EUROPEAN SCIENCE FOUNDATION
SETTING SCIENCE AGENDAS FOR EUROPE

Deutsche
Forschungsgemeinschaft
DFG

Shared Responsibilities in Sharing Research Data: Policies and Partnerships

Report of an ESF-DFG workshop, Padua, 21 September 2007

www.esf.org

European Science Foundation (ESF)

The European Science Foundation (ESF) provides a platform for its Member Organisations to advance European research and explore new directions for research at the European level.

Established in 1974 as an independent non-governmental organisation, the ESF currently serves 77 Member Organisations across 30 countries.

Deutsche Forschungsgemeinschaft (DFG)

The German Research Foundation (DFG) is the central, self-governing research funding organisation that promotes research at universities and other publicly-financed research institutions in Germany. The DFG serves all branches of science and the humanities by funding research projects and facilitating cooperation among researchers.

Within the DFG, the Division “Scientific Library Services and Information Systems” has as its mission to establish effective information services and innovative information infrastructures suited to meet the needs of the research community at German universities and research institutions.

Slides available at:
www.esf.org/activities/science-policy/corporate-science-policy-initiatives/sharing-research-data-2007.html
Videos available at:
http://cabtube.cab.unipd.it/conferenze/berlin5-open-access/esf-workshop

Report published by the European Science Foundation, 2008.

Suggested citation:
ESF (2008). Shared Responsibilities in Sharing Research Data: Policies and Partnerships.
Report of an ESF–DFG workshop, 21 September 2007.

Report Editing: Alexis-Michel Mugabushaka (ESF) and Max Vögler (DFG)

ISBN: 2-912049-90-3

All views, conclusions and recommendations expressed in this report are those of the speakers. They do not necessarily reflect the views of DFG nor those of ESF or its Member Organisations.

Contents

Preface

On 21 September 2007, the European Science Foundation (ESF) and the German Research Foundation (DFG) organised a one-day workshop ***“Shared responsibilities in sharing research data”***. The workshop was held in the frame of the 5th follow-up conference of the Berlin Declaration on Open Access which took place at the University of Padua (Italy).

The Berlin Declaration on Open Access, signed by many European research organisations, has triggered a wide range of efforts and initiatives to facilitate access to research publications. Yet the Berlin Declaration on Open Access goes beyond scientific publications. It covers also “ .... raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material”.

It is against a background of growing consensus that enabling access to research data is an equally important task, and that a shared vision and sense of responsibility is needed among the stakeholders to make “open data” a reality, that this workshop was organised.

The objectives of the workshop were:

- to acquaint research organisations in Europe (primarily ESF member organisations) with on-going and planned initiatives for open access to research data;
- to present and discuss policies and practices on open access to research data of selected research funding organisations;
- to identify areas in which research organisations could collaborate on this issue.

The speakers, coming from the scientific community, funding organisations, data centres, and universities, met an equally diverse audience in a lively debate about the tasks which need to be undertaken and the challenges to be addressed in order to secure research data for the future generations of researchers.

We are very thankful to the speakers and the more than 80 participants who, through their contributions or questions and interventions, made this debate possible.

We are also grateful to the local host, the Conference of Italian University Rectors and the organising team headed by Ms Antonella De Robbio, University of Padua, for the effective and the valuable support provided during the workshop.

We hope that this report shows the breadth and various angles of issues raised and captures the essence of the recommendations made. We believe its input will guide future efforts in sharing research data.

Dr. John Marks

Deputy Chief Executive
Director of Science and Strategy
ESF

Dr. Beate Konze-Thomas

Head of Department
Research Programmes and Infrastructure
DFG

September 2007

Introductory Session – Sharing Research Data: benefits and responsibilities

Introductory session

Sharing research data: what can research funding agencies do?
Max Vögler

Sharing research data: a shared responsibility
John Marks

After welcoming the participants, John Marks (ESF) and Max Vögler (DFG, on behalf of Beate Konze-Thomas, who could not attend the workshop) presented the structure of the workshop and provided a background overview of the main issues to be discussed.

Sharing research data: what can research funding agencies do?

Max Vögler

Research data, in the variety of formats in which they are collected and stored, are the very fundament on which research is built. **Dr. Vögler** called attention to the growing consensus around the potential benefits of sharing research data:

- Research data are the "infrastructure of science": their re-analysis helps validate or correct previous results and, in an interdisciplinary setting, can also open up new research avenues well beyond the initial context in which the data were collected;
- Sharing research data ensures the efficient use of (public) funds and resources: the unnecessary (re-) collection of data is minimised and data collection becomes a collective exercise;
- Sharing research data is also a reliable way to safeguard research integrity. It counteracts misconduct related to data fabrication and falsification;
- Replication studies are also a powerful means of training new generations of researchers.

Advances in information and communication technology (ICT) have made it easier to handle and store large amounts of research data and to share them across the globe. Yet considerable effort is still needed to use this potential and reach the goal of "open access to research data".

This does not entail making all data freely available to everyone at all stages of the research process – there are many legitimate reasons for restricting access to data – but to encourage a system whereby quality-ensured data sets can be easily re-used and verified for legitimate purposes.

Funding agencies can play a crucial role in such a process.

- They can encourage the open verification of data sets as a part of the "good research practice" they promote.
- They can fund projects which define good practices in sharing research data and develop tools and techniques to facilitate it.
- They can define effective data sharing policies for the projects and institutions they fund.

The presentation concluded with examples of selected DFG initiatives to promote the sharing of research data.

Box 1: Selected DFG funded initiatives in sharing research data

The DFG-funded project STD-DOI has developed standards for publication and citation of scientific primary data in the Earth and Environmental Sciences. The project implemented a system making primary scientific data citeable by attributing data to its investigators (by means of the Digital Object Identifier).
www.std-doi.de

Established in 2007, a network of DNA banks is funded by the DFG. The network provides high-quality, long-term storage of DNA materials and maintains the online documentation of each sample. It also provides access to DNA materials (at self-cost). It was initiated by the Global Biodiversity Information Facility (GBIF).
www.dnabank-network.org

The HyperNietzsche Project (funded by the DFG among others) aims to provide open access to the digital archive of facsimile editions of Nietzsche's work. It also acts as a repository for critical editions of Nietzsche's texts.
www.hypernietzsche.org

Sharing research data: a shared responsibility? *John Marks*

A look at various initiatives and activities to foster sharing of research data shows that several types of organisations and stakeholders currently work to realise the vision of shared data resources. **Dr. Marks** showed three levels on which the various partners are addressing the underlying issues.

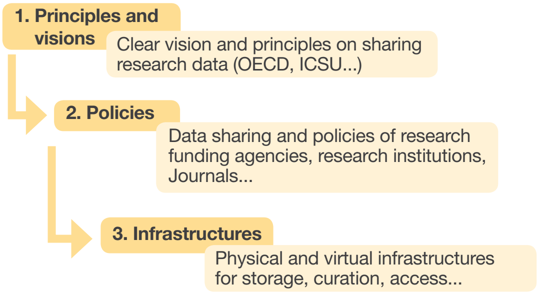

1. Principles and visions
Clear vision and principles on sharing research data (OECD, ICSU...)
2. Policies
Data sharing and policies of research funding agencies, research institutions, Journals...
3. Infrastructures
Physical and virtual infrastructures for storage, curation, access...

Figure 1: Three perspectives to address data sharing

At the level of **vision and principles**, we see that different organisations have taken up the task of developing a clear vision and communicating clearly the principles behind it. Notable examples are the OECD Principles and Guidelines for Access to Research Data from Public Funding and the ICSU report on Scientific Data and Information (see section 4.2 of this report).

At the level of **policies**, we see policies being developed by a growing number of the organisations, such as research funding agencies, journal publishers, research performing organisations and universities, which are developing data sharing (see section 4.4).

As the data also need to be physically stored and maintained, the third level deals with **research data infrastructures**. They include not only physical facilities to ensure the collection (submission), storage, curation and distribution but also "soft infrastructures" such as technical standards to ensure interoperability etc.. While some research fields have well-established facilities and standards (e.g world data centres for geosciences), there are sustained efforts to build new ones and overhaul the existing ones (see sections 2.1 and 2.3).

The three perspectives, principles and visions, policies and infrastructures, are complementary and they require cooperative efforts of various partners:

- the research community (and professional and learned societies);
- universities and research performing organisations;
- data sharing facilities;
- research funding agencies;
- international organisations;
- scientific publishers.

Partnerships between those stakeholders can help answer the three leading questions of the workshop:

(1) How can various stakeholders work together to realise the common vision of sharing research data (as articulated e.g. in the recent OECD Principles and Guidelines for Access to Research Data from Public Funding)?

(2) With increasing international collaboration and increasing complexity of science (and co-funding of research projects), how to ensure that researchers are not confronted with different regulations and that the policies are compatible with each other?

(3) How can the partners work together to ensure the sustainability of research data infrastructures?

The urgency of this last question can be illustrated by a relatively recent controversy over access fees for the Yeast Protein Database (YPD) (see Box 2)

Box 2 : Sustainability of business models: the case of the Yeast Protein Database (YPD)

A July 2002 *Nature* article reported the anger of the scientific community in life sciences as it learned that the Yeast Protein Database (YPD), a database in which it has been sharing data about protein structure and functions was changing its business model. From one day to another fees had to be paid to access those data. Most of the data has been generated by publicly-funded researchers and has been deposited in YPD in order to comply with requirements from publishers.

