Opportunities for Data Exchange

SEVENTH FRAMEWORK PROGRAMME

Baseline Report on Drivers and Barriers in Data Sharing

October 28th, 2011

Angela Schäfer[*a], *Heinz Pampel*[a], *Hans Pfeiffenberger*[a], *Sunje Dallmeier-Tiessen*[b], *Satu Tissari*[c], *Robert Darby*[d], *Krystina Giaretta*[e], *David Giaretta*[d, e], *Kathrin Gitmans*[a], *Heikki Helin*[c], *Simon Lambert*[d], *Salvatore Mele*[b], *Susan Reilly*[f], *Sergio Ruiz*[b], *Marie Sandberg*[c], *Wouter Schallier*[f], *Sabine Schrimpf*[g], *Eefke Smit*[h], *Max Wilkinson*[i], *Michael Wilson*[d]

[a] *Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany*

[b] *CERN, CH1211, Geneva 23, Switzerland*

[c] *CSC - IT Center for Science Ltd., P.O. Box 405, FI-02101 Espoo, Finland*

[d] *STFC Rutherford Appleton Laboratory, Harwell Science and Innovation Campus, Didcot OX11 0QX, United Kingdom*

[e] Alliance *for* Permanent Access, 2 High Street, *Yetminste, Dorset DT9 6LF, United Kingdom*

[f] *LIBER – Association of European Research Libraries, Koninklijke Bibliotheek, National Library Of The Netherlands, Po Box 90407, 2509 Lk The Hague, The Netherlands*

[g] *Deutsche Nationalbibliothek, Informationstechnik, Adickesallee 1, 60322 Frankfurt am Main, Germany*

[h] *The International Association of STM Publishers, Prama House, 267 Banbury Road, Oxford OX2 7HT, United Kingdom*

[i] *The British Library, 96 Euston Road, LONDON NW1 2DB, United Kingdom*

**Corresponding author: Angela.Schaefer@awi.de*

Opportunities for Data Exchange (ODE) is a FP7 project of members of the Alliance for Permanent Access (APA), which is gathering evidence to support the right investment in a layer for data sharing, re-use and preservation, in the emerging e-Infrastructure. The main concern of the ODE project is to talk with key players in the field. In dialogue with relevant stakeholders, views and opinions on challenges and opportunities for data exchange are collected and documented. To gain a broad and common understanding the ODE project collected 21 stories, in which significant stakeholders describe their experiences and their view on the topic. The focus is on the following four perspectives: scientific communities, infrastructure initiatives (e. g. data centres and data repositories), management (e. g. funding agencies and policy makers) and other relevant stakeholders (e. g. citizen science projects). This report provides an introduction, documents the stories and combines the key barriers and drivers for the permanent access to research data.

TABLE OF CONTENT

1. ODE PROJECT

The transition from science to e-Science is happening: a data deluge is emerging from publicly-funded research facilities; a massive investment of public funds into the potential answer to the grand challenges of our times. This potential can only be realised by adding an interoperable data sharing, re-use and preservation layer to the emerging eco-system of e-Infrastructures. The importance of this layer, on top of emerging connectivity and computational layers, has not yet been addressed coherently at the European Research Area (ERA)[1] or global level. All stakeholders in the scientific process must be involved in its design: policy makers, funders, infrastructure operators, data centres, data providers and data users, libraries and publishers. They need evidence to base their decisions and shape the design of this layer.

The Opportunities for Data Exchange (ODE)[2], a FP7 project, is gathering evidence to support the right investment in this layer for data sharing, re-use and preservation. ODE partners, all member of the Alliance for Permanent Access (APA),[3] collectively represent all these stakeholder groups and have a significant sphere of influence within those communities. The project is identifying, collating, interpreting and delivering evidence of emerging best practices in sharing, re-using, preserving and citing data, the drivers for these changes and barriers impeding progress. ODE will:

- Enable operators, funders, designers and users of national and pan-European e-Infrastructures to compare their vision and explore shared opportunities
- Provide projections of potential data re-use within research and educational communities in and beyond the ERA, their needs and differences
- Demonstrate and improve understanding of best practices in the design of e-Infrastructures leading to more coherent national policies
- Document success stories in data sharing, visionary policies to enable data re-use, and the needs and opportunities for interoperability of data layers to fully enable e-Science
- Make that information available in readiness for HORIZON 2020

1.1 Partners

ODE partners are:

European Organization for Nuclear Research (CERN): CERN, "where the Web was born", is funded by 20 European Member States with a budget of around 1,000 MCHF/yr.[4] CERN has 2,500 permanent staff and hosts some 10,000 HEP scientists from more than 250 institutes in 85 countries. CERN offers a unique complementary perspective of a producer of unique primary research data, as well as a major player in the design and construction of e-Infrastructures. CERN, a founding member of the Alliance for Permanent Access, is contributing to several FP7 projects relevant to the topic of data sharing.

[1] http://ec.europa.eu/research/era/index_en.htm

[2] http://ode-project.eu

[3] http://www.alliancepermanentaccess.orga

[4] http://www.cern.ch

Alliance for Permanent Access (APA): APA was set up as a non-profit organization, initiated as a Foundation under Dutch Law in 2008.[5] The goal of the Alliance is to align and enhance permanent information infrastructures in Europe across all disciplines. It is a networking organisation and a sustainable centre for advice and expertise on permanent access. The Alliance brings together seventeen major European research laboratories, research funders, and research support organisations such as national libraries and publishers. All its members are stakeholders in the European infrastructure for long-term preservation of and access to the digital records of science.

CSC, the Finnish IT Center for Science: CSC is a non-profit limited company whose shares are fully owned by Finnish state, and governed by the Finnish Ministry of Education. It is the largest national center in Northern Europe with a staff exceeding 200 (2011) providing modelling, computing and information services for academia, research institutes, the public sector and industry. CSC is also active in data management e.g., Radio and TV archive, national digital library and national long term storage) and maintains Funet, the Finnish University and Research Network, enabling fast connections between researchers. CSC has close connections to e-Infrastructure providers globally and represents Finland in key e-Infrastructure development projects.

Helmholtz Association: Helmholtz Association is with 33,000 employees in 17 research centres and an annual budget of approximately 3, 3 billion Euros, Germany's largest scientific organisation. Helmholtz research contributes to solving grand challenges in the fields of Energy, Earth and Environment, Health, Key Technologies, Structure of Matter, Aeronautics, Space and Transport. Helmholtz provides access to its infrastructures to researchers from all over the world. The development, construction and operation of large-scale facilities and complex infrastructures for data-intensive research is one of the Helmholtz Association's central tasks.

Science and technology Facilities Council (STFC): STFC is keeping the UK at the forefront of international science and tackling some of the most significant challenges facing society such as meeting our future energy needs, monitoring and understanding climate change, and global security. As a multi-disciplinary data producer, STFC has connections across a wide range of disciplines including space, earth observation, materials science and fundamental physics; in this role STFC also supports the work of many thousands of researchers across Europe. In terms of research infrastructures STFC plays a leading role in the development of e-Science in the UK and Europe.

The British Library: The British Library is one of the largest research libraries in the world.[6] It has a statutory responsibility to acquire, preserve and make accessible the UK national published archive. It holds over 150 million items ranging from historic manuscripts to modern electronic journals, digital music files and patents and is leading international collaborations to find solutions to ensure this rich and varied collection is sustained far into the future.

Deutsche Nationalbibliothek (DNB): DNB is the national library and national bibliographic information centre for the Federal Republic of Germany.[7] It is responsible for the collection, processing and bibliographic indexing of all German and German-

[5] http://www.alliancepermanentaccess.org

[6] http://www.bl.uk

[7] http://www.d-nb.de

language publications issued since 1913. The DNB is involved in several projects in the field of long-term preservation of digital data.

The International Association of STM Publishers (STM): STM has over 100 scientific publishers as members.[8] These range from the large international ones to a long list of small and medium-sized publishers. The mission of STM is to create a platform for exchanging ideas and information and to represent the interest of the STM publishing community in the fields of copyright, technology developments, and end user relations. By taking a role in digital archiving, STM fully endorses the commitment of the publishing industry to knowledge preservation.

The Stichting LIBER Foundation LIBER is the principal association of the major research libraries of Europe.[9] Its current membership includes 400 research libraries from more than forty countries, mainly but not only, in Europe. E-science and primary data are a priority in the LIBER Strategy 2009-2012. Within the area of scholarly communications LIBER concentrates its activity on Open Access and E-Science.

[8] http://www.stm-assoc.org

[9] http://www.libereurope.eu

2. DATA SHARING TODAY

This Chapter gives a short summary of the broad discussion of data sharing and describes the work of the ODE project documenting stakeholder's views on the challenges and opportunities of research data sharing.

2.1 Status Quo

Research data are valuable and ubiquitous. Research data are produced regardless of academic discipline e.g. in satellite missions by remote sensing, in text analysis in linguistics or in surveys in social sciences. The types and quantities of research data vary between the disciplines.

Since the Organisation for Economic Co-operation and Development (OECD) published their "Principles and Guidelines for Access to Research Data from Public Funding"[10] in 2007 the discussion about the permanent access to research data has grown in importance. Funders, scientific communities, libraries, data centres and publishers face the challenges and opportunities of data sharing.

In 2010 the European Commission established a High-Level Group on Scientific Data. The experts released the report "Riding the Wave: How Europe can gain from the rising tide of scientific data". The report describes long term scenarios and associated challenges regarding research data access and preservation as well as a strategy to realise the vision of a scientific data e-Infrastructure in 2030. In the introduction Neelie Kroes, European Commissioner for the Digital Agenda and Vice-Presidents of the European Commission, draws attention on the sharing of scientific data: "My vision is a scientific community that does not waste resources on recreating data that have already been produced, in particular if public money has helped to collect those data in the first place. Scientists should be able to concentrate on the best ways to make use of data. Data become an infrastructure that scientists can use on their way to new frontiers."

Over the last few years funders and science organization took up the discussion. To cite just one example: In Germany the "Alliance of Science Organisations" published in 2010 national "Principles for the Handling of Research Data". In this paper the science organisations "supports the long-term preservation of, and the principle of open access to, data from publicly funded research."[11]

There is an on-going discussion in the scientific community on the challenges of data sharing. Special issues of leading scientific journals like Nature[12] and Science[13] showing the relevance of the topic. In some disciplines learned societies are setting the themes of discussion.

Step by step libraries, data centres and other infrastructure units are intensifying their activities in the field of research data management over the last few years. Initiatives

[10] *OECD Principles and Guidelines for Access to Research Data from Public Funding*. Paris: OECD Publications; 2007.

[11] http://www.allianz-initiative.de/en/core_activities/research_data/principles/

[12] http://www.nature.com/news/specials/datasharing/

[13] http://www.sciencemag.org/site/special/data/

such as DataCite,[14] an international consortium for data citation, or the rise of research data repositories like PANGAEA[15] or Dryad[16] are examples for this trend.

Publishers are beginning to develop strategies to support the sharing of research data. In light of the "Brussels Declaration" from 2007 the STM publishers "encourage the public posting of the raw data outputs of research. Sets or sub-sets of data that are submitted with a paper to a journal should wherever possible be made freely accessible to other scholars." Cooperation between publishers like Elsevier and PANGAEA, the partnership between Dryad and a number of journals in the field of biodiversity research and the new breed of data publishing journals, such as ESSD[17] and GigaScience[18] – which build on the existence of reliable data repositories - are an indicator of the increased awareness of data sharing.

Further, stakeholders from the public and commercial sector involve themselves in the discussion. In the context of Open Access, taxpayer associations and stakeholders of the Open Data community demand broader access to publically funded data. An example is provided by the vision of the Open Knowledge Foundation (OKF): "for research to function effectively, and for society to reap the full benefits from research activities, research outputs should be open."[19] As well actors from the commercial sector emphasize the value of open research data. In the report "Big data - the next frontier for innovation competition and productivity" the research department of McKinsey & Company, a global management consulting firm, notes: "Access to data will need to broaden to capture the full potential for value creation. Increasingly, companies will need to acquire access to third-party data sources and integrate external information with their own, to capture the full potential of big data. In many cases, efficient markets are yet to be set up for trading or sharing data" [20]

These developments demonstrate the broad discussion on sharing research data. Nevertheless it must be noted that data sharing is still not the standard in science. Several studies focus on sharing practices in science. Some examples:

- Campbell EG, Clarridge BR, Gokhale M, et al. Data Withholding in Academic Genetics. Evidence From a National Survey. *JAMA*. 2002;287(4):473-480. Available at doi: 10.1001/jama.287.4.473
- PARSE.Insight. *Insight into digital preservation of research output in Europe. Insight Report.*; 2010. Available at: http://www.parse-insight.eu/downloads/PARSE-Insight_D3-6_InsightReport.pdf.
- Savage CJ, Vickers AJ. Empirical study of data sharing by authors publishing in PLoS journals. *PloS one*. 2009;4(9):e7078. Available at doi: 10.1371/journal.pone.0007078
- Tenopir C, Allard S, Douglass K, et al. Data Sharing by Scientists: Practices and Perceptions Neylon C, ed. *PLoS ONE*. 2011;6(6):e21101. Available at doi: 10.1371/journal.pone.0021101
- Vogeli C, Yucel R, Bendavid E, et al. Data withholding and the next generation of scientists: results of a national survey. *Academic Medicine*. 2006;81(2):128-36. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16436573

[14] http://www.datacite.org

[15] http://www.pangaea.de

[16] http://datadryad.org

[17] http://www.earth-system-science-data.net/

[18] http://www.slideshare.net/GigaScience/gigascience-a-new-resource-for-the-bigdata-community

[19] http://okfn.org/about/vision/

[20] http://www.mckinsey.com/mgi/publications/big_data

All these studies show that data sharing holds many challenges. Despite the high level and general enthusiasm for data sharing, re-use and preservation, successful implementation will require detailed understanding of a complex landscape of intertwined issues, which are related to data sharing.

