other activities, such as summer workshops (held in July 2001, 2002 and 2003), which aim to educate PhD students and researchers in ‘fMRI informatics’. Members of the Center also hold talks and give presentations about the database to a number of relevant audiences within neuroscience (Human Brain Project Meetings, Society for Neuroscience, Society for Cognitive Neuroscience, NIMH fMRI Experience Workshop, AAAS annual meeting).

---

5 *Journal of Cognitive Neuroscience* 2000 Nov; 12 (suppl. 2).
Present Data Sharing Practices and Regulations

Obtaining data

Currently, authors of papers published in the *Journal of Cognitive Neuroscience* are required to submit their data to the fMRIDC as a condition of publication. Material from other journals has also been received (Interview, van Horn, 2002) but was still being processed as of January 2003. Articles in the *Journal of Cognitive Neuroscience* include an ‘accession number’ in the acknowledgement section, which identifies the data in the fMRIDC. In the electronic version of the journal consulted via an electronic journal provider (EBSCO), a hyperlink to the fMRIDC website is also provided. Surprisingly, the instructions to authors on the website of the MIT Press (which publishes the journal) do not mention the requirement to submit data, nor does the journal website have a link to the database. The function ‘table of content alert’ forwards emails of the table of contents of the journal to those who register, but does not make mention of the availability of the data at the fMRIDC either.

In terms of accessibility, the fMRIDC can be considered a hybrid database, insofar as it is fully digital but has both online and offline components. Only some ‘annotations’ to the data are available via the web interface. The actual experimental scanning data can be requested via a webform, and is then sent to the requesting researcher by the post on CD-ROM or tape, according to the requester’s preference. The following fields of ‘annotation’ can be searched: author, abstract, accession number, keywords, PubMed id (a code assigned to documents included in the PubMed database of biomedical and life science literature) and title. This information can be searched via a search dialogue box, modelled on the PubMed interface.

Goals of the Data Center

The fMRIDC is a research-oriented database. Its goals are to ‘help speed the progress and the understanding of cognitive processes and the neural substrates that underlie them by:
The principle of data sharing is an apple pie issue.

- Providing a *publicly accessible* repository of peer-reviewed fMRI studies.
- Providing *all data* necessary to interpret, analyze, and replicate these fMRI studies.
- Provide *training* in neuroinformatics for both the academic and professional communities.\(^7\)

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 include commercial elements (for example, where access is tied to a subscription).

**Databases in the field and actual use of shared data**

The fMRIDC is one of several initiatives in the brain mapping, though the only one to be both centralised and to offer ‘raw’ data. Anecdotal evidence indicates that none of the repositories can be said to be widely used at present. Only a small handful of papers have been published, in which databases of research findings are used. The fMRIDC itself has attempted to stimulate use of shared data by organising a competition, with a monetary incentive and publication as reward for the winning paper, based on re-use of fMRI data.

It is indeed extremely difficult to follow the exact re-use (if any) of data from databases. The fMRIDC keeps track of the number of data sets shipped out, but does not have any other information on the use of the data shipped. It also keeps track of the usage statistics derived from its log files, but the use of this information is quite limited. Researchers making use of the data deposited in the fMRIDC are asked to cite the original publication from which the data has been used, and to indicate the accession number of this data set. Tracing such an accession number cannot be done using traditional bibliometric techniques (i.e. these are not included in references). This is an important lacuna for data base developers and funding agencies, since it hinders systematic evaluation of the very purpose of databases.

\(^7\) http://www.fmridc.org/ (accessed 29 May 2002).
THE CONTROVERSY AND POLICY OF THE FMRIDC

While, in the words of the editors of Nature (Anonymous, 2000b), the principle of sharing primary data is a mom and apple pie issue – no self-respecting researcher can be against it – objections to it can be formulated, however, when it comes to discussing it as a practice to be implemented. Controversies often bring parties to make their standpoints more explicit and dramatically expose problems and unexpected consequences. The following sections attempt to describe the practices of the FMRIDC in relation to discussions about data sharing in the research community. In each section, a particular aspect of the controversy is analysed, and its implication for data sharing explored.