Referring back to this incident and similar ones, a May 2006 Editorial of *Nature Cell Biology* called for urgent action "*to select appropriate databases for funding on a stable, long-term basis (…) this selection should not be executed at the national level but rather in an international setting that reflects the origin of the research contained in the database*".

The Editorial proposed to set up an international database panel with the authority and resources to "*award indefinite funding to key community databases (… and) ensure that databases remain open access resources*".

Source:
Alison Abbot (2002) Biologists angered by database access fee. *Nature*, Vol. 418, July 2002, p. 357.
Sharing science, Editorial in *Nature Cell Biology*, Vol. 8, May 2006, p. 425.

1. Open Data: promises and untapped potentials [keynotes]

Keynotes

Avian Influenza :
Why Do We Need to Share What We Know?
Ilaria Capua

Open Data
Peter Murray-Rust

The workshop began with two keynote addresses from researchers, illustrating the importance of data sharing with concrete case studies from their respective research fields.

Using the example of the avian flu genetic data, Ilaria Capua made a strong case for data sharing even before publication in the interest of speeding up research progress in areas critical to human well-being.

Peter Murray Rust illustrated how removing current restrictions on accessing full text publications (in their raw form) will harness new research opportunities in chemistry.

1.1 Avian Influenza: Why Do We Need to Share What We Know? *Ilaria Capua*

In her keynote address, **Dr. Capua** began by describing how the avian influenza viruses of the H5N1 subtype have become widespread in vast areas of Asia, the Middle East, Europe and Africa. This opportunity given to the virus has greatly increased its potential to affect the health of wild and domestic animals, and humans. It has become a global threat for animal and human health and an issue of grave concern for food security (as it reduces the primary sources of protein to the undernourished population).

Currently, various measures, including education programmes and veterinary support to farmers, have been undertaken to curb its spread.

Great research efforts are also underway. Yet the medical, veterinary and agricultural scientific communities are challenged with a virus that is modifying itself as it adapts to different species and combines with other influenza viruses of avian and potentially mammalian origin, infecting new species along the way.

In order to understand how the virus evolves, it is crucial to compare the sequence data of newly-isolated virus strains with those already available. Yet those data are often not openly accessible because researchers tend to release them only after publication or for other intellectual property-related considerations.

It goes without saying that the delayed access to this crucial data undermines research progress in an area crucial to global human and animal health.

In 2006, Dr. Capua's laboratory at the Istituto Zooprofilattico Sperimentale delle Venezie (IZSVe) isolated the H5N1 African strain. It declined to deposit the gene sequences in a password-protected database to which only 15 laboratories had access and opted instead to deposit the full sequence in GenBank, a publicly-available database of genetic sequences.

Depositing the data in the restricted database would have meant two advantages for the IZSVe team. Data deposited in a closed system would have given the team a well-deserved "head start" in publishing articles related to the sequence. With the data freely available, any researcher worldwide now had access to the data. Second, the team would have gained access to the rich holdings of the restricted database as one of only 15 laboratories worldwide.

Dr. Ilaria Capua

For Dr. Capua, the international health risk associated with avian influenza far outweighed those considerations.

Her decision launched an avalanche of sympathy in the scientific community as well as the broader public. Not only did her decision become the subject of multiple articles in *Nature* and *Science*, it was also prominently featured in the *New York Times*, the *Washington Post* and other leading publications.

Dr. Capua then became one of the initiators of the GISAID initiative which aims to foster sharing of avian and other influenza virus sequences and related data. GISAID has the potential to become a model for other areas. (Box 3). As Dr. Capua stressed "do we really need a crisis of public health proportions in every single discipline to change the perception of data sharing issues?".

1. Open Data: promises and untapped potentials [keynotes]

Box 3: GISAID initiative

GISAID is the Global Initiative on Sharing Avian Influenza Data.

It is a platform on which researchers share their data (genetic sequences, clinical manifestations in humans, epidemiology, observations in poultry and other animals).

The usage agreement foresees that the data a researcher uploads are made accessible to other researchers who agree to the same terms of use.

The data can be used to publish results, provided the authors agree to collaborate with the data provider in further analysis and research. They can also be used to develop vaccines and other interventions.

Although initial angel funding was provided exclusively from the private sector, the initiative's sustainability is expected to be carried in part by public funding, either via grants or operating funds from governmental organisations, similar to the way previous influenza resources were funded.

By focusing on the social cost of not acting, Dr. Capua was able to generate a great deal of momentum in this individual case.

Dr. Capua closed her presentation by pointing out the many challenges the new GISAID initiative faces as it tries to become a mature initiative. One of the greatest challenges involves funding; especially obtaining sustainable funding is a major hurdle in creating such initiatives and databases.

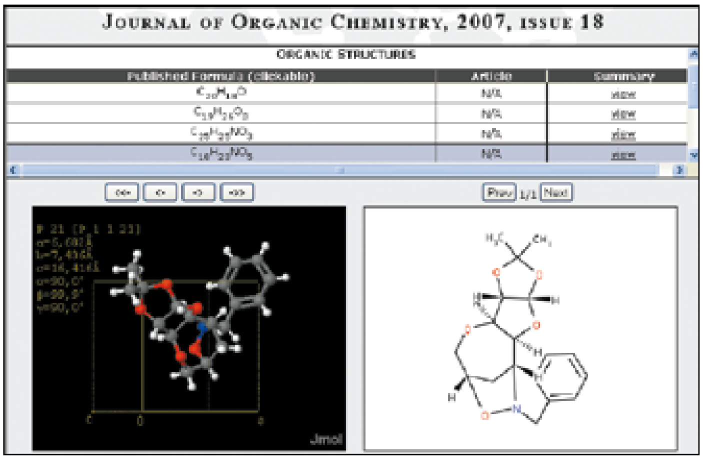
1.2 Open Data, *Peter Murray-Rust*

The concept of "Data-Driven-Science" describes a new line of inquiry whereby existing data – often found in publications - is used as a primary resource in driving scientific research.

Professor Murray-Rust gave an example from the field of chemistry in which data do not only refer to "laboratory data" but also to data descriptions which can be found in scholarly publications. Over one million new chemical compounds are published yearly, but these are scattered through hundreds or thousands of journals, monographs, papers and dissertations. By gaining access to such resources and subjecting them to a wide variety of data- and text-mining, researchers can ask new and different sorts of scientific questions.

He gave a demonstration of CrystalEye[1], a data-mining tool developed by his team. CrystalEye searches for crystal structures through published materials on a daily basis. It then reads, extracts, and aggregates them. Anyone can then access and visualise this information on a dedicated portal.

This and similar efforts are often actively hindered by a series of "permission barriers" put in place by most commercial publishers.



Journal of Organic Chemistry, 2007, issue 18

ORGANIC STRUCTURES

Published Formula (clickable)	Article	Summary
$C_{20}H_{18}O$	N/A	view
$C_{19}H_{26}O_5$	N/A	view
$C_{20}H_{23}NO_3$	N/A	view
$C_{18}H_{23}NO_5$	N/A	view

Figure 2: Screenshot from CrystalEye

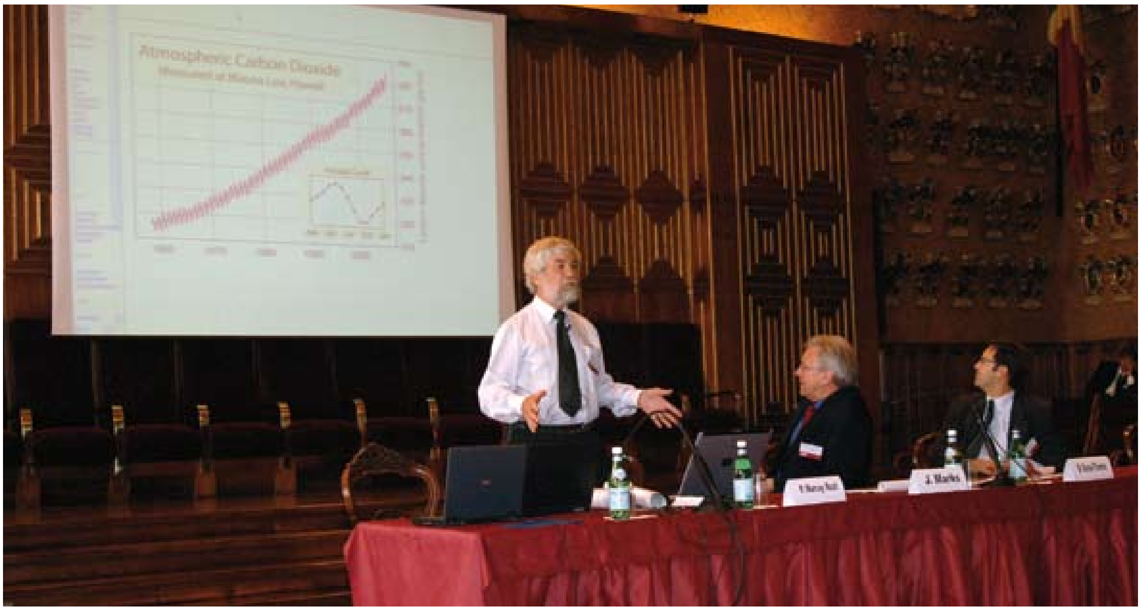
The restricted access to the research data occurs in a variety of ways.