2.2 Survey

To ensure a broad and common baseline picture of opportunities and challenges of data sharing covering all themes and aspects identified, the ODE project collected meaningful interview stories as "success stories", "near misses" and "honourable failures" in data sharing, re-use and preservation to form the baseline to help us analyse the drivers and barriers to data sharing.

After a thorough discussion and selection procedure involving all partners, 21 successful interview stories were finally completed, in which relevant stakeholders describe their experiences and their views on drivers and barriers to data sharing and re-use. The aim was to collect and identify evidence to form a reliable information baseline about the status quo of data sharing and re-use, including:

- attitudes of pioneering scientific communities,
- policies of funding agencies and relevant initiatives in the Member States,
- co-ordination of emerging pan-European e-Infrastructure initiatives,
- access of data e-Infrastructures to researchers in emerging economy,
- extension of data e-Infrastructures to the educational system.

Instead of re-evaluating and warming up results from previous or running studies the ODE partners decided to get a fresh first hand impression on practical experiences from stakeholders that are, or have been, directly involved in the activities listed above. To meet these requirements stakeholders were consulted via personal interviews that could represent different perspectives and characteristics in a profound qualitative manner.

During the first face-to-face meeting the ODE partners decided to focus on the following four perspectives to get a current and broad picture of the challenges and opportunities of handling research data:

- Scientific communities: the perspective of the scientists and their disciplines
- Infrastructure initiatives: the perspective of stakeholders from e-infrastructure departments and initiatives (e. g. data centres and data repositories) on national as well as EU and global level
- Management and policy initiatives: the perspective of stakeholders from the management and policy area (e. g. funding agencies and policy makers)
- Others: additional relevant stakeholders (e. g. citizen science projects)

Potential interview partners were discussed and suggested jointly by all partners. From a list of 45 suggested interviewees representing these four perspectives around 30 potential interviewees were chosen and contacted personally by the assigned project partners. From this group, 21 persons finally agreed to give an interview.

Each interview took one hour on average. The backgrounds of these interviewees cover a wide range from scientific communities of different disciplines, scientific infrastructures and management perspectives concerning funding and policy making.

The interviews were either conducted via face-to-face meetings or via conference calls or via videoconferencing. To support the interview process a recommended guideline was kept by starting with an adequate introduction of the interviewees' position, tasks and background, followed by the nature of research data and the perceived state of dealing with those data in the person's sphere. All interviews focused on the following aspects of data sharing:

- Highlights in data sharing
- Lowlights in data sharing,
- Unforeseen events in data sharing
- Intentions for the future sharing of data

Further financial, technical, legal, natural and social factors, which influence the aspects mentioned, were queried.

Finally every interview has been reviewed and approved by the interviewee. For each interview, a comprehensible and narrative interview story was written. These individual stories of "success", "near misses" and "honourable failures" in data sharing form the baseline to analyse the drivers and barriers to data sharing.

The following persons, grouped accordingly to the four perspectives, were successfully interviewed. Since most of the persons held several roles in their career and are active in several areas a clear distinction is not always possible.

Scientific community:

Person and topic	Institution and position	Research field
Prof. Dr. Peter Braun-Munzinger: The cultural challenges of data sharing in high energy physics.	Scientific Director of the ExtreMe Matter Institute at the GSI Helmholtzzentrum für Schwerionenforschung and Professor of Physics at the Technical University in Darmstadt, Germany.	Physics (High Energy Physics)
Graham Cameron: Handling the increasing size and complexity of data in molecular biology.	Associate Director of the European Bioinformatics Institute (EBI) of the European Molecular Biology Laboratory (EMBL). The EBI is based at Hinxton, UK.	Life Science (Biology)
Dr. David Carlson: A lesson in sharing.	Director of the International Polar Year 2007-2008 International Program Office (IPO) at the British Antarctic Survey in Cambridge, UK.	Geosciences (Polar Research)
Prof. Dr. Peter Lemke : Lessons learnt from data sharing in meteorology and Intergovernmental Panel on Climate Change (IPCC).	Head of the Climate Sciences Division at the Alfred Wegener Institute for Polar and Marine Research (AWI) and Professor of Physics of Atmosphere and Ocean at the Institute of Environmental Physics at the University of Bremen,	Geosciences (Climate Research)

	Germany.	
Prof. Dr. Karen Helen Wiltshire: „Data are our gold“	Biologist and Head of Biologische Anstalt Helgoland; Wadden Sea Station Sylt and deputy director of Alfred Wegener Institute for Polar and Marine Research (AWI) in Bremerhaven, Germany.	Geosciences (Biodiversity)

Infrastructure initiatives:

Person and topic	Institution and position	Research field
Dr. Libby Bishop & Veerle van der Eynden: Data sharing constraints in Social Sciences and Humanities.	Libby Bishop is Senior Officer Research Data Management Support Services at UK Data Archive and Veerle van den Eyden is Research Data Management Support Services Manager at UK Data Archive.	Social Sciences and Humanities
Dr. Michael Diepenbroek: PANGAEA, a data publishing system for Earth & Environmental Science	Managing director of PANGAEA and responsible for the operation of the World Data Center for Marine Environmental Sciences (WDC-MARE) at University Bremen and Alfred Wegener Institute of Polar and Marine Research (AWI) in Germany.	Geosciences
Neil Holdsworth: Data management in the context of the International Council for the Exploration of the Sea (ICES).	Head of the Data Centre at the International Council for the Exploration of the Sea (ICES) in Copenhagen, Denmark.	Geosciences (Marine Sciences)
Prof. Dr. Peter Igo-Kemenes: Costly efforts due to lacking data preservation	Professor of Physics a the Gjøvik University College in Norway and Senior Scientific Advisor of CERN.	Physics (High Energy Physics)
Dr. Leif Laaksonen: Recommendations of the e-Infrastructure Reflection Group (e-IRG)	Director at CSC - the Finnish IT Center for Science. Chair of e-IRG board during 2006-2010.	General
Eberhard Mikusch, & Katrin Molch: Work of a remote sensing data center	Eberhard Mikusch heads the department of information technology at the German Remote Sensing Data Center (DFD) at the German Aerospace Center (DLR). Katrin Molch is responsible for the DFD data services.	Geosciences (Remote Sensing)
Dr. Tommi Nyrönen & Dr. Andrew Lyall: ELIXIR - a sustainable data storage infrastructure for biological information in Europe.	Tommi Nyrönen is project coordinator of ELIXIR collaborator in Finland at CSC - the Finnish IT Center for Science. Andrew Lyall works as project manager of ELIXIR at the European Bioinformatics Institute (EBI) in Cambridge, UK.	Life Science (Biology)

Dr. Heather Piwowar: Data repositories for research communities.	Postdoc research associate, funded by the NSF-funded DataONE cyberinfrastructure project at the National Evolutionary Synthesis Center, Nescent in Durham, USA.	Life Science
Dr. Andrew Treloar: The potential of data publishing to avoid suspicion of fraud.	Linguist and Technical Director of the Australian National Data Service (ANDS).	General

Management and policy initiatives:

Person and topic	Institution and position	Research field
Dr. Andrew Treloar: The potential of data publishing to avoid suspicion of fraud.	Linguist and Technical Director of the Australian National Data Service (ANDS).	General
John Doove & Wilma Mossink: Hesitation in data sharing despite existing infrastructures.	John Doove is project coordinator at the SURFfoundation in the Netherlands with responsibilities in Enhanced Publications and Collaboratories Wilma Mossink is Project Manager with responsibilities in Permanent Access to Data.	General
Dr. Toby Green: Usable standards and services for the reuse of research data.	Head of Publishing at OECD in Paris, France.	Social Sciences
Dr. Simon Hodson: Data management plans are necessary.	Program Manager at JISC in London, UK, responsible for digital infrastructure and managing research data.	General
Finnish task force for utilization of electronic data in research	National cross-sectoral task force set by the Finnish Ministry of Education and Culture	General
Dr. Stefan Winkler-Nees: A funders view on data sharing.	Program officer at the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation).	General

Other relevant initiatives and projects

Person and topic	Institution and position	Research field
Dr. Carolin Liefke: The challenge of discoverability in data deluge	Astronomer at the Haus der Astronomie, Heidelberg's center for astronomy education and outreach.	Astronomy , (Citizen Science)
Prof. Dr. Karin Lochte: Perspective from the EC-expert group on Research Infrastructures	Director of the Alfred Wegener Institute for Polar and Marine Research (AWI), Germany.	General

2.3 Stories of Success, Near Misses and Honourable Failures in Data Sharing

The following evidence of "success stories", "near misses" and "honourable failures" present a comprehensive inventory of stakeholders' views on data sharing. These stories will be explored subsequently in the analytical phases of the ODE-project concerning drivers and barriers to data sharing through a European eco-system of data repositories.

The interviews were carried out in the first half of 2011 by:

- Suenje Dallmeier-Tiessen (CERN)
- Heinz Pampel (Helmholtz Association)
- Hans Pfeiffenberger (Helmholtz Association)
- Angela Schäfer (Helmholtz Association)
- Satu Tissari (CSC)

2.3.1 Libby Bishop & Veerle Van der Eynden (UK Data Archive)

Libby Bishop is Senior Researcher Liaison at UK Data Archive and Veerle van den Eynden is Research Data Management Support Services Manager at UK Data Archive. The UK Data Archive curates the largest collection of digital social and economic research data in the UK. It not only acquires, curates, and provides access to these datasets, but also provides the support and technical infrastructure for the community to "fulfil" the requirements set by the policies from funding bodies and research councils. Currently it hosts several thousand datasets in total. The Archive is largely funded by the ESRC, the JISC and the University of Essex.

What kind of research data is being handled at the UK Data Archive?

The UK Data Archive deals with research data from academic research, governmental data, and commercial data. The interviewees deal with the first type of research data, produced by individual researchers and research groups in the domain of the wider social sciences and humanities (SSH).

The needs of research data management in the SSH can be very particular as it is data related to people. When it comes to qualitative data for example, some interview data may need particular handling. In this instance, one cannot simply take a dataset and ingest it into a data repository. Further pre-processing is needed in order to make the research dataset suitable for publication, e.g. anonymising personal information or ensuring consent for data sharing or data publishing is in place. In addition, data management for this kind of research data requires a lot of engagement with researchers to ensure that attention is paid to data preparation, licensing, consent, and access rights during research.

What do they do in regard to research data sharing?

In the daily routine the work comprises of a lot of "hands-on" engagement: researchers who wish to publish their data in this domain usually need consultancy by human beings. There are lots of different subtypes of research data where different factors before publication need to be considered (for example to preserve anonymity). Thus, in this domain a lot of human intervention is needed and automated processing and ingest

of data is rather limited. The consultancy work is as diverse as the SSH data, and thus it is important to have specialists in place to deal with all the specific cases.

The support of the community and the individual researchers is crucial in this domain. There is widely varied experience with research data sharing. For many researchers it is their first time sharing their data. They don't know how to do it, they know there are some important things to consider before sharing it, but don't know the details. Here, consultancy is needed. It is important to note that for certain datasets open data sharing is not possible and specific access controls need to apply, e.g. to enable the sharing of confidential or sensitive data.

With more and more policies from funding bodies and research councils emerging it is even more important to guide the researchers though the "do's" and "don'ts in data sharing, so that they comply with the guidelines and share data in an appropriate manner.

Highlights

According to the interviewees, one highlight is certainly the emerging awareness of data sharing throughout the community. Previously, the UK Data Archive organized conference sessions to promote this topic in the community. Now, there are more secondary analysis projects, meaning that there is increased data reuse. Moreover, this has become a topic that is raised by the community itself, in the sense that they organize re-use events independent of the UK Data Archive.

Challenge

Based on extensive experience in, and communication with, the research community both interviewees report that there is still a lot of hesitation in the research community when it comes to sharing their material. They are busy with research and publishing, and sharing research data is often not yet on the agenda; in particular because their preservation and sharing are not considered for promotion and research assessment.

Now, there is the "carrot and stick" question: researchers must preserve and share their data because they are obliged to do so by funding bodies etc, but they are not yet really seeing the benefit. This is a long-term development and is changing, but slowly. Such changes will need more time and more consultancies.

More projects and more challenges ahead…

For the UK Data Archive, one upcoming project is persistent identification via DOI (Digital Object Identifier), which will make datasets citable. This is in discussion and will commence in the near future.

A challenge ahead is certainly the financial situation which will impose financial cuts on academia in the UK. This is an unfortunate situation, as data need proper treatment and preparation. The researchers need consultancy, which becomes even more important with more and more policies by funding bodies. If one wants to encourage researchers to share their data, one also needs to support this with the corresponding infrastructures and services.

2.3.2 Peter Braun-Munzinger (GSI Helmholtz Centre for Heavy Ion Research)

Prof. Dr. Peter Braun-Munzinger is Scientific Director of the ExtreMe Matter Institute at the GSI Helmholtzzentrum für Schwerionenforschung[21] and Professor of Physics at the Technical University in Darmstadt, Germany. Braun-Munzinger has been spokesperson for several different nuclear physics experiments worldwide. Since 2011 he has chaired the Collaboration Board of ALICE,[22] at the Large Hadron Collider (LHC) at CERN.