The launch of the FMRIDC and the first reactions

In June 2000, a letter by Michael Gazzaniga, editor of the Journal of Cognitive Neuroscience, indicated to contributors that in future, authors would be required to submit their experimental data to the National fMRI Data Center. This letter seemed to imply that other journals (Science, the Journal of Neuroscience, Cerebral Cortex, and Neuron) would also be requiring this from authors, making recipients fear that journals were trying to force the FMRIDC onto researchers (Aldhous, 2000). In response, a letter, signed by over 50 fMRI researchers was sent to several journals, as well as to the bodies that provide financial support to the fMRI data center (Keck Foundation, NSF/NIH). The authors requested from editors that they take a clear position with regards to mandatory submission. This letter had as an effect that authors were able to have a time lag in the submission of their data (Aldhous, 2000) This controversy came to the attention of science journalists and led to further discussions and clarifications about the policies of the parties involved.

---

8 The official publication of the Cognitive Neuroscience Institute, a non profit organisation
9 It has been noted that the term national was perceived as provocative (Anonymous, 2000b). The term ‘national’ has not been used in subsequent communications.
10 According to Jack van Horn, editors of other journals had more or less agreed to implement this policy, but later, in the face of the backlash from researchers, ‘decided not to decide’ and effectively did not implement this (Interview, van Horn, 2002)
11 http://www.psy.vanderbilt.edu/faculty/gauthier/fmridc_letter.html (accessed 29 May 2002). This webpage has since been unlinked and later removed from the website.
Michael Gazzaniga, in response to the letter from researchers stressed the communicative role of a database, in an interdisciplinary field, and the support role of the database, as an enabling tool that would ease the burden of sharing data for researchers (digital letter, 8 August 2000). Most editors whose responses have been made public distanced themselves from the announcement, and stated that they had no intention of requesting submission of data by authors. Some also noted that they did require this for other types of data, but that the functional brain imaging community did not seem to have formed consensus that this would be desirable (Current Biology, Nature, Nature Neuroscience). Stephen Hyman, director of NIMH, commented in Nature that while archiving is a good idea, ‘in all cases where sharing is mandated, it has been after extensive consultation with the community (Anonymous, 2000a).’

KEY ISSUES IN DATA SHARING

Lack of trust and role of data in brain mapping

In the controversy around the launch of the database, it became clear that data sharing was not a trusted initiative, and researchers perceived increased risk in having to participate in data-submission to a database. This phase of the debate illustrates vividly that data is not easily separated from brain mapping experiments, for the community of researchers. The separation, implicit in the policies of the database, seemed to run counter to the ways of working (and especially of publishing) of researchers, for whom brain mapping experiments do not lead to a single paper. By having to provide all data from an experiment upon publications, researchers argued that this might involve the risk of being scooped. Researchers would not be able to benefit from the data they had themselves gathered, because sharing at an early stage would weaken their claim to publications on (different) data from the same experiment.

Other parties may also run a risk in pursuing this strategy. While the researchers who signed the letter clearly felt that editors held the big end of the stick, journals may also risk losing contributors if they make too many demands on authors. Up to now, this does not seem to have affected the rate of submission at the Journal of Cognitive Neuroscience (Interview, van Horn, 2002).
Journals may also further be affected, in terms of their ability to maintain the peer-review system. Part of the strategy of the fMRIDC was to use the association with publication as a trust-building mechanism, to give potential users of the database some assurance of the quality of data. But coupling the data to publication in this way further involves pressures on the reviewers and the peer review system, since the scope of what is to be reviewed may increase. Brain mapping is perhaps especially liable here, since the data sets of experiments are notoriously large.

This part of the debate therefore illustrates that making submission of data a compulsory activity in relation to publishing, a range of factors come into play, besides the willingness of individual scientists to participate. The release of data is not (perhaps ‘not yet’) seen as separable from other research activities such as publishing. 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. The controversy is therefore not only about ‘submission’, but also about the consequences of this submission. In contrast to other fields (high energy physics, molecular biology, biodiversity), functional imaging does not have a strong distinction between research and data management as separate sets of activities.

‘Ownership’ versus ‘control’ and institutional resources

Another set of arguments brought forth in the controversy highlights the difficulty in separating research and data management, because of the consequences this separation can have on the organization of research. The actual circulation of data might actually challenge the relative ‘power’ of different groups, privileging some and disadvantaging others.