First, publishers routinely block access to individual IP addresses that download or index large numbers of articles. Indeed, such procedures are often explicitly forbidden in existing access agreements.

Second, the format in which the data are shared matters. A PDF file is often an image of the text and may "destroy valuable information". This format might not be easily amenable to an analysis performed by text-mining tools. XML format would be better suited.

Third, scientists who make data available often forget to specify the terms. Even many scientists who publish and archive their findings via Open Access do not take the copyright agreements and rights clauses for their data seriously enough. Often, data-reuse clauses in copyright agreements are stringently guarded against commercial reuse. For data to be truly open, the terms of use need to be specified. A "fuzzy" copyright that seeks to discourage some forms of reuse while permitting others does not work well. All actors in the system – scientists, institutions, publishers, funding agencies, libraries – must pay more attention to creating the right types of usage agreements, ones

[1] http://wwmm.ch.cam.ac.uk/crystaleye/



Professor Peter Murray-Rust

with Open Data as their goal. Professor Murray-Rust recommends using licenses such as *Science Commons* to realise the vision of Open Data.

In this context, it is important to differentiate between access barriers and permission barriers. The former are essential for Open Access: the broadest public possible should have access to scientific information. When it comes to research data, it is not just a question of access but also of permission – not just if data can be accessed but how data can be used – that becomes essential. In fact the term "open access" is a weak tool when describing access to, and re-use of, data. The concept of Open Data is better in describing the need to consider data as a critical resource and in removing access and permission barriers.

Indeed, permission barriers are most frustrating when they remain ambiguous. When one scientist sees data he or she wants to use, sends an email or letter asking for permission and then hears nothing in return, what can they do? Can we define a moratorium on "data permission silence"?

There are other issues that also make the question of Open Data essential.

- Addressing the issue of heterogeneous data. Formats and standards to ensure interoperability are needed, especially in work involving data from "between" the disciplines.
- Developing mechanisms to reward those who share their data and work to make them available. Citation metrics credit only article publications and do not properly recognise the contribution to science made in the form of creating data sets or in developing tools to work with such data. The system of Scholarly Communication as a whole must begin to find ways also to credit researchers who put their energies into such forms of research output.
- Developing discipline-specific repositories: the research communities generally know what is best for their data.

As Professor Murray-Rust wrote on his blog during the conference, data is "a critical resource which needs political and legal activity"[2]. Achieving the goal of open data will require the joint effort of many stakeholders: young people as "they understand the future better than we do"; university leadership; and the research funding agencies.

He closed his address by remarking "We have to change the way we manage our scientific data, or we're simply not going to be using it for the benefit of humanity".

[2] http://wwmm.ch.cam.ac.uk/blogs/murrayrust//?p=608

2. Data Sharing: perspectives of key stakeholders [session 1]

Session 1

Making Data Accessible: suggestions from the scientific community
Gerold Wefer

International Initiatives in Data Sharing: recent developments (OECD, CODATA and GISCI)
Yukiko Fukasaku

Data Sharing Infrastructures in the ESFRI Roadmap: A Perspective from the Social Sciences and Humanities
Peter Doorn

Data Policy of Scientific Journals:
Nature perspective
Maxine Clarke

Sharing research data requires collaborative efforts from a variety of stakeholders. The first session of the workshop was devoted to learning from representatives of the research community, international organisations, research data centres and publishers on how they foster data sharing.

2.1 Making Data Accessible: suggestions from the Scientific Community, *Gerold Wefer*

Using his own research field, marine sciences, as an example, Professor Wefer offered the perspective of a researcher on the existing efforts to share data and the related challenges.

Sharing data and developing tools to utilise data is an established practice within the marine sciences community. In recent years, the benefits of long-term preservation and data sharing have been made evident by the importance of old data series in the understanding of current and future changes in the climate.

A concrete example of a data sharing facility in the geosciences area is the PANGAEA System.

PANGAEA (Publishing Network for Geoscientific and Environmental Data) acts as an openly accessible library for archiving, publishing and distributing geosciences data, as well as ensuring their long-term preservation.

The system is hosted by the Centre for Marine Environmental Sciences at the University of Bremen and the Alfred Wegener Institute for Polar and Marine Research in Bremerhaven, Germany. It has received funding from the DFG; the European Commission; the German Federal Ministry of Education and Research; and the International Ocean Drilling Program (IODP).

PANGAEA holds a variety of data from the various fields of earth sciences. They include time series of observations, sea bed photos, distributed samples, complex data, air photos and audio records. As of July 2007, PANGAEA had more than 500 000 data sets, totalling more than 1.8 billion data items.

The services of PANGAEA are:

(1) the maintenance of data infrastructures on which data are archived and made available;

(2) data management and data curation; and

(3) data publication through online distribution and data reports.

The PANGAEA System is used also by the World Data Center for Marine Environmental Sciences (WDC-MARE) as its archive and publication unit. The WDC-Mare is one of more than 50 existing World Data Centers (See Box 4).

Reflecting on good practices and critical factors for the success of data sharing systems, Professor Wefer formulated the following recommendations.

(1) In order to assure sustainability, data storage must be managed by established centres and systems that have a competent grasp of the necessary technical expertise.

(2) The acceptance of a data system stands or falls with the simplicity of locating the system, ease of access, and its content.

(3) The acceptance of data systems stands or falls with the simplicity of finding the data, ease of accessing data sets and, of course, the quality of its contents.

(4) The data must be accompanied by standardised descriptions, so that the user can evaluate their quality and source (no data without metadata, no metadata without data).

(5) Scientists are motivated to provide data if they are appropriately referenced. Every data set must include in its description a usable citation. The citation should include a permanent identifier that is presently in conventional use by established publishing companies (for example, Digital Object Identifier, see Box 1).

(6) Funding agencies, institutes and projects should formulate their data policies with appropriate explanations and regulations.

Box 4: World Data Centers

The World Data Centers system was created by the International Council for Science (ICSU) for the International Geophysical Year (1957/1958). Committees of scientists were asked to develop data management strategy for each of the geo-physical disciplines. Currently more than 50 centers exist in Europe, US, Russia, Japan, India, China and Australia.

The World Data Centers system is overseen by the Panel on World Data Centers of the ICSU.

Professor Gerold Wefer

Apart for the perennial issues of ensuring long-term preservation of the data, the main challenge the initiative faces is to develop sustainable systems and systems which are *"user-driven and user-controlled to avoid a technical end in itself".*

Professor Wefer concluded the presentation by calling research funding agencies to establish clear data sharing policies:

"We already have well-established data information systems. What we need are more data submitted to the centres".

2.2 International Initiatives in Data Sharing: recent developments (OECD, CODATA and GISCI),

Yukiko Fukasaku

One of the main objectives of the workshop was to acquaint participants with ongoing international initiatives to promote sharing of research data. **Dr. Fukasaku** from Innovmond s.a.r.l (previously at the OECD office) presented three international data sharing initiatives: OECD guidelines; CODATA activities; and Global Information Commons for Science Initiative (GICSI).

OECD Principles and Guidelines for Access to Research Data from Public Funding

The OECD recommendations on access to research data were developed in response to a request of the OECD ministers in 2004 to "develop (...) guidelines based on commonly-agreed principles to facilitate optimal cost-effective access to digital research data from public funding".

An Expert Group, tasked to draft the guidelines, launched a survey of existing practices and policies; organised expert workshops; and undertook a wider consultation of stakeholders.

The consultation showed that the institutional frameworks to facilitate access were still lacking and that the policies and practices in place varied considerably. The stakeholders consulted expressed support for international guidelines which in their view could provide guidance to institutions in need of policies and facilitate international research cooperation.

The OECD recommendations on sharing research data from publicly-funded research are articulated in 12 principles and guidelines which are reproduced in Box 5.

These recommendations are not legally binding but set collective standards or objectives that member governments can implement.

In 2004, governments of 34 countries and the European Commission committed themselves to work towards the establishment of access regimes for digital research data from public funding in accordance with these OECD principles and guidelines.

2. Data Sharing: perspectives of key stakeholders [session 1]

Box 5: OECD principles and guidelines for access to research data from public funding

Openness: access to research data for the international research community at the lowest possible cost;

Flexibility: take into account characteristics of different research fields, legal systems, cultures and regulatory regimes;

Transparency: information on data to be made available through the Internet;

Legal conformity: conform to the national legal requirements on national security, privacy, intellectual property rights;

Formal responsibility: promoting formal institutional practices pertaining to authorship, usage restrictions, financial arrangements, ethical rules, licensing terms, liability and sustainable archiving;

Professionalism: observe relevant professional standards embodied in the codes of conduct of the scientific communities involved;

Interoperability: pay due attention to relevant international data documentation standards;

Quality: adopt good practices for methods, techniques and instruments employed in the collection and archiving of data;

Security: pay attention to the use of techniques and instruments to guarantee the integrity and security of research data;

Efficiency: improving overall efficiency of scientific research by avoiding duplication of data collection efforts;

Accountability: evaluation of access arrangements by user groups, responsible institutions and funding agencies;

Sustainability: taking measures to guarantee long-term access to data.