ALICE is one of the four major detector experiments at the LHC at CERN. The ALICE Collaboration has built a dedicated heavy-ion detector to exploit the unique physics potential of nucleus-nucleus interactions at LHC energies. This project involves an international collaboration of more than 1000 physicists, engineers and technicians, including around 250 graduate students, from 105 physics institutes in 30 countries across the world. The ALICE experiment produces 160 GB of information per second. In an automatic selection process, the incoming data is filtered at a rate of 1.25 GB per second. To organize this enormous amount of data, an elaborate data infrastructure is necessary. The management of the data is organized by multi-tiered computer architecture, called the Worldwide LHC Computing Grid (WLCG).[23] WLCG is a global collaboration of more than 140 computing centres in 35 countries. The mission of the WLCG is to maintain data storage and analysis infrastructure for the entire high energy physics community in the context of LHC. Data from the LHC experiments are distributed world-wide, with a primary backup recorded on tape at CERN. After initial processing, this data is distributed to eleven large computer centres with sufficient storage capacity for a large fraction of data and with round-the-clock support for the computing grid. These so-called "Tier-1" centres make data available to over 160 "Tier-2"centres for specific analytic tasks. Individual scientists can then access the LHC data from their home country, using local computer clusters or even individual PCs.

While the high-energy physics community is a pioneer in the field of open access to scientific articles, the sharing of data still brings many challenges. "We have an excellent data infrastructure, but no culture of raw data sharing," says Braun-Munzinger and continues: "There is a huge competition between the different collaborations and also in the experiments. This competition affects all options of data sharing."

The processing of the ALICE data is very complex. During the data processing, many parameters are taken into account. Only after complex processing can the data be analysed. "There is a lot of work behind the data," says Braun-Munzinger and points out the long way from analysis of processed data to published articles. Rigorous and time-consuming internal and external peer review processes of the data and the article are required before publication. "If data will be published before their description appears in an article, then we need to find proper ways of recognizing the work behind the data."

Currently the ALICE collaboration makes only processed data available after their description in a scientific article. "For this purpose we use the 'Reaction Database' of the 'Durham HEPData Project'." The HEPData Reaction Database is a repository for data from particle and nuclear physics experiments hosted at Durham University.[24] In

[21] http://www.gsi.de

[22] http://aliweb.cern.ch

[23] http://lcg.web.cern.ch/lcg/public/default.htm

[24] http://durpdg.dur.ac.uk

contrast to the GRID infrastructure, this database is maintained by only a few people. The data can be accessed in different formats via a web interface. In addition, the data is published on the ALICE website. The HEPData Reaction Database links to the arXiv-ID.[25] This linkage connects the article, which describes the according scientific results, and the processed data.

Braun-Munzinger points to the ongoing discussion on data sharing in the high-energy physics community: "The community discusses this topic in various contexts. I think we have to face the cultural challenges of data sharing. And this could touch a lot of serious questions of our work in high-energy physics. For example in an open science world, we have to accredit the intellectual work of the many persons who do the work behind to make data originally fit for scientific usage. This is usually done via scientific notes which may or may not stay internal in the collaboration. And we also need to avoid misinterpretation of data. Last but not least, we have to ensure the processes of quality assurance. So, there is some way to go."

[25] http://arxiv.org

2.3.3 Graham Cameron (European Bioinformatics Institute - EBI)

Graham Cameron is the Associate Director of the European Bioinformatics Institute (EBI),[26] which is part of the European Molecular Biology Laboratory (EMBL).[27] EMBL-EBI is based on the Wellcome Trust Genome Campus at Hinxton, near Cambridge, in the UK. Graham is responsible for several EU projects and oversees the institute's vast range of services, in particular the data libraries. He describes himself as a "data sharer" rather than a "classical" researcher.

Graham started working for EMBL in Heidelberg in 1982. There, he established and managed the EMBL Data Library, which grew to become EMBL-EBI. He played a major role in conceiving and developing EMBL-EBI, and became its second staff member. Today, EMBL-EBI has more than 500 members of staff.

What is your and EMBL-EBI's experience with research data?

Managing research data has always been a challenge, and one that EMBL has tackled from its very beginnings. In the 1970s they started to collect data from research projects, and in 1981 EMBL established one of the first data libraries in the world for nucleotide sequence data. At first, the goal was simply to extract data from journals. But with the acceleration of DNA extraction and growing efficiency of high-throughput methodologies, the focus shifted to attracting direct data submission by the researchers themselves. Journals were initially rather reluctant to expand their involvement in data extraction and sharing, but over time this has changed.

Similar developments were happening at the same time around the world, in particular in the US with GenBank. In 1986 the International Nucleotide Sequence Database Collaboration (INSDC) was signed, which was the beginning of the successful cooperation of the DDBJ in Japan, GenBank in the US and EMBL-EBI's Nucleotide Sequence Database. These three databases exchange data and synchronize daily, thus making it easier for researchers to access up-to-date data from around the world. The agreement will hopefully expand in the next year to include Chinese partners.

How is research data shared in the domain of molecular biology?

Because research data are published in the public domain, they could potentially be aggregated and sold by commercial users. The decision to place the data in the public domain is driven by the communities' demand for easy access to – and reuse of – the information they need for advancement. Sometimes, data is first submitted and accepted to the database with a delay in the actual publication date. Such a delay is usually driven by the submission and acceptance of a publication in a journal that requires an accession number to the data at the time of submission. But there are cases when the data producers do not want to see their data openly available before the publication of their paper.

In the very early days, the databases only published datasets that were discussed in peer-reviewed publications, in the belief that these data were quality controlled. This has changed, as the data are not integral to the classical peer-review process. Within the

[26] http://www.ebi.ac.uk

[27] http://www.embl.de

databases at EMBL-EBI, there is quality control upon data submission. It is mainly an automated process but also needs some "hands-on" curation by human beings. This could mean that the data producers are contacted by email or even by telephone when the submission team has questions regarding the dataset. This feedback is highly appreciated by the researchers.

What are the challenges associated with data sharing in the field of molecular biology?

The development of methodologies and data production in molecular biology has been accelerating rapidly. For example: the work of the Human Genome Project took 10 years to complete – that same work could now be redone within 10 minutes. Linked to this growth is also the variety and sizes of databases, which hold data ranging from little experiments to whole genomes. Over time data has come to be considered as an established scientific record. Data access is undoubtedly beneficial for the community. For instance, biomedical data access could accelerate scientific advancements for the wellbeing of humans, while data access to molecular forestry data could feed back directly to the environment.

In recent years, data production has been accelerating faster than ever. Thus, the extension of data storage has become a new challenge and there are some initiatives working on, for example, data compression.

With the increasing size and complexity of the data being produced, one of the major bottlenecks today is the contextualization and integration of data – a big challenge for bioinformatics. A user who is interested in a particular topic might not only be interested in one specific analysis, but also in other research results related to this topic. How can these materials be integrated and displayed?

A new development in molecular biology research is projects that concentrate solely on data production. The analysis and interpretation of these data is separate from the project that produces the data. Usually the data produced in a project is submitted to the database immediately. This facilitates early usage, but also asks for new discoverability tools to facilitate easy reuse of the massive amount of new material available – again a challenge for bioinformatics.

Another challenge is commercial data production. Even though an estimated 15-20% of the database users work in commercial enterprises, they hesitate to share their data openly. Based on the EBI's activities, data sharing within different commercial sectors has been stimulated. However, issues like patenting are still considered constraints.

Why is the molecular biology community (in comparison to many others) so successful in sharing their research data?

This certainly relates to the question why science, and this discipline in particular, is so successful. One answer could be that genes are everywhere. It is obvious to the involved communities that the entirety of the record is needed publicly. Unless everyone shares their data, they're of no use to anyone.

Moreover, it is easy to work with the data. The science is international, and so are the databases. In the past, paper publications were the main place to find scientific results.

But when journals started to require the accession numbers for submission, the relevance of the databases and research data increased.

In addition, the reuse of data is potentially very powerful - just browsing through datasets could lead to new hypotheses that could be tested.

In summary – what are the highlights or challenges that are experienced in the sharing of bimolecular research data?

Certainly one highlight is the early international agreement among the three international projects in the US, Japan and Germany that facilitated molecular biology data exchange from the very beginning. Being interoperable and following the standards one had agreed on, the three databases together became a powerful tool to search their domain.

The biggest challenges the community is facing are the data deluge and access to chemical information. Chemical information is an integral part of bimolecular research and even though biological information is shared rather openly, chemical information is not. They are often proprietary data and access is limited and costly. As for the data deluge: the information overload for researchers is a challenge. Now there is a need to integrate the different research materials from the different databases and serve it to the users - but how? It is important to respond to the needs of researchers and build usable interfaces that facilitate easy reuse of the materials.

EMBL-EBI in numbers:

- Visitors to the EMBL-EBI website in 2010: 3.4 million unique IP (which could represent either an individuals or whole organization)
- Data stored by EMBL-EBI as of July 2011: 10 petabytes
- Data submissions per second: 2
- Growth rate of datasets last year: Doubles every 18 months
- Growth rate of datasets this year: Doubles every 10 months
- Per cent of EMBL-EBI users at companies: 15-20 (conservative estimate)

2.3.4 David Carlson (International Polar Year)

David Carlson served as director of the IPY International Program Office. He is now education and outreach director for the non-profit geodesy consortium UNAVCO in Boulder. Dave Carlson gave no full interview per se, but he suggested treating his Nature article as an "interview story" for ODE:

Carlson, D. (2011). "A lesson in sharing" Nature 469(7330): 293-293. http://dx.doi.org/10.1038/469293a

2.3.5 Michael Diepenbroek (World Data Center for Marine Environmental Sciences - WDC-MARE)

Dr. Michael Diepenbroek has been managing director of the scientific information systems PANGAEA[28] since 1998 and is responsible for the operation of the World Data Center-MARE,[29] based at the Center for Marine Environmental Sciences (MARUM) at University Bremen and the Alfred Wegener Institute of Polar and Marine Research (AWI) in Germany. From 1992 to 1997 he elaborated the conception and implementation of PANGAEA at the AWI. Michael Diepenbroek was strongly engaged in transforming the World Data Centre system (WDC) into the new ICSU World Data System (WDS) ratified by International Council for Science in 2008.[30]

What is PANGAEA?

PANGAEA is a 'data publishing system' for Earth & Environmental Science and, as such, partner in numerous projects (European and international) covering all fields of geo- and biosciences. Since 1996 data management services are supplied on an international level. During the last years PANGAEA also became engaged in projects supporting spatial data infrastructures (SDI), as well as becoming a lead partner for the implementation of data portals and infrastructures in several NoE's initiatives. In this context PANGAEA assembled substantial knowledge and practical experience in the implementation of international standards and web technologies.

What is the success of PANGAEA and which draw backs did you experience in scientific data management?

The overall aim of PANGAEA nowadays is making scientific data available for re-use. In that process we had and we still have to cope continuously with two separate main challenges: technical installation and software management. (Of course besides running after the data personally, since data storing and sharing is not a naturally understood commitment for all scientists.)

In the very beginning of our unstructured data management attempts at the AWI in the 1990s we concentrated on individual scientific splinter groups. Hence we tried to deliver individual solutions for them. At this we could neither fulfil specially defined requirements, nor generally accepted requirements, in one go. Also we could not guarantee sustainability for only small individual groups since that kind of long-lasting framework was far too sizeable and costly to construct.

From these individual small scientific groups (e.g. Prof. M. Sarnthein's working group at Kiel University) data analysis as well as data management was demanded. Hence scientific interpretation data, analytical result data and derivates were mixed ineffectively with raw data management. Learning from this predicament we skipped analytical tasks and concentrated purely on the curatorial functions in data management.

[28] http://www.pangaea.de

[29] http://www.wdc-mare.org

[30] http://www.icsu-wds.org

What data are worth for storing and how to make data qualitatively fit for storing!

While we saw no efficiency in storing uncorrected and unproved raw data (e.g. automated underway data from e.g. the DSHIP system of RV "Polarstern"), we needed to define 'the' principle unit of a 'data set' worth archiving. It became evident very early that a 'data set' has to be a publishable and citable entity described by substantial metadata to ensure data-reusability. Together with our customers (data provider and data user) we assigned a guideline: The original data set that we ingest into the repository, should be retrievable as exactly the same fixed and defined unit - open accessibly and fit for re-use!

Since reliable data quality became more and more an issue we tried to ensure it with a defined quality flagging system to depict outliers, ranges and additional tests of variances that belong to our plausibility check during data ingest into the information system.

How to guarantee qualified repository services and true scientific reusability of data?

In the course of storing scientific data from all kinds of multidisciplinary scientific programs and publications PANGAEA became an agent for homogenization of analytical measurements assigned to define and (by the scientific community) accepted parameters. These parameter definitions are crucial for data management and data storage. It needs assigned data repositories with trained scientific data curators to assure true scientific parameter homogenization. Furthermore in terms of data quality not the originally submitted data are ingested, but an assembled data set back and forth communicated by PANGAEA data curators to be finally reassured and validated by the responsible author (principle investigator). Very often a time-consuming and tedious task!