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 (Whose scans are they, anyway?; (Aldhous, 2000), 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 (Brighton, June 2001), the issue of ownership was rather less prominent, and the consequences of a new circulation of data were foregrounded. Objections mainly concerned the circulation of information in relation to the organisation of research. The argu-
ment 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 therefore be much greater for smaller labs (Gauthier, personal communication, June 2001).

The fMRIDC has responded to these objections by maintaining that publicly funded data is considered a public good, beyond the personal efforts that are involved in collecting it, and thus should be shared. Recognising this increased risk for some researchers is important, however, since it may reduce trust in data sharing initiatives, especially if the initiative comes from what peripheral centers consider to be the ‘center’. Due to persistent protest, however, the fMRIDC has agreed to the following provision: researchers may request a timelag in release of the data (‘6 month data hold’). During this period, only the aggregated data would be released, thereby diminishing the likelihood that others might use the primary data. It is not clear, however, how this can be implemented, since data from studies are released as ‘packages’ (along with all other information from the study) on cdrom.

However the technical details are solved, these arguments highlight a legitimate fear about the consequences of separating research and data management. If a group’s data management capacities are smaller, this group may feel more vulnerable and resist participating in a flow of data that might harm one of its main research activity, publishing. In other words, a group may feel that the new circulation of data may be more risky. Given the structure of fMRI research, where a multitude of smaller centres have arisen in the last 5-6 years, this hurdle may be difficult to overcome without providing additional incentives or protections.
Data is not enough: how to describe an experiment?

Following on the initial reactions directed at the issue of publication, another difficulty with data sharing was amply discussed among researchers. Many questioned the feasibility of understanding data from functional imaging experiments, for anyone not directly involved in their production (PNAS letter, Nature, etc.). The need to describe the data in such a way that it is intelligible is a point that often arises in databasing and meta-analyses. But besides the general issue of the need for context to understand any data, there are problems that are specific to brain imaging data. An important part of the methodology in cognitive neuroscience is informed by a comparative strategy. The set up of experiments thus often involves comparative analysis of differences between two activations in the brain (task minus control, for example, reading meaningful words compared to reading nonsense words). These tasks therefore require careful description if comparisons of different ‘subtractions’ are to be made.

Furthermore, the analysis methods in brain mapping are extremely complex. A map, in brain mapping like in any other type of mapping, involves the coordination of different types of information in relation to a representation (e.g. functional and structural information). While there is general agreement on the principle for doing this (Beaulieu, 2002), there is still enormous variability in terms of what is considered relevant to understand brain mapping data from an experiment. The sub-group of cognitive psychologist in this community is particularly interested in pursuing methodologically sophisticated and subtle contrasts in activations. These experiments therefore especially require careful description of the conditions of the experiment, of the tasks being performed in the scanner and of the relations between them.

In the fMRIDC, raw data is stored, as well as ‘information deemed necessary in describing various aspects of the experiment.’ As well, authors can provide additional information ‘they deem necessary to fully describe their experiment. In order to accommodate this, the fMRIDC database is designed so that it can be extended to store a researcher’s unique descriptors.’ This accommodation of authors could be critiqued as running against the principle of archiving data for meta-analy-

---

12 These responses were consulted on the webpage dedicated to the letter-writing initiative by researchers (http://www.psy.vanderbilt.edu/faculty/gauthier/fmridc_letter.html) (accessed 29 May 2002) (This page has now been removed.)
sis. From the point of view of information retrieval, a unique describer is a ‘weak’ one when isolated in a large dataset. When understood in terms of the controversy about the very possibility of meta-analysing, this accommodation may be read as a necessary concession to reassure researchers about the possibility of maintaining the integrity of their experiments.

This point of contention can also be understood in terms of the underlying tension between research and data, and the difficulty in separating the two for this community. The very production of the data is closely tied to experimental conditions, so that it is difficult for researchers to conceive of standardisation of the data. To force this uncoupling would either mean that the data is meaningless because too decontextualised, or that experiments themselves become standardised. Obviously, neither option is attractive to researchers. The fMRIDC has tried to take a middle road, providing freedom of description and the possibility to make up a package of related material to accompany the data.

**Quality control**

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*, though contributions that have appeared in *Journal of Neuroscience* and *Cerebral Cortex* are currently being processed (Interview, van Horn, 2002).