CODATA Activities in Sharing Data

CODATA is one of the interdisciplinary Scientific Committees of the International Council for Science (ICSU). It was established in 1966 to promote and encourage compilation, evaluation and dissemination of numerical data in science and technology.

The objectives of CODATA are:

- to improve quality and accessibility of data, especially for developing countries;
- to facilitate international cooperation of data experts and researchers;
- to promote increased awareness of the importance of data sharing;
- to consider data access and IP issues.

It operates mainly through task groups and working groups to address relevant data issues. It also holds bi-annual conferences and edits the peer-reviewed journal CODATA *Data Science Journal*.

Examples of CODATA initiatives include:

- CODATA Task Group on the International Polar Year Data Policy and Management which aims to define data policy, strategy, and overall management approach for the International Polar Year (IPY) 2007-2008. It will also facilitate international cooperation for open data access among the IPY projects.
- CODATA guidelines for sharing data from the Global Earth Observation System of Systems (GEOSS).

CODATA also works with the Global Biodiversity Information Facility (GBIF) and Science Commons to define common-use licensing of scientific data products. GBIF is an international organisation which aims to facilitate free and open access to biodiversity data worldwide. It was established in 2001 following a recommendation of the OECD Global Science Forum.

Other ICSU panels dealing with data and information include the Federation of Astronomical and Geophysical Data Analysis Services (FAGS) and the World Data Centers (see Box 4).

Global Information Commons for Science Initiative (GICSI)

The GICSI is a multi-stakeholder initiative launched during the World Summit on the Information Society (Tunis 2005). Its objectives are:

Dr. Yukiko Fukasaku

- to improve the understanding and increase awareness of the costs and benefits of data access through research and analysis of good practices in data access and sharing in the existing initiatives;
- to identify and promote the adoption of successful policies and institutional and legal models for providing open availability on a sustainable basis;
- to encourage and coordinate efforts of the stakeholders in the world's diverse scientific community, particularly through the creation of "information commons" by defining conditions of "common use" licensing approaches.

The challenges in international organisations

Dr. Fukasaku concluded her presentation by highlighting two challenges that international efforts in data sharing have to address: (1) the level of awareness and adoption of "data sharing culture" and policies vary greatly between countries (and research communities); and (2) the positions in respect to intellectual property, data protection and security are also often different from country to country.

Box 6: ICSU strategic framework for "Scientific Data and Information"

In 2003, ICSU appointed a committee of independent experts to define the overarching mission and role of ICSU in the area of Scientific Data and Information and to propose a strategic framework for this area.

The report "Scientific Data and Information", published in 2004, recommends, among other things, that "financial support for data and information management" become "a routine component in all research budgets" and "the evaluation criteria for assessing research funding proposals should include evaluation of data management".

Source:
International Council for Science (2004) ICSU Report of the CSPR Assessment Panel on Scientific Data and Information

2.3 Data Sharing Infrastructures in the ESFRI Roadmap: A Perspective from the Social Sciences and Humanities, *Peter Doorn*

Data archives in the social sciences and humanities already have a long tradition in Europe. The first data archives in the social sciences were set up in the 1960s. Similar efforts in the humanities followed in the 1970s and 1980s.

Peter Doorn, Director of the Dutch Data Archiving Networked Services (DANS), presented information on efforts to foster European collaboration among research data centres in the humanities and social sciences. He began by presenting several networks that currently exist at the European level.

The Council of European Social Science Data Archives (CESSDA) is an umbrella organisation of social sciences data archives in 20 European countries. It aims to promote the acquisition, management and distribution of data and their integration throughout Europe.

CESSDA[3] has been involved in the development of an integrated data catalogue, in developing transborder data agreements, and in defining metadata standards and tools such as Data Documentation Initiative (DDI), an XML-based tool for the description of survey.

It has also facilitated two EC–funded projects. The NESSTAR (Networked Social Science Tools and Resources) project[4] resulted in a user-friendly tool for publishing, sharing, viewing and downloading data over the Web. The Madiera (Multilingual Access to Data Infrastructures of the European Research Area) project[5] developed a portal to access data from 13 social science data archives across Europe.

European organisations also have a long tradition in cooperation in large-scale surveys such as the European Social Survey (ESS), the European Values Study, and the European Election Studies.

In the humanities, the development of international data sharing infrastructures started later and was less comprehensive than in the social sciences, but research data centres in the humanities span a wide range of research fields and include various data types:

Language and text archives

- Oxford Text Archive
- Lexical corpora
- Libraries, e.g. digitised newspaper collections
- Linguistic research centres

humanities (ssh) in Europe
– Social Sciences: the Council of European S
Science Data Archives (CESSDA)
– Humanities: networks of data repositories
hubs in various fields
– Towards permanent European
infrastructures for the ssh
– The role of the ESFRI-Roadmap as a "
stone"
DANS

Dr. Peter Doorn

Archaeological data archives

- Archaeological Data Service
- E-Depot of Netherlands Archaeology

Historical data archives

- History Data Service
- Netherlands Historical Data Archive
- Historical Data Hubs
- The Integrated Public Use Microdata Series International (IPUMS)

Although initial steps in international data sharing in the social sciences and humanities have been taken, many challenges still exist. Despite notable achievements, existing infrastructures are primarily national and European activities have been, until now, funded on a project basis and carried out as voluntary activities by national centres.

Although much has been accomplished in this way, stable, truly pan-European data infrastructures for the social sciences hardly exist.

There is a growing awareness of those problems. ERA-NETS are addressing the data management issues in their action plans and the European Strategy Forum on Research Infrastructures (ESFRI) has just launched its first Roadmap. ESFRI was

[3] www.nsd.uib.no/cessda/
[4] http://nesstar.com/
[5] http://www.madiera.net/

created to support a coherent strategy of Research Infrastructures policy in Europe. The Roadmap identified needs for new or major upgrades of pan-European Research Infrastructure in all areas (RI).

Six of the 35 proposed Roadmap initiatives are in the domain of the social sciences and humanities and deal– in the widest sense – with data sharing:

- DARIAH: Digital Research Infrastructure for the Arts and Humanities – to support digital access to all surviving humanities and cultural heritage information for Europe, and its preservation in the long term;
- CLARIN: Common Language Resources and Technology Infrastructure;
- EROHS: European Research Observatory for the Humanities and the Social Sciences - aiming to make language resources and technology available to scholars of all disciplines, in particular the humanities and social sciences;
- ESS: European Social Survey;
- SHARE: Survey of Health, Ageing and Retirement in Europe; and
- CESSDA: Council of European Social Science Data Archives.

Box 7: Selected data repositories recommended by *Nature* Journals

Nature Journals require authors to make data and materials available in a publicly-accessible database. Only where no such databases exist are the authors requested to provide the data and materials to readers directly. The following are some of the data repositories recommended by *Nature* Journals (by type of data and materials types).

Protein or DNA sequences and molecular structures: Genbank/EMBL/DDBJ, Protein DataBank, SWISS-PROT.

Structures of biological macromolecules: Protein DataBank, Nucleic Acids Database or Biological Magnetic Resonance Databank.

Proteomics data sets: the International Molecular Exchange consortium, PRIDE, IntAct, PeptideAtlas; Tranche, and the Global Proteome Machine Organisation.

Microarray data: GEO and Array Express databases.

Mutant strains and cell lines: Jackson Laboratory, Mutant Mouse Regional Resource Centers, American Type Culture Collection, UK Stem Cell Bank.

Dr. Maxine Clarke, *Nature* Journal

2.4 Data Policy of Scientific Journals: *Nature* perspective, *Maxine Clarke*

An inherent principle of publication is that others should be able to replicate and build upon the authors' published claims.

Dr. Clarke, editor at the journal *Nature*, presented the policy on availability of data and materials of the Nature Publishing Group (NPG).

In a nutshell, the policy states that authors should retain all original materials, data and associated protocols; they should provide evidence of deposition in recognised repositories on submission of articles and be prepared to provide any additional data that referees (pre-publication) and readers (post-publication) may require. The various *Nature* journals recommend specific databases or repositories for the data types in their respective fields (see Box 7).

The policy also requests that the section of the manuscript describing the methods include details of how materials and information may be obtained, including any restrictions that may apply.

Nature Journals also require that any supporting data sets for which there is no public repository be made available to any interested reader after the publication date from the authors directly. They can also upload it to the *Nature* internet site.

If, after publication, readers encounter a persistent refusal by the authors to comply by making their data and materials available, they can contact the chief editor of the *Nature* journal concerned who may refer the matter to the author's funding institution and/or to publish a statement of formal correction, linked to the publication.

Dr. Clarke also highlighted major challenges in the practical implementation of those policies. From the data repositories side, the main challenges are:

- some fields or tools have no community-accepted repositories;
- there seem to be no standards for structured submissions of data and materials (beyond what the data's creator needs for his or her own purposes);
- tools for analysis of data need to be user-friendly and open to depositors and browsers alike;
- the knowledge environment surrounding the data needs upkeep (incl. ontologies, analytical tools, links to other environments).

Dr. Clarke sees a critical role for research funding agencies in facilitating data sharing. They are recommended:

- to provide reliable support for data and material repositories;
- to support educational efforts;
- to devise incentives to credit data quality and sharing;
- to encourage efforts to develop universal unique identifiers for data and researchers, useful for authors, data generators, readers and journals.