Consequently the data set editors (scientific data curators) are working in-house at PANGAEA - a data publishing system – since the semantic background and their expertise has to be assured throughout the whole procedure. To encompass the whole life cycle of data from gathering to storing to reuse, we always operate best internally, within the scientific project itself - first to assure quality and second to assure financing via the same project. In the same course, we keep the scientific status quo and we are well embedded in actual science. Normally we participate simultaneously in about 12 international and national major projects, besides the daily contact with our affiliated institutes' scientists or independent requests.

Effort and financial aspects of data storing and sharing

The idea that a 'data set' has to be a publishable and citable entity described by substantial metadata was already appreciated by Springer in 1994, but condemned for not providing financially profit! Of course a data archive with such a public assignment to the scientific community cannot work from a pure economic perception. Therefore we, as a data archive, started to cooperate with international publishers during the last fifteen years.

Still our financial pillar is the direct participation in scientific projects with their additional part of funding that recognizes the need of data archiving. But project based data curation and storage alone does not cover the full cost. It needs additional financial acknowledgment for developing and maintaining future integrative data related e-

infrastructures for coping with the exponentially increasing flood and complexity of data nowadays. These data are produced by data intensive sciences that of course trigger and exploit the development of improved sampling and high resolution sensor technologies. And all of this in the context of international cooperative networks (e.g. real-time monitoring programs), and, of course, everyone wants the data to be integrated, visible, accessible and reusable.

Since 1995, when the original data model behind PANGAEA was developed, it is still the same in principle, but the whole middleware (the part that breaks down and reassembles the matrices), and the back and front end services had to be created from scratch and adapted continuously. These are huge IT-development tasks that are not yet fully perceived by either the scientific community or the funding machinery.

How can you measure the success of PANGAEA?

PANGAEA is very well known globally in the Earth and marine environmental sciences. Our web statistics show tens of thousands of unique users per year, and, on average, nearly 500 data are downloaded per day. For the geoscientific and, in particular, the oceanographic community PANGAEA is very important and unique by means of its specific developed method to handle multifarious interdisciplinary data. Furthermore we deliver synoptic data views of projects for financiers and reviewers especially for EU-funded projects.

What is the central driver of PANGAEA?

Since our overall aim is focused on the meta-analysis of data (re-use!) we usually participate first hand in projects to cooperate directly with the scientists to ensure quality and topic scientific standard. In addition we provide accredited citability and long-term preservation associated with persistent and globally resolved digital object identifiers. Subsequently we build up reputation and trust - the back bones of good scientific practice.

2.3.6 John Doove & Wilma Mossink (SURFoundation)

John Doove is SURF Project Coordinator with responsibilities in Enhanced Publications and Collaboratories (VRE) as well as a member of the Knowledge Exchange Working group. Wilma Mossink is SURF Project Manager, with responsibilities in Permanent Access to Data and also chair of the Dutch Research Data Forum.[31]

What kind of research data is being handled and what do they do in regard to research data sharing?

SURF acts as a funding body which established the program SURFshare in which different projects focused on research data are supported. Within this framework (and also the national coalitions SURF is participating in) all kinds of research data from different disciplines are considered and supported. The two interviewees support two different aspects in regard to research data, Wilma is in charge of a work package that concentrates on the organizational aspects of permanent access to research data, whereas John takes care of the program "enhanced publications", focusing on the linkage between publications and research data (and other relevant research output).

One of the core activities in the SURFshare program which closely relates to the topic of access to research data is "Enhanced Publications". Development on Enhanced Publications started during the DIRVERII project[32] followed by calls for tender in 2008, 2009 and 2011. The projects[33] were carried out in different disciplines, ranging from the humanities to the "hard sciences".

The technical infrastructure is similar across the different disciplines, facilitating easy exchange of information across the systems. It became very clear from the beginning of this model that there are different habits and needs within the different disciplines, for example in archaeology and musicological science. Thus, in order to serve these needs customized tools for creation and front ends for visualisation are in place which supports the individual workflows.

Currently the repository infrastructure is being upgraded to support the creation, storage, visualization and exchange of Enhanced Publications. This has resulted in a common data model[34] that is used by the different tools for creations developed in the different projects' Enhanced Publications (for example: ESCAPE[35]). Additionally all created Enhanced Publications will be aggregated in the Open Access portal for scientific output in the Netherlands; Narcis.[36]

Another focus of the SURFshare program is permanent access to research data. SURF started with the program for Enhanced Publications, but realized that there is no "Enhanced Publication" without proper data preservation and data access models and that more effort is needed in these domains as well. Thus these fields became an individual work package within SURFshare and a close collaboration exists between the

[31] http://www.surffoundation.nl
[32] http:/www.surffoundation.nl/enhancedpublications
[33] http://www.driver-repository.eu/Enhanced-Publications.html
[34] http://wiki.surffoundation.nl/display/vp/1.1+Information+Model+for+Enhanced+Publications+whitepaper
[35]http://www.surffoundation.nl/en/projecten/Pages/ESCAPE-Enhanced-Scientific-Communication-by-Aggregated-Publications-Environments.aspx
[36] http://www.narcis.nl

two. SURF discusses the concept of data preservation and data access after the silo model by A. Treloar.[37] In addition, data licensing and related aspects play an important role when discussing data access. Both interviewees underline the need to understand the researchers' habits and needs in order to launch services that are really valuable for their workflows. Thus, they commissioned, amongst other reports, a report on "what researchers want"[38] [7] in regard to research data and have focused their approach on close cooperation with researchers (see for example the CARDS project[39]).

Lessons learnt from their activities in the field of data sharing

- A continuous development of infrastructures and services is needed, it is required to specify the disciplines' needs as there are different publication cultures and different handling of materials within communities
- The Enhanced Publication is an example that proofs that there could possibly be only one (technical) data publication model in the backend that serves (with adapted frontends) different disciplines
- The researchers' hesitation is one big challenge that needs to be tackled by many projects, e.g. by developing and offering new tools and services.

Highlights and challenges in the framework of their data sharing experience:

Within the experience of the work package Enhanced Publications one highlight is the publication of qualitative data integrated with a digital publication, e.g. in the "Veteran tapes EP project" which is being reused across disciplines. It is considered an exceptionally successful approach in which interview data have been made available to the public. The data are considered very valuable historical documentation and have been preserved in a labour-intensive way in order to make them reusable for future generations.

On the other hand, both interviewees consider the advancement of data sharing as a big challenge. Researchers appear to be scared to share their data, they hesitate to publish it. This is a challenge for the national and international initiatives. There are some questions that need to be solved:

- How do you convince researchers to publish research data?
- What are the conditions? One proposition could be "open where possible, closed when needed"
- What are the licenses?

To solve these questions it is necessary to exchange expertise in research data management on both a national and on an international level. That's why the Dutch research data forum has been initiated, which is a national coalition that currently

[37] Treloar, A.: Data management and the curation continuum: how the Monash experience is informing repository relationships. http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf

[38] http://www.surffoundation.nl/nl/publicaties/Documents/What_researchers_want.pdf

[39] http://www.surffoundation.nl/en/projecten/Pages/CARDS.aspx

consists of 35 members. SURF is also collaborating in many international initiatives, such as Knowledge Exchange which has a dedicated group for research data [10].[40]

John concludes that the development of data publication is under way. Data publication is not yet considered an independent contribution in scholarly communication. They do not yet count towards promotion or research assessments. The hesitation is apparent across disciplines: enhanced publications could be considered as a way to raise awareness of the fact that there is more to share than just the article.

[40] http://www.knowledge-exchange.info

2.3.7 Toby Green (Organisation for Economic Co-operation and Development - OECD)

Toby Green is currently the Head of Publishing at OECD in Paris. He has more than 25 years experience in scholarly and STM publishing. He has held several positions, starting with Academic Press, then Applied Science Publishers, then Pergamon Press and Elsevier Science. Toby Green joined OECD as Head of Marketing in 1998 and was promoted to Head of Publishing in 2007.

In 2001, OECD launched the world's first combined e-books, e-journals and dataset service, SourceOECD. This platform was re-launched as OECD iLibrary in 2009 and now also includes working papers.

Toby Green is currently Chair of ALPSP, the largest international association of non-profit scholarly publishers. He is the author of the white paper "We Need Publishing Standards for Datasets and Data Tables"[41].

What does the OECD and research data bring together?

The mission of the Organisation for Economic Co-operation and Development (OECD) is to promote policies that will improve the economic and social well-being of people around the world. The OECD provides a forum in which governments can work together to share experiences and seek solutions to common problems. The fruits of the OECD's research, analysis and data gathering are published as a series of reports and datasets. This output is highly relevant for policy makers, researchers in civil society, academia and some commercial sectors.

Every year OECD publishes approximately 250 reports and 100 working papers alongside 700 datasets. All reports published since 1998 are available online and those since 2005 are available in print via print-on-demand channels. Datasets are also available online with annual archival editions on CD-Rom.

All these publications, working papers and datasets are available online via OECD's publishing platform, OECD iLibrary[42].

What kind of research data is being handled?

Generally speaking two types of data:
Firstly, there is "live" research data that is being updated regularly. These so-called longitudinal, time-series, datasets
Secondly, there are one-off datasets gathered for particular research projects. These datasets do not change over time; it could be considered "frozen" data.

How does the OECD publish research data?

OECD considers datasets as published 'objects' in much the same way as a book or journal article is a published 'object'. Therefore, just as a book or journal article has a cite-able bibliographic and catalogue record, so each and every dataset has one too.

[41] http://dx.doi.org/10.1787/603233448430

[42] www.oecd-ilibrary.org

This is evident, when looking into the detailed presentation of the datasets: first of all they do have their own MARC records; secondly they are presented with an individual DOI (digital object identifier). In addition, subsets of data and data collections receive their own DOI. A downloadable, ready-made, citation is offered for each dataset that includes the DOI to encourage end-users to cite data in the same way they would cite a journal article or book.

"Live" datasets retain the same bibliographic record from year-to-year, but if the older data is revised significantly (a rare event), the current dataset is frozen and a new dataset is released with a new bibliographic record and DOI, with links between them (a parallel to the way journals are managed when they change title.)

Data is published in one of two ways. It is either published as a stand-alone dataset (which might include data sub-sets within a collection) or it can be published as supplementary data linked to a particular publication. The links to supplementary data are called 'StatLinks'.

In both cases, OECD's data editors work with the data producers and authors to help prepare the data for publication. This quality-assurance work ensures that the data is accessible and understandable for a wide range of end-users. A central concept of OECD iLibrary is to help users find content – whether data or analysis – as quickly and as simply as possible. All the content available (text, tables, figures) is displayed in search results, sorted by relevance not by content type. It could be that the supplementary data is found first, leading the user to the chapter, not necessarily the other way around.

What would you consider a personal highlight and lowlight in your experience with data sharing?

One highlight is the OECD Factbook. This compendium of 120 indicators drawn from across the breadth of OECD's data collection is presented in a variety of ways: print, USB key, online and as an App. Each indicator is a double-page spread containing data and an explanation in simple, accessible, terms. The underlying data for each indicator is available as a spreadsheet, even from the print edition..

Another success is the OECD Better Life Index, launched in May 2011. The Better Life Index allows end-users to adjust the weighting of eleven parameters so they can build (and share) their own index based on OECD's data.

Less successful has been the development of a generic visualization tool because too many features have been crammed together so the tool often 'gets in the way' of the data and storyline. A simpler version is now being developed.

What are the projects and challenges ahead?

The central challenge is to find a business/funding model for publishing that is sustainable in the long-run. The cost of publishing is increasing and the march of technology means continuous investment in publishing systems will be required for the foreseeable future.

Another challenge concerns long-term archiving of data – who will ensure that datasets available today will continue to be available on 50 or 100 years' time?

Regarding the vision, it is important to embed data publishing in scholarly communication even further, it needs to be a seamless experience for users.

Links: OECD iLibrary[43], Statlinks[44], OECD Better Life Index[45], OECD Fact book[46], Visualisation tool[47]

[43] http://www.oecd-ilibrary.org

[44] http://oe.cd/

[45] http://www.oecdbetterlifeindex.org

[46] www.oecd.org/publishing/factbook

[47] http://stats.oecd.org/OECDregionalstatistics

2.3.8 Simon Hodson (Joint Information Systems Committee - JISC)

In the UK a rather larger number of policies by research funders, even law, exist which exert a growing influence on researchers practises toward data. Simon Hodson is the programme manager for Digital Infrastructure, Managing Research Data at the JISC ("historically": the Joint Information Systems Committee) and oversees a large number of projects which deal with the multiple aspects and necessities of data sharing.

The interview started out with some questions about the FoI (Freedom of Information) law, as applied to publicly funded research data. Hodson noted that, while on the whole public opinion in the UK finds FoI "a good thing", researchers have some misgivings. Nevertheless, researchers and universities need to adapt because FoI is the law!

FoI law provides some protection, such as: the need to protect personal information trumps FoI. The Scottish FoI Act also provides a degree of protection against premature release of data, which could damage research, where there is an ongoing research project. In England and Wales, which has a separate act, this protection is only available if the research project has a pre-existing publication plan. On the other hand, the perceived, potential damage by misinterpretation of data is no valid objection. Besides the so called "Climategate", the FoIA formed the background to the JISC-funded project ACRID (Advanced Climate Research Infrastructure for Data) by the UEA (University of East Anglia) and the STFC eScience centre. The project is based on the UEA climate data, which are indeed available, but deserve of improved access to documentation, e.g., full provenance information and software codes.