Staff of the fMRIDC performs a further check of the coherence of the data submitted. Authors are specifically involved in a final check of the data and information to be included in the database, and have the final word as to the exact contents of the data and descriptive information submitted. The attention paid to maintaining author control throughout the process may arise from the concern on the part of authors at the launch of the database. These processes are quite labour intensive, and require the position of a data submission officer for the database.

The uproar in the community drew attention to the fact that publishing and data sharing are considered different activities. Similarly, assuming that peer review of articles and quality control of data are equivalent may lead to problems. Rather, these two different levels of activity may interact, and while the consequences may be positive or negative, they should not be underestimated.
Data sharing may have a positive effect on the peer review system. Some of the editors involved in the debate have noted how difficult it is in the peer-review process to evaluate claims in the field of functional imaging. The field of imaging is perceived as vulnerable to artefacts because of its tools (a tiny movement of the head in the scanner can cause much noise). There are also a large number of methods for analysing data and integrating various types of data within experiments, which make it difficult to compare studies across labs (Anonymous, 2000a) (OHBM, 2001). In relation to these difficulties, it is possible that a data basing effort might, in the middle to long term, alleviate both problems. It may increase standardisation, and possibly allow for detection of artefacts by other researchers. Whether this potential is realised, however, depends on a number of factors. The NIH had attempted to stimulate this process by holding workshops, in the 80s. Through these, some progress was achieved, but the main standardisation in the field of mapping occurred because of clear benefits to researchers in pursuing their experiments, not because of benefits to the community as a whole (Beaulieu, 2002). Still, it remains that data sharing may help improve quality-control mechanisms, if indeed greater circulation of data is achieved (see next section for a discussion of one important condition for this to occur).

Second, this case also draws attention to possible negative interactions between quality control and data sharing. The transfer of peer review of journal articles to the building of a database raises questions about how this mechanism may work, when its context and goal change. To return to the tension between experiment and data, the peer review system functions at the level of experiment. In the case of the fMRIDC, this mechanism is taken to also be useful for quality control at the level of data. Whether this is the case should not be assumed. It can be argued that experiments that are considered to be of high quality should be based on high quality data. But reviewing an article and reviewing complex data sets and all their analyses should not be assumed to be equivalent. Some journals have already recognised this: the BMJ have experts in statistics whose task it is to review formal aspects of data for articles submitted, as a way of relieving pressure on reviewers and of improving the quality of the articles published. Given that, in the case of the fMRIDC, the goal is to make even larger quantities of data available than are available in a published article, then the challenge on reviewers may be even greater, or the data go unchecked.14

14 Another databasing initiative, BrainMap DJI, also involves peer-review; first, at the level of publication, but also in terms of the coding of the data to be submitted, which is reviewed by designated reviewers who are familiar with the databases’ coding system.
The interaction of databasing initiatives and of existing quality control mechanisms can therefore have positive or negative effects. The suitability of traditional peer-review (associated with publication) for the purposes of database building should not be assumed, however, since the conditions that gave rise to this mechanism of quality control are changing dramatically, and may need to be adapted.

**Cultivating software development and sharing**

Currently, one of the major areas of activity at the Data Center is the development of software. This is seen by the Center as a way of increasing the general acceptance of its initiative and the need to share data generally. Creating software will enable researchers to do analyses across studies and make possible the meta-analyses that are part of the potential benefits of databasing and data sharing (van Horn interview, 2002). Furthermore, development of software is an interdisciplinary endeavor, which requires specialized knowledge beyond the resources of the fMRI and brain mapping community. Data-mining tools are being developed in the Department of Computer Science at Dartmouth, and information theoretical approaches are crucial to the suite of tools being developed (Interview van Horn, 2002). The website suggests that these tools will be ready in July 2003, and they are being developed using an open source philosophy. These will go beyond the current tools used in the mapping community (which rely on the ‘general linear model’), which are most suitable for working with data from single studies.