3. Data Sharing: perspectives of research funding agencies [session 2]

The second session of the day was devoted to the perspectives of research funding agencies that have implemented data policies. The overarching questions were: what sorts of policies have been implemented ? What do these policies entail? What are the agencies doing to support data sharing efforts? Answering some of these questions were speakers from the National Institutes of Health (NIH, USA), the Natural Environment Research Council (NERC, UK), the National Science Foundation (NSF, USA) and the Unit "GÉANT and e-Infrastructures" (European Commission).

Session 2

Implementing the NIH Data Sharing Policy: expectations and challenges
Belinda Seto

The Data Sharing Policies of UK Research Councils: principles and practices
Mark Thorley

Preserving and Sharing Research Data: NSF Data Strategic Vision and US Interagency Working Group on Digital Data
Chris Greer

European Commission Support to Research Data Infrastructures
Carlos Morais Pires

3.1 Implementing the NIH Data Sharing Policy: expectations and challenges, *Belinda Seto*

Progress in scientific research depends on the free flow of information, and the exchange of ideas and knowledge, explained **Belinda Seto** from the US National Institutes of Health (NIH). Restricting information flow, which is the bedrock upon which future studies are dependent, can impede the advancement of research. Following this idea, NIH issued a policy that reaffirms the principle of wide data accessibility: *"Data should be made as widely and freely available as possible while safeguarding the privacy of participants, and protecting confidential and proprietary data."*[6]

The policy makers expect researchers who are funded by the NIH to make available final research data, especially unique data, for research purposes to qualified individuals within the scientific community. In implementing this policy, NIH is very mindful of the need to protect the privacy of individuals who participate in experimental studies and the confidentiality of data.

Data sharing can be accomplished through a number of methods. The most common method is publishing articles in scientific publications. Researchers also share data through an informal channel, by responding directly to data requests. However, when a large amount of data needs to be shared, an efficient approach is needed, generally through establishing a network of databases. One should recognise that there are challenges to creating successful networks, which may include fundamental differences in informatics infrastructure and communication tools used at various research sites. Solutions will entail standards for data collection, processing, and archiving to allow interoperability among databases and the ability to query data across databases.

The NIH encourages researchers to reuse data found in repositories to develop patents and commercial products.

The NIH requires that all projects proposals requesting more than 500 000 USD submit a plan for data sharing (or to state why data sharing is not possible). This should include statements as to the accessibility, standards and sustainability of research data. The NIH offers both a number central data repositories for research data (see Box 8) and also encourages institutions hosting the grant holders to provide such repositories. Quality control of the data occurs locally, before data is submitted to a repository, and must include the certified de-identification of research participants in clinical trials.

Dr. Belinda Seto

[6] NIH Data Sharing Policy: http://grants.nih.gov/grants/policy/data_sharing/

Box 8: NIH Central Data Repositories

NIH maintains a variety of data repositories which are openly accessible to the research community.
They include:

GenBank: a database of nucleotide sequences from over 160 000 organisms. The GenBank, DDBJ (DNA Data Bank of Japan), and EMBL (European Molecular Biology Laboratory) databases share data on a daily basis.

RefSeq (Reference Sequence): a source for well-annotated sets of sequences (incl. genomic DNA, transcripts, and proteins).

PubChem: database for chemical structures of small organic molecules and information on their biological activities.

Source:
Resource guide of the National Center for Biotechnology Information
http://www.ncbi.nlm.nih.gov/Sitemap/ResourceGuide.html
or the list of databases and electronic resources
http://www.nlm.nih.gov/databases/

In general, NIH differentiates between “open access data” and “controlled access data”. However, there are cases in which more “layers” of access are necessary. For example, in the longitudinal study of adolescent health (Project AddHealth) there is a basic and accessible layer of “public use data”, which contains only a subset of cases from a given trial. A second layer consists of “restricted-use contractual data”, which is available to a core group of researchers and institutions that have agreed to certain confidentiality conditions and continually pay a fee to cover the costs of providing data and user support to the data set. Finally, a third layer consists of “cold room data” – meaning that data can be accessed only onsite. This example shows that principles need to be balanced against privacy and sustainability issues and that solutions should be tailored to fit the specific situation. To date, the AddHealth data have been used in more than 1 000 documents (reports, journal articles, theses).

Dr. Seto explained that the NIH encountered a number of challenges while implementing a data policy. The first consisted of the difficulty involved in formulating a policy that can keep up with technical developments. A second challenge is security. Security requirements continually change and any system that contains sensitive or confidential data must be regularly updated. Small institutions are particularly handicapped by the problem of creating secure environments for restricted data. Then there is the issue of specific media types. Images especially come in a wide range of formats and ensuring that data work flow guarantees some degree of open access, preservation and interoperability remains a challenge.

3.2 The Data Sharing Policies of UK Research Councils: principles and practices, *Mark Thorley*

Mr. Mark Thorley

The second talk in the session was given by Mark Thorley from the Natural Environment Research Council (NERC) on “Data Sharing Policies of the UK Research Councils: principles and practices.” As emphasised by Mr Thorley, data is an inherent part of the scientific record. It should be maintained to allow reproduction and validation of research results, and thus be seen as a publication in its own right. This especially applies to large-scale or long-term studies, the results of which can be used by many researchers in almost as many disciplines. The high cost and broad scientific need for such studies means that there is an inherent public interest in ensuring that such data sets are openly available and sustainable.

For data policies, one size does not fit all. The data policies of NERC and of other research councils in the UK (see table 1) have evolved over time and there are key differences. All the UK research councils recognise certain principles associated with research data: that it is a valuable, long-term public good; that data sharing can improve opportunities for data exploitation; that grantees should have an acknowledged right of first use; and that data management is essential for success. However, the research councils differ greatly in how they implement these principles. Some provide for a central national facility (such as the UK Data Archive in the Social Sciences) whereas others delegate infrastructure needs back to the community.

Table 1 : Data sharing policies of the UK Research Councils

ESRC - Economic and Social Research Council

Formal data policy – currently being updated.

- Joint JISC- & ESRC-supported UK Data Archive, including the Economic and Social Data Service.
- Applicants must carry out a data review to ensure funds are not requested for data that are already available. Data must be offered to the archive within 3 months of end of award.
- Developing ‘new thinking’ in data management and sharing. For example, QUADS: Qualitative Archiving and Data Sharing Scheme.

NERC - Natural Environment Research Council

Data policy handbook and guidance. New version under development.

- All data must be offered to a NERC data centre to enable long-term management and re-use.
- Recognition of rights of investigator teams.
- NERC supports 7 data centres for long-term management of environmental data.

BBSRC - Biotechnology and Biological Sciences Research Council

Data sharing policy and implementation guidelines.
Endorsed by Council July 2006, apply from April 2007.

- Applicants must produce a data sharing plan. Data sharing encouraged in all research areas where there is a strong scientific need and it is cost effective to do so.
- Funds can be requested to support data management and sharing activities.

MRC – Medical Research Council

Data sharing and preservation policy – applies to new grants awarded from January 2006.

- Applicants must produce a plan for data sharing and preservation and include costings in grant applications.
- Implementing data management facilities at MRC- owned centres (as part of corporate responsibility for data).

AHRC - Arts and Humanities Research Council

De facto policy - detailed in funding guidance.

- Any significant electronic resources or datasets created as a result of research funded by the AHRC must be made available in an accessible depository for at least three years after the end of the grant.
- Can request resources to support management and sharing.
- Archaeology – special case. Must use the AHRC-supported Archaeology Data Service.

EPSRC - Engineering and Physical Sciences Research Council

No formal policy, does not overly intervene in the research dissemination process.

- Encourages PIs to manage primary data as the basis for publications securely and for an appropriate time in a durable form under the control of the institution of their origin.

STFC - Science and Technology Facilities Council

Policies to develop following merger of PPARC and CCLRC.

- Facilities (i.e CCLRC) – well-developed policies and facilities on a per-project basis.
- Grant holders (i.e PPARC) – Data curation policy agreed in principle.

Each approach – centralised or delegated – has its advantages and disadvantages. National centres of course offer longer-term support for research data and can serve as single points of contact. They are, however, also comparatively costly, often fall behind in supporting the newest developments (standards, access management, etc.), and they risk getting "further away" from the scientific communities they serve. Delegated infrastructure, on the other hand, is "closer" to the science and thus can be more responsive to scientists' needs. Such a delegated approach risks the lack of a clear, long-term vision, firm mandate and sustainable financing.

In closing, Mr. Thorley pointed to a number of issues that he sees funding councils grappling with as they implement data policies. "Formulating a policy", he said, "is the easy bit."

- How involved should a Research Council be in setting up and managing data centres? How much of the driving force must come from the community in order for such a centre to be an accepted part of the community?
- How can Research Councils monitor whether the policies they have mandated are actually upheld by grantees and data centres?
- What regular consulting structure with the scientific community has the Research Council put in place?
- What sort of accompanying (funding) programmes have been put into place to ensure capacity building of data management and curation skills?
- How can a funding council help to make sure that data sets are found?
- Long-term commitment to data access: any long-term strategy that depends on PIs to ensure data management and sharing is headed for trouble.