Hodson observed that regarding retention times there is currently a wide spectrum of positions, depending on discipline: The BBSRC (Biotechnology and Biological Sciences Research Council) currently requires raw and original data to be retained for 10 years (although this is apparently under review), while in social sciences, widespread current practice is such that subjects of interviews may have been promised that interviews would be destroyed after 5 years.

Therefore, selection, appraisal and retention of data need to happen on a case-by-case basis. For this, guidelines have to be developed by funders and data archives in consultation with scientists and learned societies. Hodson expects that a general guideline will hold: "Unless there is a good reason to destroy data, it should be preserved and shared". Obviously, this includes the possibility that data management plans in proposals can contain the action "destroy".

As part of a JISC-funded project, the UK Data Archive examined data management practice in major programmes and centres funded by the Economic and Social Research Council (ESRC). Although there were some, generally individual, examples of good practice, the study found that there was considerable room for improvement. A particular issue lay with longterm investments, where, often as a result of repeated or extended funding, the requirement to deposit data at project end had been overridden and, as a result, data produced in the early life of a 15 year centre had not be deposited and risked being lost. In response, and working closely with the ESRC, the project produced data management guidelines, model data management plans for such large, ongoing investments and made a set of recommendations. These included the importance of Principal Investigators ensuring that a senior owner takes ultimate responsibility for ensuring good data management practice, pointing out the benefits of a

resources hub of useful information on data management and, above all, the need for allocation of sufficient resources and personnel to good research data management. As a consequence, such large scale and long term investments projects might even become subject to an audit of the adherence to the data management plan (and disciplinary guidelines). Notwithstanding resource issues, UK research funders are increasingly needing to consider how best to monitor compliance with research data policies.

Thus, the Freedom of Information law and the funder-imposed requirements of data management plans (DMP) are or will become strong drivers for data sharing, which work top-down. But this driver must be matched by support for researchers: The JISC-sponsored DMP-guidelines for individual disciplines have been considered helpful in this regard.

Matching the councils' policies, guidelines need to be complemented by tools, systems and teaching materials (this being the role of the DCC, Digital Curation Centre) to help researchers implement the plans. Hodson emphasises that this is the principle of all projects the JISC Managing Research Data Program is funding.

However, he adds that beyond funders' requirements there should also be a positive message to convince scientists of the benefits of data sharing – such as increased citation[48] - and advocates a systematic collection of examples of these benefits.

[48] Piwowar HA, Day RS, Fridsma DB, 2007 Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308

2.3.9 Neil Holdsworth (International Council for the Exploration of the Sea - ICES)

Neil Holdsworth has been head of ICES Data Centre since 2007.[49] He ensures the ICES data strategy, data policy and business plan are implemented and reflect the changing needs of the ICES user community. While managing relationships with key partners in the marine network he participates in he also takes a lead role in international data standards activities. Neil Holdsworth has wide experience as Data Systems Analyst and has worked on making marine data more readily available to scientists and the public. He has developed automated systems available online to control the quality, validity and format of marine data. Since 2008 he has been an assigned member of the Marine Observation and Data Expert Group, MODEG advising the European Commission in Brussels

ICES – International Council for the Exploration of the Sea

The International Council for the Exploration of the Sea coordinates and promotes marine research on oceanography, the marine environment, the marine ecosystem, and on living marine resources in the North Atlantic. Members of the ICES community include all coastal states bordering the North Atlantic and the Baltic Sea, with affiliate members in the Mediterranean Sea. ICES is a network of more than 1600 scientists from 200 institutes linked by an intergovernmental agreement (the ICES Convention, 1964) to add value to national research efforts and gather information about the marine ecosystem. This information is developed into unbiased, non-political advice. The 20 European and American member countries that fund and support ICES use this advice to help their governments and international regulatory bodies manage the North Atlantic Ocean and adjacent seas.

ICES maintains some of the world's largest databases on marine fisheries, oceanography, and the marine environment, and its Data Centre is part of a global network of distributed data centres. ICES operates an open access data policy adopted by the ICES Council in 2006. This Data Policy conforms to the IOC Oceanographic Data Exchange Policy.[50] ICES publishes its scientific information and advice in open accessible reports, publications, its own Journal of Marine Science and on the ICES website.

What was the beginning of ICES - the initial sharing of information and data?

The beginning of ICES goes back until 1902 (Inaugural Meeting in Copenhagen), where a group of dedicated scientists started to share information and data to know more about fish distribution, oceanography and the marine ecosystem beyond borders.

The founding members were Denmark, Finland, Germany, The Netherlands, Norway, Sweden, Russia and United Kingdom. The initial exchange of information and data was driven by scientists - not politics! It started with sharing log books of fisheries, landings and with collecting information consistently over a period of time to make more information available – nowadays in digital format. The Copenhagen declaration 1964 – the ICES Convention - as an official intergovernmental agreement solidified ICES finally as an advisory board to add value to national research efforts.

[49] http://www.ices.dk

[50] http://www.iode.org/index.php?Itemid=95&id=51&option=com_content&task=view

What are the main obstacles in sharing data internationally?

International guidelines seem to be too complicate and not very practical. People tend to follow traditional rules and standards based on national or federal regulations. But these regulations are diverse; hence the national conventions can limit the ability for international cooperative data sharing.

But we cannot criticize these national conventions for not being generally cooperative or homogenized on a European level since the main funding comes from national dedicated funding of regional/national driven programs.

ICES data sharing today - Why is it not as good as it should be?

Scientific disciplines, to some degree, still work separately, since their data traditionally had particular uses unique to themselves (1950s to 1980s) e.g. fisheries and physical oceanography. These disciplines grew side by side, but separately, in science as well as within ICES. Biologists in particual are less advanced in wide-scale data sharing. They have a more regional, hence small scale, approach to their research compared to e.g. oceanographers or meteorologists. Biologists need to couple their investigations on a higher scale to tackle comprehensive global environmental problems.

Later on in the upcoming ecosystem approach a fundamental need for integration and consequently data sharing emerged. But different standards and guidelines and distinct traditions still exist today and need to be resolved. During the 1980s to 1990s scientists and politics still did not meet on a practical level.

But since the formation of OSPAR, HELCOM and the EU integrated and cross-border environmental data are increasingly needed everywhere.

How does ICES help to overcome obstacles in data sharing?

ICES follows a top down and bottom up approach. On the one hand intergovernmental and political alliances like the EU, OSPAR and HELCOM need special advice and integrated approach. ICES helps to answer their questions and gives advice. On the other hand Scientists themselves organized in ICES working groups bring up new questions and solutions across disciplines and interact with other science groups. Therefore in ICES both parties find a meeting and communication platform.

Funding hindrances still to overcome

So far national, regional or local funding does not consider international concerns adequately, but should do so right from the beginning. Furthermore R&D funding should not only produce immediate short-lived results, but should generate and steer sustainable integrated research efforts. This is still a tremendous task.

Strategic barriers for data and information sharing

1. Traditionally disciplines developed separately and differently. Hence, many problems in communication, standards and mutual understandings exist. Therefore more interdisciplinary working and standardization groups and education programs are needed.

2. National and regional competitiveness still exists. Hence, protection of national interests, resources and political power are causing distinct barriers for international data sharing. Often national funding interests overrule international integrative approaches. Concerning mentality and legality, there is still a certain European North-South divide to overcome, not to mention the adaptation of Eastern Europe.

3. Another severe cause restricting Open Access to data are legal problems on national and international levels such as ownership, copyright and protection of once acquired possession. Slowly we are overcoming these obstacles through international and interdisciplinary committee work e.g. Open Access data policy adopted by the ICES Council in 2006 conforming to the IOC Oceanographic Data Exchange Policy.

4. Traditionally research side and political advisory side did not develop adequate communication structures. This led to a misbalance between scientific expertise and political decision making. This resulted in lack of cross-border information exchange and data sharing infrastructures. This is addressed today via international expert groups and interdisciplinary commission work. The outcomes of these activities need to be realized more effectively.

5. In the wake of international and national integration programs the burden of reporting and delivering of data has become huge. It may cause a hindrance to properly addressing those who must be reported and what must be delivered. There are too many addressees to be reported to. This seems to be caused by an overall steering problem.

2.3.10 Peter Igo-Kemenes (European Organization for Nuclear Research - CERN)

Peter Igo-Kemenes, of Hungarian origin, holds a PhD in physics from the University of Leuven (Belgium). After initial positions at Heidelberg University (Germany) and CERN (Geneva, Switzerland), he spent two years as a visiting professor at the Columbia University (New York). After his stay in the US he returned to Heidelberg University, finished his "Habilitation" (1984) and joined the OPAL experiment on the LEP at CERN (the pre-cursor to the LHC) where he spent the larger part of his scientific career. During the mid-90-s he became the leader of the LEP Higgs Working Group, which had the mandate to combine the data of the four big LEP collaborations ALEPH, DELPHI, L3 and OPAL in matters of Higgs boson search. Currently he holds a professorship at Gjøvik University College in Norway and acts as Senior Scientific Advisor to CERN, mainly in matters of Open Access publishing and long-term data preservation. Recently he participated in the two European FP7 projects: Parse. Insight (Permanent Access to the Records of Science in Europe) and SOAP (Study of Open Access Publishing) and helped laying down the foundations of the SCOAP3 project (Sponsoring Consortium for Open Access Publishing in Particle Physics).

Highlights - success stories in data exchange:

The LEP Higgs Working Group worked on statistically combining the data of four large-scale experiments with the purpose of improving the overall sensitivity of the search for the Higgs boson. This enterprise lasted for about 10 years and resulted in essential publications which marked the end of the LEP era for the Higgs boson searches.

The data have been kept alive since the end of LEP (in 2000), together with the analysis software, and are currently reformatted and stored such that it can be reused in combination with future search data. The data will be published soon on INSPIRE. Reanalysis of the data is anticipated in the near future, for example combination with similar data from the Tevatron accelerator experiments (Fermilab/USA), which will tie up with the subject where LEP left it. Increasing interest in the LEP data can also be anticipated from the LHC experiments, which are in their start-up phase.

Another success story is the combined analysis of two datasets, produced by two experiments, separated by about 20 years. The data have been used in a single analysis to determine the energy dependence of a fundamental physical parameter, namely the strength of the so-called "strong" (or nuclear) interaction. For the low energy part, the results from the JADE experiment at DESY in Hamburg (finished in the early 80s) and, for the high energy part, the results from the OPAL experiment (LEP, CERN, finished in the year 2000) were used. During JADE there was no effort at all to preserve/ conserve data in a way that made it re-useable for such combined analysis. The success of the combined analysis relies on the dedication of two people from JADE who painstakingly studied old logbooks and computer printouts to revive the JADE data. They eventually became members of the OPAL cooperation for the purpose of producing the combined analysis. This "archaeological" work took several years but the resulting publication became a fundamental document on the subject.

Obstacles in data-exchange / data preservation for re-use in HEP

Sociological aspects: the environment of concurrent experiments dealing with similar subjects can be described as a precarious balance between competition and cooperation.

This was indeed the case within the LEP Higgs Working Group constituted by members of the four LEP experiments. Concurrent experiments do not putt down all their cards, just the minimum necessary. This may sometimes be in conflict with the full insight that is needed for producing reliable combined results. Such conflicts will certainly continue to exist when it comes to compiling data today.

Difficulties in preservation: one challenge within data preservation is of course the rapidly changing technology. The LEP data for example cannot be re-run on currently existing computing "platforms" without a major "revival" effort. In general, old hardware and software soon becomes outdated or unreadable. Migration to new platforms and virtualization of software are some of the efforts that have to be invested in for long-term preservation and re-use.

The conservation of internal knowledge and understanding of all the experimental details: without this knowledge it is very difficult to take the data and analyze it. Detailed documentation needs to accompany the data. There is a balance to be struck between the levels of detail of the data offered for conservation. On one hand, a fine "granularity" of the data requires more detailed knowledge of the exact meaning. On the other hand, a coarser "granularity" imposes severe limitations on the possibilities of re-use. Particularly for HEP experiments, dealing with very complex data, some internal knowledge will always be necessary. Even though the LEP Higgs data will be made open access (together with accompanying documentation), one should seek the expert knowledge of former LEP collaboration members, as long as they are available, for successful re-analysis.

Lowlights

The LEP experiments, which ended in 2000, did not invest the necessary effort in to allowing data to be preserved on a large scale for possible reuse. As a result, re-analysis will be possible only in some specific domains of physics. Most of the results produced during the lifetime of the experiments could not be reproduced. In order to avoid this happening again, experiments worldwide in the process of closing down try to invest in this effort and avoid a similar situation. The main initiative is in the hands of the "Study Group for Data Preservation and Long Term Analysis in High Energy Physics" (DPHEP).[51]

DPHEP (Study Group for Data Preservation and Long Term Analysis in High Energy Physics)

Thus far, there has been almost no data preservation during the experiments' time. As seen in the cases mentioned above a lot of manual work was needed to revive data. In order to avoid this happening again, DPHEP has been started by major experiments that have finished data collection (for example: the Tevatron experiments CDF and D0, experiments at DESY (Hamburg); BaBar at SLAC/US, Belle at KEK/Japan). These experiments, together with the current LHC experiments, may represent the last generation of their kind. Hence, ensuring the possibility of reuse at a later stage may become vital.

[51] http://www.dphep.org

An important aspect of data preservation is the fact that within the lifetime of an experiment one never fully exploits the data. Only the future can tell what has been overlooked. New theories, for example, can generate new interest in old data.