A range of tools and analytic approaches are under development, at Dartmouth and elsewhere. While they will differ from the traditional ‘single study tools’, it is useful to briefly note how software is generally developed in the community, since these practices may influence the generation of these new meta-analytic tools. Generally, software for the analysis of single studies is either developed as ‘packages’ or as ‘suites’ of programmes in which the various components are more autonomous. Computer science expertise is furthermore a standard requirement of the staff of brain imaging centers, and the field of neuroinformatics has close ties to brain mapping and experiments with brain imaging technologies since its early developments. Still, in spite of the fact that this

---

15 Brain mapping was identified as one of the promising fields for the integration of neuroscientific data (the goal of the Human Brain Project), due to early efforts at standardisation and at interdisciplinary projects (Pechura & Martin, 1991). The Data Center’s grant from the NSF falls under the Neuroinformatics theme, as part of the Human Brain Project. The National FMRI Data Center receives the following: NSF BCS-9978116. See description of neuroinformatics at http://www.nimh.nih.gov/neuroinformatics/index.cfm. Some of the history of the HBP is detailed in (Beaulieu, 2001).
expertise is embedded in research, it is also frequently overstretched. Users of software therefore complain about the lack of support and maintenance available. While grateful for the feedback they receive, developers of software consider that requests for support from fellow researchers are endless, however, and that they must be careful with the amount of time they invest in this activity.

The functional imaging community has noted these tensions, and the NIH and other US institutes have launched the Neuroimaging Informatics Technology Initiative (NIfTI). Its main activities have been workshops and attempts to enumerate the existing software (which it feels unable to do with confidence, a sign of the diversity of practices in the field), and in 2002, a funding scheme for the development of neuroinformatics tools. Another international group working under the aegis of the OECD Global Science Forum, the Neuroinformatics group has also been active in surveying the available software tools (among other activities). These initiatives are quite recent, and the impact of their role as coordinating bodies or ‘clearing houses’ is not yet prominent.

There is therefore awareness of the need for larger-scale initiatives in software development, as well as a strong presence of informatics expertise in labs and a culture of sharing of software in the functional imaging community. Most activities, however, are still concentrated at the level of tools for the analysis of data from single experiments, in contrast to the needs for developing data-mining and meta-analyses across studies. The development of such software has been identified as a key hurdle to be overcome (Eckersley et al., 2003). One should not be overly optimistic, given the other difficulties noted in this case study. The existence of these tools, however, will form the basis of a further incentive to share data, since they will make it possible to use shared data.

**Coordination of standards: regulatory mechanisms, professional bodies, and funding organizations**

The importance of standards for data and for reporting analyses of functional imaging experiments has been noted above. The major efforts at coordinating standards are reviewed in this section.

Professional bodies and consortia of researchers involved in the brain mapping community have begun to address the issue of standards in data sharing. The grantees of the Human Brain Project organised a survey on the topic of inter-operability at the annual meeting of the Organisation for
Human Brain Mapping in Dusseldorf, 1999, and NIfTI has a questionnaire on its website, since at least July 2001. The Subcommittee on neuroinformatics of the OHBM has also recommended that it play a role in the development of procedures for sponsoring discussion and disseminating expertise about databasing, assigning credit to original authors, standards for informed consent and confidentiality of patient data (OHBM, 2001). This recommendation marks a new phase in discussions on this topic. In earlier discussions, the issue of standardisation met with a more divided response. Some parties argued that good science (and not standards) should dictate decisions made by scientists, and that the OHBM should support good science, not become an enforcer of standards. The fMRIDC controversy seems to have served as a catalyst to setting this issue as a serious agenda point, worthy of consideration.

At the organisational level, the issue of data sharing and of the need for standards is also receiving increased attention. Furthermore, because the scope of databases has from early on been defined as necessarily international in scope, a network of agencies has evolved around the topic. The Human Brain Project has played an important role in this, and the HBP has had meetings with Medical Research Council of the UK and the European Science Foundation (Koslow, 2000). A joint US-European Community neuroinformatics steering committee has also been established. This has led to funding in Fifth Framework, to support training and database creation. At the level of the OECD Megascience Forum, the topic of Neuroinformatics has been tackled by the Working Group on Biological Informatics. The group was established in 1996, and reported to the forum in 1999. Now working under the Global Science heading, the Global Science Neuroinformatics group (GSF-NI) has 3 goals. It will (1) collate a list of neuroinformatics resources, (2) develop database guidelines for content, organization, quality control, and standards to ensure interoperability and longevity of the neuroscience databases and (3) create a Neuroinformatics Portal. The emphasis of this body is on making tools that exist visible, but it will also consider the evaluation and reward structures for scientists in relation to the evolving context of databases and collaboration.