3.3 Preserving and Sharing Research Data: NSF data strategic vision and US Interagency Working Group on digital data, *Chris Greer*

Dr. Chris Greer

In his presentation, **Chris Greer**, who was recently appointed as director for the National Coordination Office for Networking and Information Technology Research and Development (NITRD, USA), focused on "Preserving and Sharing Research Data: NSF Data Strategic Vision and US Interagency Working Group on Digital Data." Dr. Greer began with some central figures to highlight the role of data. Data production is growing exponentially and researchers are doing more and more of their work online. Almost 80% of legal research occurs online, and 36% of the resources humanities scholars use in their research are available online[7]. "We are in the digital era and are playing a game of catch-up," Dr. Greer emphasised. "Data are routinely deposited in a well-documented form, are accessible to specialists and non-specialists alike, and are properly protected while being reliably preserved."

For Dr. Greer, there are a number of central questions that a scientific community must ask itself – nationally, internationally, discipline specific – when confronting the issue of data curation and

[7] http://www.clir.org/pubs/reports/pub126/pub126.pdf (page 12)

preservation. All of these issues need a good deal of further research, which should be coordinated internationally:

- What should be saved, for how long, and who decides?
- What means can be used to assess the value of data whose most important use may lie years in the future for unforeseen purposes?
- For preservation of software (including models and simulations): When is emulation adequate and when is preservation of the original hardware and operating systems required?
- What information (metadata) about data processing, filtering, transformations, workflows, and other manipulations should be saved and how can it be linked reliably to the original data?
- What physical media are the best choice for digital archiving?

Dr. Greer, like the other speakers, highlighted the often complex legal situation that surrounds data. The NSF, for example, allows grantees to retain the principal legal rights to intellectual property developed in a grant. However, the NSF also balances these rights with an obligation it passes on the researchers to "make results, data and collections available to the research community." To this end, the NSF includes language in its Grant Proposal Guide expecting grantees to "share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work."[8] Dr. Greer clarified that cost sharing should operate "in dissemination mode, not cost recovery mode" and thus researchers should not charge more for access than the cost of transporting the data.

Dr. Greer also emphasised that when it comes to Open Data, one size does not fit all. Different communities will have different needs relating to data policies, data preservation and data access.

In closing, Dr. Greer briefly mentioned the recently established US Interagency Working Group (IWG) on Digital Data, which includes representatives from 27 US Government Departments and Agencies, an unusually high number for such a working group. The working group has the task of developing and implementing "an interoperable framework to ensure reliable preservation and effective access to digital data."[9] This will be quite a challenge and he looks forward to also working together with European partners to tackle many of the issues raised.

Box 9:
Examples of data sharing policies in US

Science of Science Policy
http://www.nsf.gov/pubs/2007/nsf07547/nsf07547.htm

Earth Sciences
http://www.nsf.gov/geo/ear/EAR_data_policy_204.pdf

Social and Economic Sciences
http://www.nsf.gov/sbe/ses/common/archive.jsp

Polar Programs
http://www.nsf.gov/publications/pub_summ.jsp?ods_key=opp991
Long Term Ecological Research (LTER)

Community Data Policy
http://www.lternet.edu/data/netpolicy.html

Ocean Sciences
http://www.nsf.gov/pubs/2004/nsf04004/nsf04004_1b.htm

[8] NSF Grant Proposal Guide NSF 04-23, p. 50. <http://www.nsf.gov/pubs/gpg/nsf04_23/nsf04_23.pdf>
[9] WG Terms of Reference, January 2007, <http://iwg.cfa.harvard.edu/twiki4/pub/IWGDD/IwgddTermsOfReference/>

3. Data Sharing: perspectives of research funding agencies [session 2]

3.4 European Commission Support to Research Data Infrastructures,

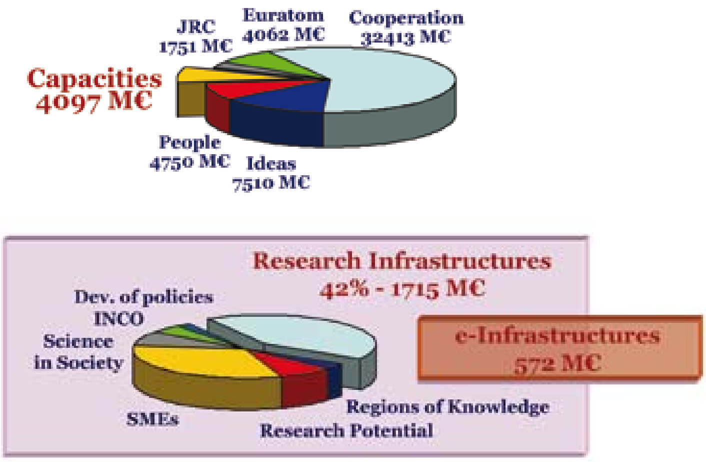
Carlos Morais Pires

In the final talk in this section, **Carlos Morais Pires**, head of sector "scientific data infrastructures", in the unit "GÉANT and e-Infrastructures" of the DG Information Society, presented the activities of the European Commission (EC) to support data sharing facilities.

Within the EC, scientific data sharing infrastructures are supported mainly in the "e-Infrastructures", scheme of the "capacities programme" of the Seventh Research Programme (FP7) 2007-13. The "capacities programme" has a budget of 4 billion € (8 % of the total FP 7 budget), out of which 42% (1.7 billion €) is devoted to research infrastructures in general. Within this total, "e-Infrastructures" account for 572 M€ (about 1.1 % of the total FP 7 budget).

Dr. Morais Pires

Chart 1: EC support for e-Infrastructures in the FP7



The e-Infrastructures scheme supports, among others, the development and deployment of scientific data infrastructures through commissioned studies such as eSciDR (a study to guide the policy to drive forward the development and use of digital repositories in Europe) and through grants (following open, competitive calls).

Two calls were planned in 2007 and 2008 with 15 and 20 M€ respectively.

At the time of the workshop, the first call had been closed. From the experience of the first call, the following types of proposals ranked highest in the peer-review based selection of projects:

- Proposals that adequately address the challenge of managing growing (massive) amounts of experimental data (with quality-verification concerns);
- Repositories that are used to manage, harvest data and metadata (incl. simulation models/ software, experimental output, published papers by peers to strengthen existing and formulating new hypotheses);
- Scientific repositories that are used to make a wealth of information useful and re-usable for scientists and researchers in various communities.

Scientific infrastructures generating experimental data and information face a number of challenges. They range from validation and quality assurance to long-term preservation. Each community or institution will undoubtedly have the tendency to focus on its own requirements and start shaping its 'own' infrastructure.

What has to be discussed is the need to see beyond these frontiers and use e-Infrastructures to really put into practice economies of scale at infrastructural level.

Dr. Morais Pires offered for reflection some questions to be addressed in the next years in order to realise the vision of first-class research data infrastructure in Europe.

- What funding models can be applied to the maintenance of repositories, for their own efficiency, sustainability, and the preservation of content?
- What can be done to harmonise and simplify authentication and authorisation mechanisms across Europe to gain access to e-Science resources?
- Which mechanisms of incentives are needed to encourage data generators to deposit (and thus share) their data, and to provide good-quality metadata?

Box 10 shows some projects related to the research data infrastructures funded b ythe EC.

Box 10: Projects related to "research data infrastructures" selected following the first and second call of FP 7 – e-Infrastructures

GENESI-DR
(Ground European Network for Earth Science Interoperations - Digital Repositories)
A project to establish Earth Science digital repository access for European and world-wide science users.

Euro-VO AIDA
(European Virtual Observatory (EURO-VO) and AIDA (Astronomical Infrastructure for Data Access)
Aims to deploy an operational Virtual Observatory (VO) in Europe by networking observation facilities, data centres and technology centres.

METAFOR
(Common Metadata for Climate Modelling Digital Repositories
Aims to develop a common information model to describe climate data and the models that produce them.

NMDB
(Neutron Monitor Database)
Aims to manage high resolution data from Neutron Monitor Stations (to measure cosmic ray variations).

IMPACT
(IMproving Protein Annotation through Coordination and Technology)
Aims to create a database called "InterPro" which will bring together a vast array of resources which are used to search genomes and proteomes for "protein signatures" (these are entities used to recognise a particular domain or protein family).

DRIVER II
(Digital Repository Infrastructure)
Aims to network existing institutional repositories (for publications). One component will deal with the linking of publications to experimental or observational data on which they are based.

PARSE.Insight
(Permanent Access to the Records of Science in Europe)
Will produce a Roadmap which focuses on specific parts of the overall e-Infrastructure needed to support the long-term preservation of records of science.

4. Discussions and Conclusions

The objectives of this one-day workshop, attended by more than 80 people, were threefold:

- to acquaint research organisations in Europe with on-going and planned initiatives for open access to research data;
- to present and discuss policies and practices on open access to research data of selected research funding organisations;
- to identify areas in which research organisations could collaborate on this issue.

The workshop confirmed the potential benefits that can accrue from a wider culture of sharing research data. "Open Data" could advance science as interdisciplinary data re-use opens up new fields of analysis; it could enable a more efficient research process as data is not reproduced unnecessarily. The more open the data, the harder scientific misconduct (especially data falsification and fabrication) becomes. Sharing data could also help in the training of new generations of scientists through replication studies.

The workshop participants learned from international efforts to promote data sharing. Among the various initiatives which were presented and seen to complement each other, the ***"OECD Principles and Guidelines for Access to Research Data from Public Funding"*** clearly stand out. To date, more than 30 governments and the European Commission have committed to their implementation. The OECD guidelines and principles published in 2007 provide a robust frame for any other initiative to foster an open data culture.