The effort within DPHEP is centralized. Its aim is to develop standards and methods and to work out technologies for data preservation specifically for HEP.

DPHEP is interacting with astrophysics, where the data is less complex than the HEP data. In astrophysics some standards for data exchange are already in place. HEP can learn from astrophysics even though the levels of complexity are not comparable.

The size of the effort of conserving HEP data should not be underestimated, either from the manpower point of view or from the financial point of view. Keeping data alive (migration of the data to new supports, keeping software alive...) is a huge load and it is unlikely that the experiments alone can provide for this over the long term from their research budgets.

Future perspectives

For HEP the lessons learnt should be taken into account and a parallel effort in data preservation should be made while the experiments are alive (and produce data). Such efforts probably need to be run by data preservation experts. Data preservation should not only happen after the shutdown of an experiment. The awareness is already there but actions are still lagging behind. (Positive sign: the LHC experiments, CMS in particular, are joining the DPHEP effort). In conclusion it is important to keep in mind that HEP is a very exceptional field with its huge and complex data output.

Aspects of interdisciplinary

HEP grew out of nuclear physics, which was grown out of atomic physics. All these fields are cognate to HEP. However, direct interaction and exchange mainly happens at the level of results and not on the level of sharing "raw" data. Recently cosmology and astrophysics also became kin to HEP. The goals are the same: to find the most precise description of the beginning of the universe with its elementary particles and interactions using, however, widely different technologies. Today HEP and astrophysics is merging into what is called astro-particle physics, speaking the same language.

HEP as a discipline is at the frontier of technology: each piece of equipment is a "prototype" demanding new standards from industry in fields like vacuum technology, magnets, superconductivity, laser technology, material sciences, etc.. There is a lively exchange between the industry and science on the level of development of equipment and generating spinoffs, which find their applications in everyday life. The impact on information exchange technology and medical sciences are well-known examples.

2.3.11 Leif Laaksonen (CSC - IT Centre for Science)

The main idea behind e-IRG is to provide well-prepared information and recommendations on matters in the e-Infrastructure field to a broad range of actors and stakeholders in the policy field, ranging from national governments to the European Commission. e-IRG has succeeded in contributing remarkably to the e-Infrastructure requirements of the ESFRI Roadmap research infrastructures through its Blue Paper. Leif Laaksonen believes that it is essential to create forums for an open discussion on issues relating to how to advance in this important topic. The message from him to all data related actors is to initiate a cross-cutting international forum for research data management and to initiate strong co-operation with the existing forums in e-Science and e-Infrastructure.

The e-Infrastructure Reflection Group (e-IRG) was founded in 2003 to define and recommend best practice for pan-European electronic infrastructure efforts.[52] It consists of official government delegates from all the EU countries and the European Commission. The e-IRG produces recommendations, roadmaps, white papers, and blue papers, and analyses the future foundations of the European Knowledge Society.

The main objective of the e-Infrastructure initiative is to support the creation of a political, technological and administrative framework for easy and cost-effective shared use of distributed electronic resources across Europe. Attention has been directed towards high performance and grid computing, networking, and in particular data storage, availability and access of data essential for the research process.

The former Chair of the e-IRG Leif Laaksonen describes the influence of the group:

"The main idea of this forum is to provide well-prepared information and recommendations on matters in the e-Infrastructure field to a broad range of actors and stakeholders in the policy field, ranging from national governments to the European Commission. An important collaborator for the e-IRG is the ESFRI (European Strategy Forum on Research Infrastructures).[53] This collaboration strengthens the understanding on the topics in building a research e-Infrastructure in Europe and underpins the collaborative effort involved in promoting the progress on a practical level. The e-IRG produces alternately a new White Paper and Roadmap, providing an e-Infrastructure vision for the future. Challenging e-Infrastructure issues are tackled through task forces and reports. The ultimate goal of the work is to pave the way for a general-purpose European e-Infrastructure supporting the research and researchers in Europe. The e-IRG makes suggestions and prepares recommendations for the European Commission and the Member States.

It is important to note that the open access and free availability of data is central to the research process. The public authorities have a significant amount of useful data in their registries, which should be more effectively available for research or even for creating new business. The legislation, which varies from country to country, is a severe barrier for effective utilization of that valuable data. In EU there is no common legislation for the utilization of governmental data but recommendations from the European

[52] http://www.e-irg.eu

[53] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri

Commission are slowly making their way in national legislation. Various countries are in very different phases on the path for updating of their legislation.

On the basic research level the lack of a habit among researchers to deposit their own research data is still seen as a central barrier for data sharing and the understanding of the benefits of doing so must be promoted. There are of course huge differences between scientific disciplines; as some scientific disciplines are already very aware of the value of their data and a lot has been done on promoting efficient sharing and utilization of data. A reason for the lack of data sharing tradition in some fields may also be in the requirements set by the funding organizations as these might only require that the research publications must be published in open access archives but not the data which the results are based on. However, this is slowly changing as the infrastructure is providing better services for enabling these types of activities.

The research fields, not handling personal data or not involved in the commercializing of data, are the most advanced in data sharing. Drivers for the data sharing are usually very practical like the high costs of local data and the needs to store large amounts of data. Some fields also have an advanced tradition to deposit and share data.

Data sharing should also be promoted through a common system where research projects and researchers receive credit for sharing their data. Research projects should be supported to include planning also for their data management efforts already at the project starting phase. The projects should be aware to include the work and device costs involved in managing and maintaining these valuable data sets when applying for funding.

A further barrier to fully utilize the possibilities in employing data sharing infrastructures is to ignore the customer point of view, which has to be accounted for already at the building phase.

Despite several recognized barriers there are also successes. e-IRG has succeeded in contributing remarkably to the e-Infrastructure requirements of the ESFRI Roadmap research infrastructures through the Blue Paper. Leif Laaksonen sees that it is essential to create forums for an open discussion on how to advance this important topic. The message from him to all the data related actors is to initiate a cross-cutting international forum for the research data management and to strongly co-operate with the existing forums in e-Science and e-Infrastructure."

2.3.12 Peter Lemke (Alfred Wegener Institute for Polar and Marine Research - AWI)

Prof. Peter Lemke is head of the Climate Sciences Division at the Alfred Wegener Institute. He is also Professor of Physics of Atmosphere and Ocean at the Institute of Environmental Physics at the University of Bremen.

He has been working on the observation and modelling of climate processes since the mid 1970s, particularly on the interaction between the atmosphere, sea ice and the oceans. He has participated in seven polar expeditions - mostly as chief scientist. Due to poor monitoring conditions in the Polar Regions, he was committed to developing new measuring technology, especially for remote sensing.

Peter Lemke was an active member of the Joint Scientific Committee for the World Climate Research Program (WCRP) 1995 - 2006. This is the highest international committee for climate research and he acted as its chair for six years. Furthermore he heads REKLIM, the climate initiative of the Helmholtz Association, in which eight research centers are collaborating; a big challenge concerning data sharing and model development.

Professor Lemke was instrumental in preparing the World Climate Report of the Intergovernmental Panel on Climate Change (IPCC), which was awarded the Nobel Peace Prize in 2007.[54] In June 2010 he was announced as one of the experts for IPCC's Fifth Assessment Report, where he will act as Review Editor responsible for the chapter on the Earth's cryosphere.

Why and how did real data sharing start in your community?

For me personally data sharing started right with my doctoral thesis. For that task I had to digitize analogue paper maps (sea ice charts). After completion of the work, the digital data set was submitted to the World Data Center for Glaciology in Boulder, USA, for use by the wider scientific community.

Within the meteorological community data sharing started with the beginning of the international coordination of weather forecasts through the International Meteorological Organization in 1873. In other environmental disciplines a data sharing process had been established since the first Geophysical Year assembly in 1957/58.

Personally already in 1979 I was urged by my supervisor to feed our data into the World Data Center for Glaciology in Boulder while taking part in the World Climate Program (WCP) implemented by WMO (according to convention by the International Council for Science - ICSU) - not least by our deep integration in this international research program. The WCP-data sharing endeavour turned out very positive to stimulate collaborative science right from the beginning, since the repository of the WDC of Glaciology digests globally huge amounts of relevant data for international research and meteorological services, e.g. also from ESA and NASA. Most of its data is open access. Even the NASA is a declared principle data investor to that WDC.

[54] http://www.ipcc.ch

To what extent was data sharing an essential issue in preparing the IPCC-report?

The mission of the Intergovernmental Panel on Climate Change (IPCC) is to determine at regular intervals the state of the climate system and its impacts on ecosystems and human society and to point out potential political countermeasures. The IPCC was instituted by the World Meteorological Organization (WMO)[55] and the United Nations Environment Program (UNEP)[56] in 1988 when the possibility of global climate change became evident. The IPCC does not conduct its own research, nor does it provide data. Hence to prepare the IPCC-report we did not request data directly - if at all, only by means of control or adjustment. Mostly we compiled relevant scientific evidence for comprehensive analysis. The IPCC-assessment is mainly based on peer reviewed and published scientific/technical literature, which is evaluated in a thorough, objective, free and transparent manner (http://www.ipcc.ch).

What kind of positive and negative experience in data sharing do you know in climate research?

Weather forecasting data have been shared already for about 150 years as an imperative necessity: because we need to prepare for any weather phenomena in time. Of course weather is not constrained by national borders. Very early on, people learnt that it is most important to know the weather upwind in London to predict next day's weather in Hamburg. Under these circumstances data sharing works basically, because fast communication exists regularly via telegraph since the first worldwide operating meteorological service has been established. Since meteorological data are naturally distributed worldwide a centralized weather forecast system was inevitable and the International Meteorological Organization (IMO) existed from 1873 until it was succeeded by the now well established, WMO in 1950. In that area, global data sets are compiled and distributed constantly. Since weather data had been exchanged worldwide right with the historic upcoming of emerging global communication techniques data sharing in meteorology has a long-standing tradition. It works out very well compared to other disciplines.

In contrast, experience shows that barrier-free access to, e.g. hydrological data, is still causing huge problems. These data are needed to relate data collected in the field (ground truth data) with remote satellite data, for their evaluation and modelling; especially for disaster risk reduction. Actual hydrological data are subject to state and national administration. If you may gain access to these data at all, it is years later, because they are of national strategic importance (resources, agriculture) and are therefore restricted. In this field, international open access data release does not seem to be possible.

In contrast, free access to data from the international World Climate Research Program (WCRP) is the normal case since its establishment in 1980. This very successful program is funded by the World Meteorological Organization (WMO), the International Council for Science and the Intergovernmental Oceanographic Commission of UNESCO. It supports progress in the prediction capabilities of operational centres in extended weather and seasonal forecasts as well as longer-term variability and climate-change projections. Scientists organized in the WCRP provide a major part of the scientific

[55] http://www.wmo.int

[56] http://www.unep.org

material assessed by the Intergovernmental Panel on Climate Change (IPCC) in its advice to the UN Framework Convention on Climate Change. These activities form the scientific basis for adaptation to climate change and for developing mitigation strategies that are eventually implemented on international and regional levels. (http://www.wmo.int)

Despite being very well organized internationally the WCRP does neither gain its own research money nor its own funding programs. But the program did turn out well as a working platform for meetings and for international data exchange. For example, WOCE (the World Ocean Circulation Experiment) was a very successful project, especially concerning data sharing. In that activity, international data bases had been implemented and substantial digital world atlases had been created.

But the urgent and essential adjustment of meteorological research and models with hydrological data collected on location is hardly possible because of insufficient access to local data like soil moisture, discharge, etc. neither on a national scale nor on an international scale.

Not alone for the IPCC-report, but for world-wide climate research and the bigger picture we urgently need to couple global meteorological data with regional hydrological ground truth data to run realistic climate models and predictions.

Data fit for re-use? Problems with metadata and homogenization

In meteorological and climate research, metadata are very important, and generating them always implies high effort. Generally this works out well for the World Data Centres. Also the German BSH (Federal Maritime and Hydrographic Agency), for example, is well positioned. At the National Snow and Ice Data Center (NSIDC) in Boulder (USA) re-usability due to appropriate metadata handling works out well.

The high effort in handling diverse calibration methods and standard verification procedures hampers data re-usability in meteorology. Even nowadays this is still causing problems for data archiving. Therefore it is essential, that only adequate climate institutions are specialized to homogenize and archive climate data. To exemplify this relevance - at the WMO historical data are reprocessed and converted to current standards. Only within this organization scientific specialists can interpret these historical data properly and implement international standards.

Furthermore quality control starting right with the individual field measurements is indispensable for re-usability of data. Even during compiling the IPCC-report data offsets had been noticed while aligning data from diverse measurement devices. Another example is that overlap and sensor ranges of more than twenty satellite operators have to be managed and the data itself need to be calibrated constantly and with each new satellite sensor. The standardization of weather, water and climate data and meta data is essential to ensure an orderly and efficient share and use of the information between WMO Members from the provider to the user. Hence tasked expert teams develop and maintain the relevant standards, and develop guidance for their implementation.

Furthermore the early installation and improvement of WMO's Global Telecommunication System (GTS) enables the usage of all weather service data world-wide. It plays a vital role in facilitating the flow of data and processed products to meet

requirements in a timely, reliable and cost-effective way, ensuring access to all meteorological and related data, forecasts and alerts. This secured communication network enables real-time exchange of information, critical for forecasting and warnings of hydro-meteorological hazards.