---

16 Imaging Interoperability Workshop, March 22 - 23, 1999, Bethesda, MD, US.
While the issue of standards for this field has found a place on the agenda of funding agencies and professional bodies, it cannot be said that at this time, either clear incentives or clear enforcement accompany this awareness of the importance of standards.

Databases and IRBs: subject protection and data sharing

A further set of regulatory bodies is involved with data sharing in functional imaging. The fMRIDC’s activities, because they involve the circulation of human subject data, are thus shaped by the practices of various bodies, often quite far removed from the Center, both institutionally and scientifically. Functional imaging data submitted to the fMRIDC falls under the jurisdiction of bodies based at the researchers’ institutions, known variously as Internal Review Boards or Institutional Review Boards (IRBs). These are effectively ‘ethics’ committees, which grant approval for research to be pursued. The degree of involvement of these bodies can vary, but it is usually quite high when human subjects are involved (as in the overwhelming majority of functional imaging studies). An important task of these bodies is to supervise procedures such as informed consent and the protection of subject confidentiality.

Immediately following its launch, the fMRIDC has seen itself obliged to align its activities to the requirements of IRBs. Indeed, making data sharing conform to these requirements seems an arduous task in the views of many researchers. The protection of human subjects was often raised as a concern and obstacle to data sharing (e.g Fox in Aldhous, 2000). The fMRIDC has therefore produced guidelines for the anonymisation of data19, and policies about subject protection are addressed explicitly in the website of the database. The memo and accompanying material on the website also point to the need to include submission to a database (and therefore, use by other researchers), when obtaining consent from subjects and IRB approval.20

---

19 A memo concerning it was issued on September 6, 2000 (fMRIDC Memo Concerning Submission of Data from Human Subjects) and explain the importance of removing elements that might lead to identification. They also explain how data is double-checked for removal of all identifiable elements (in brain imaging, not only identifiers such as names and social insurance numbers need to be removed. Anonymisation must also include stripping the scans of elements that could be used to reconstruct facial features of subjects, since more or less the entire head is scanned).

20 A letter explaining the use of data made by the database is also available on the website, for use by researchers who might need such a document to obtain permission from their Internal Review Boards in order to submit data to the database.
Data sharing, as a new activity different from research, therefore needs to be integrated into the work of regulatory bodies. In this case, these bodies are particularly important for maintaining the public’s trust in scientific research. The integration of data sharing into IRB procedures seems especially difficult because different bodies may perceive data sharing differently. Van Horn has commented that collaboration or co-authorship based on data from the fMRIDC might be undermined by the possibility that IRBs from two different institutions may not see the issue in the same way. Furthermore, many subject protection measures are embedded in national legislation. The evaluation of data sharing and the threat it may pose to human subjects also varies per country, making the fMRIDC careful about the prospect of accepting data from countries other than the US.

A further difficulty concerns the relative novelty of this kind of activity, and especially the digital and networked nature of the practice. For example, data mining implies that a resource for ‘discovery’ (Van Horn & Gazzaniga, 2002) is created, and that the uses of the data are not known a priori. To IRBs, this open-endedness may seem at odds with one of the pillars of informed consent, where use is to be stipulated. As a result of this open-endedness and the unpredictability of future use of the data, IRBs may express discomfort with asking for ‘blanket’ consent to be given.

**CONCLUSIONS**

*Research and data management: a difficult separation*

In some scientific settings, the separation of the production of data, experimental work and the management of data are visibly and functionally different, if integrated, activities. In the case of functional imaging research, this separation seems underway, though the relationship is not nearly as clear as in high-energy physics, biodiversity or molecular biology. Indeed, what the controversy around the fMRIDC allows us to see are the difficulties in achieving this separation. Potentially, data would flow in a new stream, one that is separate from the context of experimentation and publication, and from the traditional reward structures associated with these. The Center, and indeed functional imaging as a field, are much more recent than the research fields and the data repositories that serve either high-energy physics or molecular biology. It may be that such a separation between research and data management is becoming a feature of contemporary science (informational turn in science). While the current set of case studies suggests as much, much more work
needs to be done to apprehend whether this is an overarching trend. The rest of this section focuses on the conclusions that can be drawn from the evidence gathered up to now.