Two funding agencies from the US, the NSF and the NIH, presented their data sharing policies. On the European side, the approaches of the research councils from the UK were presented.

From these presentations, it became clear that we have examples of good (if not best) practices in formulating research sharing policies of research funding organisations. These policies address some of the concerns raised, such as the handling of sensitive personal data, the rights of first use by those who collected data, and they have been tested in practice.

A major difference between the funding agencies' policies lies in whether they set up research data centres/repositories themselves or whether they "delegate" this task to other institutions. As Mark Thorley pointed out, each approach has its advantages and disadvantages which have to be carefully considered in the national context and in each research field.

In most research agencies though, there seems to be no explicit mechanism to regularly monitor the implementation of data sharing policies (and, if necessary, enforce them).

One of the main recommendations on this aspect is that research funding agencies formulate clear and firm data sharing policies. Gerold Wefer said pointedly "we have well-established data information systems (in marine biology). What we need are more data submitted to the centres".

Contributions from other stakeholders (researchers, data centres, publishing community) brought forth a wide range of issues which should be given due consideration in future debates on data sharing policies.

- The keynote addresses made clear that the concept of sharing research data should be broadened to include also other innovative approaches such as sharing data even before publication (as in the case of data on Avian Flu). A challenge is how to adequately spot and support such efforts.
- Ways should be found to suitably "credit" researchers who share the data they collected. Until "data publication" is properly acknowledged as a valuable contribution to the research community, it can hardly be expected that researchers will wholeheartedly implement data sharing policies. A case study on cancer research shows that clinical trials which make microarray data available are cited about 70% more frequently than clinical trials which do not. (Piwowar et al. 2007). Perhaps such findings may help persuade authors to share their research data.
- Managing, curating and making data accessible requires significant resources and a higher level of expertise in data handling. Professionally-run data centres are in the best position to deal with the complex issues involved. The workshop learned from the efforts of the European Commission to support research infrastructures. Through these efforts, several research data infrastructures were initiated. However, concerns about the viability of those infrastructures after the end of project funding should be taken seriously.
- Data sharing policies and research data facilities should "stay close to the science and the scientists". There is a risk of data centres becoming "a technical end" in themselves. Efforts should not be spared to get researchers involved (and to remain involved) and indeed to take the lead in the development of research

sharing policies and the deployment of research data facilities.

- The workshop made clear that – as in many other research policy areas –in data sharing policies and data facilities too, one size does not fit all. The data sharing culture greatly differs among the research disciplines and some have well-developed facilities and others do not. It was suggested that the situation in different research fields should be assessed (how prone are different communities to share data? which “established” data sharing facilities exist in which disciplines?). It is important also that data sharing policies be tailored to each research discipline to take into account its specificities.

The need for collaboration between the various stakeholders was referred to on several occasions in the presentation and the discussions. Two examples of the lines of discussions are given below:

- Will (or can) researchers deposit their data if no good infrastructure exists to handle such data? On the other hand, does a large and sustained investment in research infrastructure make sense if the data repositories remain empty for lack of researchers’ willingness to share their data? It is clear that investment in research data facilities should go hand in hand with funding policies which encourage data sharing.
- Scientific journals such as Nature have developed policies requesting sharing the data on which publications are based. At the same time, research funding agencies are also increasingly asking the researchers they fund to share their data. How to make sure that those policies are consistent, e.g. in terms of types of data they have to deposit and the repositories they should use?

To tackle the numerous issues identified during the workshop and to take forward the suggestions which came up in the discussions will require collaborative efforts from various stakeholders. The research community (and professional associations as well as learned societies); universities and research performing organisations; research funding agencies; scientific publishers; research data infrastructures; and international organisations (dealing with research policy), all have their share of responsibility in the promotion of an open data culture. But only if they take their collaborative responsibility to move this issue forward jointly will the goal of open data be realised.

5. References

The following is a list of documents used during the preparation of the workshop and in the preparation of this report or referred to by the speakers.

Abbot, A. (2002) Biologists angered by database access fee. *Nature*, Vol. 418, July 2002, p. 357.

ICSU (International Council for Science) (2004) ICSU *Report of the CSPR Assessment Panel on Scientific Data and Information.* Paris: ICSU.

Klump, J. et al. (2006) Data publication in the open access initiative. *Data Science Journal*, Vol. 5, 2006,pp. 79-83.

NAS (National Academy of Sciences) (1985) *Sharing Research Data*. Washington, DC: National Academy Press.

NAS (National Academy of Sciences) (2003) *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences.* Washington, DC: National Academy Press.

NAS (National Academy of Sciences) (2006) *Strategies for Preservation of and Open Access to Scientific Data in China. Summary of a workshop.* Washington, DC: National Academy Press.

Nature Cell Biology (2006) Sharing science. Editorial, *Nature Cell Biology*, Vol. 8, Nr 5, May 2006, p. 425.

NERC (Natural Environment Research Council) (2008). *eScience: Harnessing the power of the internet for the environmental research*. Swindon: NERC. http://www.nerc.ac.uk/research/programmes/escience

NSF (National Science Foundation) (2005) NSF *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century* www.nsf.gov/nsb/meetings/2005/LLDDC_draftreport.pdf

NSF (National Science Foundation) (2007) *Cyberinfrastructure vision for the 21st Century Discovery.* Washington, DC: NSF.
http://www.nsf.gov/pubs/2007/nsf0728/index.jsp

OECD (Organisation for Economic Co-operation and Development) (2007) *Principles and Guidelines for Access to Research Data from Public Funding*. Paris: OECD. www.oecd.org/dataoecd/9/61/38500813.pdf

Piwowar, H.A. et al. (2007) Sharing Detailled Research Data is associated with increased citation rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308.

RIN (Research Information Network) (2007) *Stewardship of digital research data: a framework of principles and guidelines. Responsibilities of universities and colleges, research institutions and research funders.* http://www.rin.ac.uk/projects-list

Vickers, A.J. (2006) Whose data set is it anyway? Sharing raw data from randomized trials. *Trials*, 7:15 (accessed through PubMed http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1489946)

Wouters, P. and Schröder, P. (eds). (2003) *Promise and Practice in Data Sharing.* Amsterdam: Networked Research and Digital Information. NIWI-KNAW.

Wouters, P. (2002) Policies on Digital Research Data: *An international survey.* Amsterdam: Networked Research and Digital Information NIWI-KNAW.

6. Appendix

6.1 Workshop agenda

EUROPEAN SCIENCE FOUNDATION
SETTING SCIENCE AGENDAS FOR EUROPE

Shared Responsibilities in Sharing Research Data
Policies and Partnerships
An ESF–DFG workshop in the frame
of the Berlin 5 Conference
Padua, Friday 21 September 2007

Deutsche
Forschungsgemeinschaft
DFG

Programme
The workshop is jointly organised by the DFG and ESF in the frame of the 5th follow-up conference of the Berlin Declaration on Open Access (Berlin 5 Conference). The focus of the workshop lies on policies and practices of research organisations on open access to research data.

The objectives of the workshop are:

- to get research organisations in Europe acquainted with on-going and planned initiatives for open access to research data
- to present and discuss policies and practices on open access to research data of selected organisations
- to identify areas in which research organisations could collaborate on this issue.

9.30 - 11.00 : Introductory Session	
09.30 - 10.00	Welcoming Address by **Beate Konze-Thomas** (DFG) and **John Marks** (ESF)
10.00 - 10.30	Avian Influenza: Why Do We Need to Share What We Know? **Ilaria Capua**
10.30 - 11.00	Open Data: **Peter Murray-Rust**
11.00 - 11.30 *Coffee break*	

11.30 to 13.00 Session 1 : Perspectives of Key Stakeholders **Chair : John Marks**	
11.30 - 11.50	Making Data Accessible: Suggestions from the Scientific Community **Gerold Wefer**
11.50 - 12.10	International Initiatives in Data Sharing: Recent Developments (OECD, CODATA and GISCI) **Yukiko Fukasaku**
12.10 - 12.30	Data Sharing Infrastructures in the ESFRI Roadmap : A Perspective from the Social Sciences and Humanities **Peter Doorn**
12.30 - 12.50	Data Policy of Scientific Journals: Nature Perspective **Maxine Clarke**
13.00 - 14.30 *Lunch*	

14.30 to 15.20 Session 2 : Perspectives of Research Funding Agencies **Chair: Beate Konze-Thomas**	
14.30 - 14.50	Implementing the NIH Data Sharing Policy: Expectations and Challenges **Belinda Seto**
14.50 - 15.10	The Data Sharing Policies of UK Research Councils: Principles and Practices **Mark Thorley**
15.10 - 15.30	Preserving and Sharing Research Data: NSF Data Strategic vision and US Interagency Working Group on Digital Data **Chris Greer**
15.20 - 16.00	EC support to research data infrastructures **Carlos Morais Pires**

For further information on the workshop, please visit the Berlin 5 Open Access Follow up Conference website: http://www.aepic.it/conf/index.php?cf=10

Organising team:
Alexis-Michel Mugabushaka (ESF)
Max Vögler (DFG)

6.2 Biographies of speakers and organisers

Ilaria Capua

Dr. Capua is currently Head of the Virology Department at Istituto Zooprofilattico Sperimentale delle Venezie, Padova, Italy and Head of the National, FAO and OIE (World Organization for Animal Health) Reference Laboratories for Avian Influenza (AI) and Newcastle Disease (ND). During her career as a veterinary virologist she has been nominated OIE and FAO expert for AI and ND. She has been involved in managing several AI outbreaks on a global scale, and in particular has supported African countries affected by the H5N1 crisis. She is currently the chairman of OFFLU, the OIE/FAO Veterinary Network on Avian Influenza. In 2006 she launched the Global Initiative on Sharing Avian Influenza Data, endorsed by 70 medical and veterinary virologists and 6 Nobel laureates, which sparked an international consensus on sharing genetic information.