Final remark

Since meteorologists are strongly involved in global joint projects, they are *a priori* interested in sharing knowledge and data. Otherwise weather and climate research would hardly be possible. Hence data sharing became an implicit commitment, although no personal control exists either. A data sharing ethos developed very early due to the instantaneous need for action preceding natural weather hazards. And of course prediction of any weather condition implies global information and data exchange. In summary, the long established data sharing of the national weather services provides the basis for our global climate research.

2.3.13 Caroline Liefke (GalaxyZoo)

Carolin Liefke, born 1981, has been fascinated by the night sky since she was thirteen years old. From 2000 to 2005 she studied physics at the University of Hamburg, and specialized in astronomy. Subsequently, she worked on stellar activity and X-ray astronomy at the Hamburger Sternwarte for her PhD. Carolin is an enthusiastic amateur astronomer and member of several astronomy associations. For more than ten years, she has been involved in astronomy outreach and education. In March 2010 she turned this passion into a profession, now working at the 'Haus der Astronomie', Heidelberg's center for astronomy education and outreach.[57]

She maintains the German version of GalaxyZoo[58] and other citizen science projects in the Zooniverse, and other education and outreach activities in the field of astronomy. In the Zooniverse projects, large amounts of scientific data are handed over to laymen for special analysis tasks that require a human brain to solve, such as classifying galaxies, searching for exoplanet transits, or finding unknown asteroids.[59]

What is Caroline's experience with research data?

When studying physics she coded tools for data reuse. She is a "real research data re-user" who has searched and integrated a lot of existing research into her research projects. Even though data sharing is well advanced and data is handled "in the open" she encountered difficulties when searching for it. There have been cases when, she found out later (even after a projects end), that there were more datasets that could have contributed to the findings of her research. Thus, she thinks that some of her research papers could have been improved or accelerated by better discoverability. She knows similar stories from friends and colleagues and she is glad that the challenge of discoverability is now one of the aspects that are being worked on by the Virtual Observatories (VO).

Regarding her participation in Galaxy Zoo, it is important to address the definition of data sharing. The data sharing is limited in Galaxy Zoo in the sense that the participants do not play an active role in the sharing process. They are presented with pre-processed data and a very special task. However the raw data the project is based on are shared among the scientific community.

Her view on data sharing in the astronomy in general

According to Caroline there is a lot of data sharing in the dynamic field of astronomy. Research information is handled very openly. Data management is usually run by the missions and their institutions themselves. In the first year after the data production data access is limited to the researchers who proposed and participated in the particular project. Afterwards, the data is open access. Again, the challenge is not so much the actual data preservation, but rather the discoverability of the data. A major and ongoing initiative is the Virtual Observatories (VO). This initiative will facilitate easier discoverability of research data, more sophisticated data mining, and more complex automated analysis.

[57] http://www.hausderastronomie.de

[58] http://www.galaxyzoo.org

[59] http://www.zooniverse.org

Her highlights and lessons learnt from research data sharing are related to the challenge of discoverability in the data deluge. The major challenge of lost data, or data that appears to be lost, is being tackled by the virtual observatories. The initiative also takes care of "old" datasets from finished projects, preserving them and making them available via their interfaces. She thinks that the major challenge for the oncoming years is data management and presentation of huge projects and with it the management of the data deluge. The latter usually requires advanced automated processing and selection for the data archive.

2.3.14 Karin Lochte (Alfred Wegener Institute for Polar and Marine Research - AWI)

Prof. Karin Lochte, a biologist, is now director of the Alfred Wegener Institute, a member of the Helmholtz association. Since she has served as member of the German Science Council (Wissenschaftsrat), as Vice-President of the Helmholtz-Association, and in other roles, she is in a position to assess realistic approaches to policies regarding data.

Of course, she has been an active researcher herself – in particular, the principal investigator of a project, ADEPD, which had the objective to „build up a joint data base for deep sea biological and geochemical data from a variety of sources". To this end, "1775 published and unpublished data sets were collected in two years" .[60]

It was this project which made amply clear to her, that data known to "exist" were mostly not in the form suitable to share them - and thus, they would not be shared easily. The ADEPD solution to this was to pay research groups to prepare their (existing) data for incorporation in the ADEPD database. Lochte concludes that, as long as data delivery is not part of evaluations and not a recognized part of a scientist's reputation, one must be prepared to pay for the sharing of data.

In her role as head of the DFG Senate Commission on Oceanography, she has come to the conclusion that there must be a commitment when funds or ship time are granted, that the data must be delivered within a specified time frame. Only then, full payment of grants should be allowed (data delivery would need to be controlled). In research proposals there should be a data management plan and resources planned to implement it.

The Senate Commission will require that there must be a delivery of "early/fast data", together with the "technical" report on each cruise and it will ask when the remainder of data will be delivered. Lochte does not regard, however, a uniform deadline, such as 1 or 3 years, possible for all disciplines.

As she observes scientists in her field struggle with access to nominally available data, Lochte concludes: Beyond the technical accessibility of individual datasets, the optimum would be just one point of access to all data, which would be simple to use (e.g.: no database know how, no "way of thinking" required) and "fool proof". But she suspects it is wishful thinking to ask for Google-like simplicity when one would like to ask for "chlorophyll data in the Atlantic at 200 meters depth"

Therefore, for the time being, in order to overcome the stumbling block of familiarization with tools and portals for data search she sees the need for support (e.g.: helpdesks) and training – both at the institutional as at some central level. To render training effective, she advises some standardization in access to databases.

[60] http://en.wikipedia.org/wiki/ADEPD

2.3.15 Eberhard Mikusch & Katrin Molch (German Aerospace Center - DLR)

Eberhard Mikusch heads the department of information technology at DFD. He received the Diploma Degree in computer science and has lead several software development projects for ground segment facilities in the aerospace domain. Responsible for multi-mission data and information management, he represents DLR/DFD in the European Earth observation long term data preservation working group.

Katrin Molch is responsible for the DFD data services. She holds a Master's degree in Geosciences and has been working as a remote sensing scientist with European and North American research organisations. Leading the D-SDA Services team, she is working towards improving accessibility to the DFD data holdings and increasing the use of the data by a broadening customer base.

The German Remote Sensing Data Centre (DFD)[61] is an institute of the German Aerospace Centre (DLR).[62] DFD and DLR's Remote Sensing Technology Institute (IMF) together comprise the Earth Observation Centre EOC, which is the centre of competence for earth observation in Germany. DLR is a member of the Helmholtz Association, Germany's largest scientific organization.

DFD is involved in national, European, and international Earth observation missions. The data centre supports science and industry as well as the general public. With its national and international receiving stations, DFD offers direct access to data from missions and safeguards all data in the so-called German Satellite Data Archive (D-SDA) for long term use. The DFD operates thematic user services, for example the World Data Centre for Remote Sensing of the Atmospheric (WDC-RSAT), and the Centre for Satellite Based Crisis Information (ZKI).

Operational Earth observation missions continuously generate huge quantities of satellite data representing the momentary condition of the land surface, the oceans and the atmosphere of our planet. "We have to manage a large amount of data. In 2008 we stored 200 *terabytes*. In 2011 we store already 400 *terabytes* of data. And the mass of data is growing exponentially," says Eberhard Mikusch. This data is part of the world's cultural heritage. "The data of an ozone concentration map of a specific day can never be repeated," says Mikusch. His colleague Katrin Molch, responsible for the data services, added: "The long term preservation of this data is necessary to grant researchers access to this data now and in the future." There are more than *30 employees working at* DFD to enable permanent accessibility of remote sensing data.

Satellites have only limited capabilities to store the recorded data on-board. In general, they have to be in direct contact with a ground station to download their data in real-time. The so-called payload data are received, archived and processed around the clock and are used to generate a wide range of value-added information products reflecting spatial and temporal relationships. They are made available for the most varied applications, e.g. surface temperature maps, digital elevation models, ozone concentration maps and multispectral images. Depending on the target group, data are processed differently. The processing of the data to high-quality information products is very complex. "The preparation of information products for special needs is part of our

[61] http://www.dlr.de/caf/en/desktopdefault.aspx/tabid-5278/8856_read-15911/

[62] http://www.dlr.de

mission. For example, in scientific contexts, other data is needed than in the field of crisis management," says Molch. In compiling the information products, data from different sources are used. "We would like to see more data openly accessible. In compiling the information products, we rely on data from other sources, which are not always easily available" says Mikusch.

An essential requirement for the sharing of remote sensing data is data curation. Acquired payload data has to be preserved and handled far beyond the initial satellite platform's lifetime, since its value increases with age, for example in the domain of global change monitoring. "At DLR we have developed a multi-mission data management system, called DIMS - Data and Information Management System, to handle the data accumulating in such quantities and diversity. Long-term archiving and reuse are coupled with each other. DIMS supports the whole data lifecycle, i.e. especially reprocessing is part of the concept" says Mikusch. Preservation and migration is required not only for data but also for the systems handling the data.

The DFD offers special collections of satellite data for scientific use. Most of the collections are openly available. For example, the World Data Centre for Remote Sensing of the Atmosphere (WDC-RSAT) offers scientists and the general public a continuously growing collection of atmosphere-related satellite-based data sets.[63] The data come from a variety of missions in which the DFD organized the data management.

The WDC-RSAT is a member of the World Data Services (WDS). The WDS supports the International Council for Science (ICSU) mission to ensure the long-term stewardship and provision of quality-assessed data and data services to the international science community. The WDS has adopted a data policy: "There will be full and open exchange of data, metadata and products shared within WDS, recognizing relevant international instruments and national policies and legislation" (excerpt).

"The WDC-RSAT data is freely accessible," says Molch. If data obtained from WDC-RSAT is used as the basis for a publication or for further dissemination, it is necessary to acknowledge and reference its origin. WDC-RSAT is using the Digital Object Identifier (DOI) to ensure persistent identification of data sets.

With regard to the commercial market for satellite data, some data is only available for scientific purposes. The huge amount of data must be considered. Often special software tools are needed to work with the data. Hence Molch states: "Working with the data has to be learned." So DFD provides opportunities for the next generation of scientists by offering internships as well as jobs for young scientists.

"A major challenge for the future is financing of data management after the end of a project," says Mikusch. "Data sharing isn't possible without long-term preservation. Funders should recognise the need for sustainable financing of data management. In order to answer important questions about future climate change, scientists need as far as possible open access to our archives."

[63] http://wdc.dlr.de

2.3.16 Tommi Nyrönen & Andrew Lyall (European life science infrastructure for biological information - ELIXIR)

The ELIXIR project has a strong economic driver, as the costs of data regeneration are huge compared to the costs of data preservation. In addition, the scientific work done on the systems level by large distributed teams in basic biological research and various applied fields of the ELIXIR project is strongly dependent on a knowledge of DNA sequences and integration of different data obtained by different technologies.[64]

The ELIXIR project will finish in 2011. It is a four-year preparatory phase funded by the EU's Seventh Framework Program (FP7) as part of the European Strategy Forum on Research Infrastructures (ESFRI) process. ELIXIR's mission is to construct and operate a sustainable infrastructure for biological information in Europe to support life science research and its translation to medicine and the environment, the bio-industries and society.

The first mentioned driver for the ELIXIR project has been the massive amounts of data which is generated by high-throughput DNA sequencing technologies to reveal the generic code of life and the meaningful integration of new types of data. The growth of the amount of data has been unprecedented and it is estimated that it will accelerate. The amount of data is estimated to increase up to a million times the current rate in about 10 years. The observed expansion of data volume has been supra-exponential.

Knowledge of DNA sequences has become indispensable for basic biological research and in numerous applied fields such as diagnostics, drug development, biotechnology, forensic biology and systems biology. In addition, the way of working has also changed, from individual research groups to large distributed teams with needs for common access to extensive common data resources. A strong trend in research is to understand organisms, diseases and the behaviour of human beings at the systems level. This leads to the need for integration of different data obtained by several different technologies in meaningful ways. This other driver combines the need of computational resources and analysis tools with data storage.

The third driver to data sharing is the huge difference between the costs of data generation and preservation. Data generation and its description and quality control are laborious tasks for researchers. It is estimated that the cost of data preservation are about one per cent of that of data re-generation. Thus, there is a strong economic driver for data sharing.

The vast demands of collection, curation, storage, archiving, integration and deployment of heterogeneous biomolecular data cannot simply be handled by a single EU state, but requires strong international collaboration and coordination in Europe. There is a pressing need for a common infrastructure, and ELIXIR is the project to coordinate all up-coming scientific and technical issues in the construction of this.

In general, remarkable investments have been made in life sciences in Europe, but there is no coordination or common strategy to obtain any synergy or savings between the states or the projects. One barrier for data sharing has been the reluctance of most countries to invest in providing global data infrastructures from their national funds. If

[64] http://www.elixir-europe.org

considered at European wide level, this has led to a situation of largely fragmented resources.

Another barrier to data sharing is the fragmentation of scientific communities and the lack of collaboration between various biomedical and biological disciplines. This is relevant for both disciplines closely related to each other as well as, for example, technical and biology related fields. There is a severe lack of interdisciplinary researchers, especially between technical and non-technical fields. These are very demanding positions for which there is growing demand. Interdisciplinary researchers or professionals are necessary within infrastructures as all researchers cannot be required to have a broad knowledge on various fields.

A solution for Europe-wide technically and economically efficient data sharing is the well-coordinated construction of distributed infrastructure. This will enhancestorage capacity and make it easier for decision makers to invest national funds in global infrastructure. However, it sets remarkable requirements for common working models as well as legal and governance issues.