The reasons identified in the various sections for the relatively slow start of the fMRIDC (many of which are also valid for other databases in the functional imaging community for that matter), indicate that data sharing is not simply an activity that must be engaged in, an extra task for scientists to put on their to-do list. Rather, making data available involves changes in many aspects of scientific work. These are reviewed below.

**Conventions**

The importance of field-specific conventions about how knowledge and data can ‘packaged’ are important for understanding ‘access’ to data (Hilgartner & Brandt-Rauf, 1994; 362). 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?

In the process of setting these conventions, discussion and coordination within the research community are important. The launch of the fMRIDC has stimulated the former, and as discussed in the section *Coordinating Standards*, the importance of coordination has been placed on the agenda of various funding and professional bodies. Still, these conditions may not achieve a culture in which data sharing thrives, 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.
Incentives and obligations

Journals  An earlier effort in the functional imaging community to build a database (BrainMap) depended entirely on the voluntary contribution of researchers. This database was not particularly successful in attracting submissions. In the efforts of the fMRIDC, the incentives have been indirect from the view of the researcher. Rather, the obligation to submit data to the Data Center as a condition of publication has been stressed. This coupling of publication and submission have been used before, for example in molecular biology, where an accession number became a requirement of submission of an article for review for all the main journals in the field (Lenoir, 1999). As discussed, no such consensus yet exists in functional imaging and neuroscience journals. It must be kept in mind, however, that publishing and data sharing, while related, involve different ‘flows of data’ (Hilgartner & Brandt-Rauf, 1994) in the view of researchers. The possibility of changing the relation between the two exists, but must be handled carefully, so as not to cause backlash. Furthermore, it seems most likely that these changes will require mutual adjustments – for example, the development of reward systems that are related to the submission of data.

Funding organizations  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 (this and other) funding agencies.

New types of work  While, as noted above, sharing data is not ‘just’ another task, it is also that. By all accounts, developing good data management practices involves work and often, new skills for researchers. A recent survey of academic geneticists, for example, found that requests made directly to researchers were surprisingly often not honoured. Most often, the amount of work involved was given as the reason for this (Campbell et al., 2002). In discussions with researchers around the fMRIDC, this element was also mentioned as a disincentive. 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.
Support for new tools  The value of submitting data has not yet been demonstrated to researchers in terms of clear research benefits. As mentioned, the software tools needed to be able to handle, let alone ‘exploit’ data across studies are not yet well developed. This has implications for ‘added value’ functions of data repositories, and for the motivation of researchers to re-use data. The strong presence of informatics expertise in labs, and a culture of sharing of software in the functional imaging community, seem auspicious for the development of these tools. The incentives for developing software for the use of databases, however, may not be so obvious to software developers working in laboratories, where their agendas are set by the local needs for tools for analysing experiments. There may be the need to change the current culture to one where commercial exploitation of software provides incentives, or else tailor funding possibilities to support this kind of work.

Interactions

The second set of issues raised by this case study falls under the heading of interactions of this new flow of data with other processes and structures.

Quality control mechanisms  The main quality control mechanism of the fMRIDC involves the peer-review system associated with publications. Once a paper is accepted for publication, the data associated with it is considered to have been vetted for inclusion in the database. Quality control is key to building trust in a database. Clarity about what kind of review is performed should therefore be a priority. Two important assumptions should be queried in relation to the reliance on peer-review. First, this reliance may put added strain on the peer-review system, if data are also to be reviewed in detail by reviewers. Second, the review of data may require different type of reviewing than that 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.

Ethical issues and human subject data  Databasing and data sharing, especially in relation to the open-endedness of use and circulation of data contained in databases raise novel ethical issues. This open-endedness is desirable in the eyes of developers and users, but can be particularly prob-
lematic for bodies charged with regulation of research ethics. The creation of infrastructures for
data sharing therefore interacts with regulatory bodies (‘IRBSs’), whose approval must be obtained
to share data. As such, these bodies may block the circulation of data. This point is sometimes
made painfully clear to researchers. In a project in the field of brain imaging in the first phase of the
HBP, one of three partners in a consortium saw their share of the data gathering held up for years,
because the local IRB objected to the kind of ‘blanket’ consent which subjects were asked to pro-
vide. A more recent project21 (BIRN) which will involve sharing of imaging data has taken this issue
into consideration from the early stages and is trying to learn from the way the fMRIDC has handled
issues of anonymisation, so as to conform with IRB demands.