From 1997 to date Dr. Capua has been invited to give over 70 lectures as an international expert and as a guest lecturer at training courses in Europe, the US, Central and South America, Africa and Asia. She was awarded the Houghton Lecture Award in 2005 and the Promed 2006 Award.

From 1990 to date she has authored over 290 publications, predominantly on viral diseases of poultry including papers published in international refereed journals, papers and abstracts published in the proceedings of conferences, guest editorials, reviews, chapters of books and has co-authored an atlas and text on avian influenza.

Dr. Capua is currently coordinating two EU- funded projects, is a partner in an additional four projects and is involved in the EU Network of Excellence EPIZONE.

Maxine Clarke

Dr. Clarke is Publishing Executive Editor of the journal *Nature*, among other things responsible for editorial publication policies as well as author and referee services. She joined the staff of *Nature* in August 1984 as a subeditor. Since then, she has been a news reporter and editor, and has been the editor for the News and Views, Book Reviews, Brief Communications, Commentary and Correspondence sections of the journal. For some years, she handled manuscript submissions to *Nature* in the field of motor protein biophysics. Before joining *Nature*, Dr. Clarke did postdoctoral research at Kings College London (including visiting research at the Max Planck Institute for Biophysics in Heidelberg) and obtained a MA in Physiology and a PhD in muscle biophysics at Oxford University.

Peter Doorn

Dr. Doorn is Director of Data Archiving and Networked Services (DANS), the national centre for permanent access to research data for the humanities and social sciences. He studied human geography in Utrecht and defended his PhD there. He taught computing for historians at Leiden University between 1985 and 1997. He was Director of the Netherlands Historical Data Archive and Head of Department at the Netherlands Institute for Scientific Information Services (NIWI).

Yukiko Fukasaku

Dr. Fukasaku is Managing Director of Innovmond s.a.r.l., a research and consulting company based in France, specialising in scientific research and innovation issues. She founded the company in 2006 after leaving the OECD where she was a principal administrator in the Directorate for Science, Technology and Industry. During her career she has also worked for other international organisations, government agencies and research institutes including UNESCO, ILO, JETRO, NEDO, Centre de Sociologie de l'Innovation of the École Nationale Supérieure des Mines de Paris and Mitsubishi-Kasei Institute of Life Sciences as staff or as an independent researcher. She has published in the area of policy studies in science and innovation including research governance, data access, environment, energy and sustainable development. She has studied chemistry, history and social studies of science and holds a PhD in science and technology policy studies from SPRU, University of Sussex in UK.

Chris Greer

Dr. Greer is Senior Advisor for Digital Data in the Office of Cyberinfrastructure at the US National Science Foundation. He recently served as Executive Secretary for the Long-lived Digital Data Collections Activities of the National Science Board and is currently Co-Chair of the Digital Data Interagency Working Group of the National Science and Technology Council's Committee on Science.

Beate Konze-Thomas

Dr. Konze-Thomas is Head of Department "Research Programmes and Infrastructure" at the German Research Foundation (Deutsche Forschungsgemeinschaft). Her scientific background is in biology and chemistry. After PhD studies at the Ruhr-Universität Bochum she joined Michigan State University as a research associate, later returning to Munich University before moving to DFG. Her function includes responsibilities for coordinated and cooperative research programmes, for graduate school and for research infrastructure such as libraries, repositories and large scientific equipment.

John Marks

Dr. Marks is Deputy Chief Executive Officer of the European Science Foundation. Since January 2004, he was Director of Science and Strategy of ESF. Before joining ESF, he was the Director of Earth and Life Sciences at the Netherlands Organisation for Scientific Research (NWO). During 1992-1993 he was the acting Executive Director of the International-Geosphere Programme in Stockholm. From 1981 until 1998 he worked in the Netherlands' Ministry of Education, Culture and Science in various positions in the field of Science Policy. He has been involved in a wide range of science policy and management activities at the international level, both European and global, among others in the frame of the International Council for Science (ICSU). He holds a PhD in experimental low temperature physics from the University of Leiden, the Netherlands.

Carlos Morais Pires

Dr. Morais Pires is the Head of Sector of the area 'Scientific Data Infrastructure' in the unit dealing with e-Infrastructures in the EU 7th Framework Programme (Capacities Programme). He has been with the European Commission (Information Society Directorate General) since 1998, dealing with the development of a large number of R&D projects addressing Information and Communication Technologies. He has also been contributing actively to the Commission's bridging output from research into policy and regulatory activities. He holds a PhD in Electrical Engineering in the field of Telecommunication and Media Broadcasting technologies.

Peter Murray-Rust

Dr. Murray-Rust is Reader in Molecular Informatics at the University of Cambridge and Senior Research Fellow of Churchill College. He was previously Professor of Pharmacy in the University of Nottingham from 1996-2000, setting up the Virtual School of Molecular Sciences.

His interests have involved the automated analysis of data in scientific publications, creation of virtual communities, e.g. The Virtual School of Natural Sciences in the Globewide Network Academy and the Semantic Web. With Henry Rzepa he has extended this to chemistry through the development of Markup languages, especially Chemical Markup Language. Peter Murray-Rust holds a PhD in Chemistry.

Belinda Seto

Dr. Seto is the Deputy Director of the National Institute of Biomedical Imaging and Bioengineering (NIBIB). She is responsible for the oversight and management of all aspects of the Institute's research and training mission.

Dr. Seto earned her PhD in biochemistry at Purdue University in 1974. Following postdoctoral training at the National Heart, Lung and Blood Institute, she joined the Food and Drug Administration where she conducted research in virology for nearly 10 years. She received numerous awards for her research, including the Distinguished Alumni Award for Science from Purdue University, DHHS Secretary's Award for Exceptional Achievement, Inventor's Awards, NIH Director's Awards and she is listed in the American Men and Women of Science.

Mark Thorley

Mr. Thorley is the Data Management Co-ordinator for the UK's Natural Environment Research Council - NERC (www.nerc.ac.uk). He is responsible for coordinating activities relating to data management and scientific information strategy within NERC and has been in this post since 2002. Prior to this he was manager of the Antarctic Environmental Data Centre at the British Antarctic Survey (www.bas.ac.uk) between 1990 and 2002. He is a member of the Research Councils UK (RCUK) group which produced the RCUK Position Statement on Access to Research Outputs (see www.rcuk.ac.uk/access/default.htm) and he also represented the UK on the OECD working group which developed the Principles and Guidelines for access to digital research data from public funding.

6. Appendix

Gerold Wefer

Professor Wefer is a Professor of Marine Geology at Bremen University, Germany. After studying and working in the Geology Department at Kiel University and Scripps Institution of Oceanography he moved to Bremen in 1985. His main areas of research include: sedimentation processes in shallow water; ecology of benthic foraminifera; carbonate production in boreal and tropical seas; distribution of stable isotopes in calcareous organisms; particle flux (carbon and associated elements) in high latitudes and in the South Atlantic; paleoclimate in the South Atlantic. Professor Wefer has participated in over 30 major research cruises, many of which as chief scientist. Since 2001 he has been Director of the Ocean Margins Research Centre of the Deutsche Forschungsgemeinschaft. In 2001 Professor Wefer received the “Communicator Award”. The prize is given jointly by the DFG and the Donors’ Association for the Promotion of Sciences and Humanities in Germany (*Stifterverband für die Deutsche Wissenschaft*) and recognises scientists and academics who have communicated their research findings to the public with exceptional success.

Workshop organisers

Max Vögler

Dr. Vögler is a Program Officer in the Humanities and Social Sciences Division, and the Libraries and Information Systems division at the German Research Foundation (DFG). There he is responsible for the project Knowledge Exchange, an international cooperation between four European funding organizations in the information infrastructure area, and infrastructure policy development in the humanities and social sciences. He received his PhD in history from Columbia University in 2005.

Alexis-Michel Mugabushaka

Dr. Mugabushaka works as Science Officer for corporate science policy in the office of the Chief Executive of the European Science Foundation (ESF) in Strasbourg. In this capacity he oversees several science policy initiatives of the organisation. He held previously the positions of Officer for Statistics and Evaluation at the German Research Foundation (DFG) and Research Associate the International Centre for Higher Education Research (INCHER) at the University of Kassel, Germany. He has a PhD in Applied social sciences in the area of higher education and science policy studies.

EUROPEAN
SCIENCE
FOUNDATION
SETTING SCIENCE AGENDAS FOR EUROPE

1 quai Lezay-Marnésia | BP 90015
67080 Strasbourg Cedex | France
Tel: +33 (0)3 88 76 71 00
Fax: +33 (0)3 88 37 05 32
www.esf.org

Print run: 300 - December 2008 - Latitude :