More attention, on the policy level, to the interaction of regulatory bodies and data sharing initia-
tives can be very valuable here. 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 (see section Databases and IRBs above).
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.

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 (Anonymous, 2002). 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.

Making data ‘public’: WWW and networks as key infrastructures

The final point to be raised in this discussion is one which shines by its absence: the meaning of
making data public. In policy documents as well as in the controversy around the database, why the
World Wide Web is the best means either to share data or to make it publicly accessible regularly

seems to fall outside the boundaries of discussion – with some exceptions. The way this medium is taken for granted may lead to foreclosing the discussion about the range of possible ways of sharing data. For example, the fMRIDC currently sends out data on physical supports, but promises to make data available via the web in the near future. The Center is not the only body to state (rather than argue) the importance of the WWW.22

While the impact of the WWW on scientific communication is undeniable, it may not be desirable to make this an unquestioned assumption of data sharing, for two reasons. First, this assumption that THE WWW is the best means for sharing obscures the fact that the WWW itself may need to change to accommodate data sharing. For example, in the case of imaging data, the ability for networks to accommodate transmission of images is an issue that may require consideration. Furthermore, issues of accessibility may also be affected by reliance on networks rather than other kinds of support for digital data. In other words, if data sharing is to take place over the Web, the seamlessness of the interfaces between networks that make up the current Web, and possible changes to the functionality and architecture of networks (i.e. Grid and other variations) are also implicated in data sharing plans.

A second set of issues, regarding the impact on the data itself, is also obscured if the WWW is taken for granted as the means for data sharing. For example, the more ‘public’ a data set, the more important issues of anonymisation become. In brain imaging, this has often been resolved via the stripping or scrambling of data that could be used to reconstruct a subject’s face. These procedures, however, render the data unusable for some types of analyses. This is a clear case of a trade-off between making data available for greater circulation, and the value of the data. Other data sharing arrangements might involve a different type of trade-offs between these two desirable features: for example, requiring researchers to travel to a location to perform analyses on a secure system. Of course, there are many factors that will affect which solution is deemed most desirable, which are beyond the scope of a single case study. The purpose here is simply to signal that there may be a tendency to ignore the range of possibilities, and to posit public access via the WWW as the most

---

22 The Human Brain Project has now included, as part of its mission statement, the assumption that the resources developed will be accessible via the World Wide Web: “The goal is to produce new digital capabilities providing a World Wide Web (WWW) based information management system in the form of interoperable databases, and associated data management tools.” http://www.nimh.nih.gov/neuroinformatics/index.cfm (accessed 5 June 2002).
desirable form of data sharing, without attention to the interactions between the sharing process, notions of access and the nature of the data.

To end, a note on the theme of the research project that frames this case study: trust. The transitional phase in which functional imaging finds itself, in terms of separation of research and data management, shows how crucial it is that trust be maintained. If researchers perceive that data management is being valued over research, a backlash may occur. Similarly, new forms of data circulation may affect the public’s trust in science. Furthermore, there must also be trust in the value of having data separated from the context of production. To achieve this, trust building mechanisms, such as clear incentives as well as obligations, and regulations, are crucial. The care taken with commercial involvement in the EBI (also discussed in this volume) can also be seen in terms of this need to maintain trust. Trust of researchers in the data sharing infrastructure is maintained through carefully organising data management functions along the same principles as (publicly-minded) research. In other words, if data sharing is evaluated as being overly risky, or if the values implemented in data sharing do not build and sustain trust on the part of the producers of data, data sharing will not fare well. No matter how great the promised or potential benefits, trust must be maintained and nurtured in the field, if practices of data sharing are to develop.


Yencharis, L. (2002). Q & A With Len Yencharis, Cover Story: An interview with John Van Horn, Ph.D., M.Eng., a Research Associate Professor at the Center for Cognitive Neuroscience, Dartmouth College. *Advanced Imaging.*
The Web may not be the best means to make data accessible.