The integration of several nodes located in member countries is an essential technical task and must be realized in the most optimal way. In addition, existing computation resources in Europe need to be combined with data infrastructure, which will cause a demand for an increase of computation capacity.

In future, continuous training for the diversifying user community in the optimal use of the developing infrastructure and tools is a key role in reaping the benefit of data generation and preservation by a common European-wide infrastructure. The rapid development of various analysis tools and their integration is a huge challenge and the ELIXIR project has a key role in progressing this. Similar kind of drivers and barriers to data sharing affect tools sharing. However, a lot has been learned through working with data sharing. The efficient utilization of data generated is based on good description using standards where possible. The ELIXIR project needs to coordinate the development, implementation and deployment of standards, as well as common vocabularies and ontologies.

The ELIXIR project is making a great effort to share biological data by building a technically and economically efficient Europe-wide infrastructure for a diversified interdisciplinary user community. The prerequisite of the success of the ELIXIR project in its demanding task is sustainable funding from its member countries. A lot will be learned during the process and the experience can be utilized in other fields to promote their data sharing and its remarkable influence on society.

2.3.17 Finnish task force for utilization of electronic data in research

In 2009 the Finnish Ministry of Education and Culture created a national cross-sectored task force to draft a roadmap for Finland to develop the availability and preservation of data resources to be used in research, including e-infrastructure solutions, as well as a proposal concerning national division of responsibilities, cooperation and coordination among various actors. In addition the task force steered a national project related to electronic data or information resources for research. The task of the project was to form an overall understanding of the current situation in Finland and to compare it to international recommendations and arrangements of other countries. The task force finished its work at the end of 2010 with a roadmap to be evaluated by several related actors in Finland.

The background of the task force was a roadmap of national research infrastructures and the decision of Research and Innovation Council to find out what to do and how to proceed to promote the utilization of data resources. The project found that the framework for the utilization of various data resources in research in Finland is remarkably larger and more complex than was thought when planning the project. Currently, the subject is even more topical as several essential actors have been active in and already achieved goals relating to the area.

The project found that Finland has a notable collection of unique, public sector information, including extensive registries which are largely useful for research though not initially gathered or produced for research purposes. In addition, Finland has, by international standards, high quality knowledge and skills in the research fields that produce these vast amounts of electronic data. However, public sector information as well as research data are currently difficult to find, to access or to utilize.

From the user point of view, there are problems related to scattered storage and management of data, in addition to lack or insufficiency of metadata, quality assurance, and incompatibility of formats and tools. Complex legislation, interpretation of information security or privacy protection, unclear terms of use, high costs of use, as well as time and difficult processes for obtaining data from public sector actors not prepared to serve researchers, are severe barriers for the utilization of public sector information. A common problem is lack of infrastructures and services that support the utilization and sharing of the data.

From the data producer's or maintainer's point of view, especially for public sector organizations, there are currently contradictory requirements and targets for their operation. Commonly, for research organizations and public sector organizations, there is a lack of data policies, practices and culture of sharing data for research or open use. There is also a lack of incentives, capabilities, resources, funding and infrastructures.

Barriers at a general level are lack of coordination and co-operation of various organizations, limited resources, and absence of perseverance. There is no steering of the ensemble, compatibility of systems or role differentiation of various actors. The general atmosphere does not encourage data sharing. There are no incentives and requirements set by research funding bodies. Finally, data deluge and diversity, as well as the definitions and descriptions needed for metadata production, do not make data sharing any easier.

The competitiveness of Finnish research requires a strong commitment to the building of an information infrastructure and to the strengthening of the related knowledge and skills. These also form the basis for international research co-operation, innovativeness and the enhancement of equal opportunities for data usage between researchers.

The task force has formulated its vision to support Finnish research and innovation in a far-sighted manner as follows: Finland has a clear data policy supported by common e-services. Data resources generated with the aid of public funding are easily available for research and, in principle, without any charge, guided by legislation and uniform terms of use and taking the data confidentiality issues into consideration. A sustainable development and funding system for the information infrastructure guarantees that both existing and new data are sufficiently described and made available by using easy to use network services. A supportive and fair merit system supported by the funding bodies ensures that new, high quality data is added to the information infrastructure.

In order to achieve the vision, actions are required at many levels and in co-operation with and coordination between various actors, including sufficient resources for the actual work. We need a collective will for improving the availability and utilization of data resources as well as a data policy to realize this. The definitions of the data policy guide the revision of legislation, the development of common practices at organizational level, and the construction of the information infrastructure for research. The objective is that national data resources are widely available for the use of entire society.

The most crucial actions to be taken are: the expression of the collective will in the government platform, establishing a cross-sector coordination group to enhance the data issues from the research perspective, commencing the planning of the information infrastructure for research, and the legislative reforms that enhance the wider utilization of data resources.

Some more detailed actions are to make an inventory of public sector data available for research purposes, to proceed compatibility of various techniques, to develop uniform terms of use and adopt suitable licenses, to draft common principles for research financiers and public sector organizations, and to enhance capabilities on data issues.

Planning of the information infrastructure, including common services, needs to be started as early as possible. This must be done in close cooperation with thematic end-user communities to find out their common needs. Planning for long term preservation of research data needs to commence also.

2.8.18 Heather Piwowar (National Evolutionary Synthesis Center - NESCent)

Heather Piwowar is currently Postdoc associated with three different projects and institutions: DataOne and Dryad and the National Evolutionary Synthesis Center, NESCent. She obtained a PhD focused on research data sharing within the biomedical field and is now very much interested in patterns of research data reuse. Before her career in research, she worked in different positions in start-up companies in the US.

What is Heather's experience with research data?

Heather Piwowar's interest in research data began when she tried to reuse research data for a project and failed to find the research data she was looking for. She then started to study data sharing.

Her research so far has mainly focused on one type of research data: gene expression microarray data. She chose this data type because the data sharing standards and infrastructure within this discipline were already well established, but the archiving of gene expression microarray data was not yet universal practice.65 It is a particularly interesting data type because it is perhaps more typical of individual investigator-driven research than the biomolecular data stored and handled via Genbank, ENA etc. Gene expression microarray data are collected under a wide range of experimental conditions, on a variety of incompatible platforms, and undergo variable processing steps.

Dryad

One of the projects she is working on is the data repository Dryad. It accepts only data that is associated with published research and is open to different fields in biology. To serve more communities and to facilitate easy reuse of the materials, it will "handshake" with the main existing databases in molecular biology such as Genbank.

One of the issues that is being discussed at Dryad is the sustainability of the project. It is currently funded under an NSF grant, but it needs to have a sustainability plan which extends beyond the project's end. Different business models have been studied.66 It is likely that Dryad will begin charging journals for data submission in 2011.

Data sharing: about the carrots…

Apart from her work with Dryad, Heather has been studying data sharing in general as she thinks that currently there is a lot of guesswork involved , but very few "real" numbers on data sharing are available. According to Heather, the researchers' hesitation is a key barrier in research data sharing and in order to convince researchers to share their materials it is important to show them some numbers. She has attained some compelling results: first of all she found evidence of a citation advantage associated with research data sharing: by studying 85 papers, she found out that referencing openly available research data in an article is associated with a citation benefit of 70%.[67]

[65] Piwowar H.A (2011): Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data. PLoS ONE 6(7): e18657. doi:10.1371/journal.pone.0018657

[66] Beagrie et al. (2010): Business Model and Coste Estimation – Dryad Repository Case Study. http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/beagrie-37.pdf

[67] Piwowar H., Day R., Fridsma D. (2007): Sharing Detailed Research Data Is Associated with Increased Citation Rate,

When studying the data sharing of microarray data over time, she found that there has been an increase in the proportion of gene expression microarray datasets that have been deposited into public archives between 2000 and 2009, but the rates seem to be plateauing at about 45%. Her analysis suggests that the NIH policy requiring a data management plan for large grants is not associated with an increase in public data archiving.[68] [69]

Researchers who had already shared their data once are more likely to share their data again. Interestingly, researchers who publish in open access journals were also more likely to share their data.[67]

In another study, Heather and co-authors investigated data policies and practices in journals on the environmental sciences. They studied 500 articles across 6 journals, discovering that data citation policies were rarely articulated and lacked standardization. Even when policies are followed, they lacked attribution discovery.[70] Her research suggests that journal data policies are strongly correlated with data sharing behaviour.[67] [71]

Thus, Heather believes in the carrots that need to supplement the sticks. She concludes that if researchers could see that they are cited and attributed for their data publication and that their sharing is considered in their promotion committees this would make an important incentive. Thus, one of the core activities in the oncoming years should be the provision of evidence of research data reuse, for example in the respective data repository.[72]

PlosOne. http://www.plosone.org/article/info:doi%2F10.1371%2Fjournal.pone.0000308

[68] Piwowar H.A. (2011): Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data. PLoS ONE 6(7): e18657. doi:10.1371/journal.pone.0018657

[69] Piwowar H.A., Chapman W.W. (2010): Public sharing of research datasets: a pilot study of associations. Journal of Informetrics. Volume 4, Issue 2, Pages 148-156. http://www.sciencedirect.com/science/article/pii/S1751157709000881

[70] Enriquez V., Judson S. W., Weber N. M. (2010): Data citation in the wild. http://precedings.nature.com/documents/5452/version/1/files/npre20105452-1.pdf

[71] Piwowar H.A., Chapman W.W. (2008): A review of journal policies for sharing research data. ELPUB. http://elpub.scix.net/cgi-bin/works/Show?001_elpub2008

[72] Piwowar H.A., Vision T.J. & Whitlock M.C. (2011): Data archiving is a good investment Nature, 473 (7347), 285-285 DOI: 10.1038/473285a. http://researchremix.wordpress.com/2011/05/19/nature-letter/

2.8.19 Andrew Treloar (Australian National Data Service - ANDS)

Stefan Andrew Treloar, Ph.D, linguist and now well known as technical director of ANDS (Australian National Data Service) became involved in data sharing as the architect of the institutional repository project ARROW (Australian Research Repositories Online – http://arrow.edu.au/), in 2006. Then, a scientist approached him to ask whether the Monash University repository (http://arrow.monash.edu.au/) could store and make accessible data that could be linked to a publication.

The scientist had a reason for making available some 35 GB raw data from protein crystallography: There had been a case of scientific fraud in his field, recently and he needed to provide the data as a supplement to an article in Science[73]. The urgency was because the article needed to be submitted on that Monday.

Monash's repository manager spent the weekend on the ingest of this data, raising the volume of data by two orders of magnitude and dealing with the challenge of metadata which went much beyond the "Dublin Core", used for the article until that time. He succeeded, the data was stored, the article was submitted – and the author came back, predictably, to ask for more space, actually 1-2 TB!

This request could not be fulfilled, due to technical reasons. Of interest however was, again, the argument of the scientist: These were "unsuccessful data", which he wanted to make available so that others could make "sense" of it! He had tried and failed, and now he wanted others to have chance.

In the course of discussions further reasons emerged to make these raw crystallography data available: There was a case when a sign error in an analysis program had resulted in a completely wrong crystal structure. This could have been found out easily, had there been a straightforward way to re-analyse the raw data, and if the reviewers had been able to get access to it. This argument immediately led to consideration of the problems of software writers: How can they test those complex analysis programs if they don't have ample access to raw data and results derived by other programs? Finally, the scientist related his experience with the first manuscript-cum-dataset and its reviewers: It seemed scary to seem to be asked to review the data as well – they would have a much higher workload.

Since ARROW couldn't, and the "Protein Data Bank"[74] wouldn't accept their TBs of data, the scientists moved on and created their own repository for diffraction images, TARDIS[75] . This repository, notes Treloar, is actually known to have been used by software writers.

This and other experiences (including a growing fascination with data issues) led to Treloar's move from publications to data: DART and ARCHER[76] were complex projects about handling complex data - as well about making them public as handling them "privately" in a scientific collaboration.

---

[73] Actually, the data are referenced, among other facts, in reference 32 of *DOI: 10.1126/science.1144706*

[74] Worldwide Protein Data Bank, http://www.wwpdb.org

[75] The Australian Repositories for Diffraction ImageS, http://tardis.edu.au/

[76] Andrew Treloar and David Groenewegen; ARROW, DART and ARCHER: A Quiver Full of Research Repository and Related Projects, available from http://www.ariadne.ac.uk/issue51/treloar-groenewegen/

Today, Treloar serves as the Director of Technology for the Australian National Data Service (ANDS), which strives to make Australia's research data discoverable[77]. Through ANDS, Researchers can find data in tens of thousands of data collections. Lots of other ANDS (supported) services have been found necessary to realize the vision of "More researchers re-using and sharing more data more often"[78]

Beyond abstract vision and goals, he points out his orientation by two pre-ANDS examples:

- Biologists trying to track the extinction of species in Australia (about 200 species in 200 years) used linguistics data: Fairly good data about where and when certain aboriginal languages were spoken are available. They could use the existence of a name for a species as a proxy to its existence.
- Whaling records for the Southern Hemisphere show, since 1931, the position of every whale caught. From this proxy to the extent of pack ice cover, a decrease of its extent by 25% could be derived. Although this interpretation has been much discussed, the original data are hard to come by[79].

Treloar says: "It's this kind of research which should be enabled by ANDS"

[77] Kethers, S., Shen, X., Treloar, A., and Wilkinson, R. (2010), "Discovering Australia's Research Data". Proceedings of JCDL 2010, available from http://andrew.treloar.net/research/publications/jcdl2010/jcdl158-kethers.pdf

[78] Burton, A. and Treloar, A. (2010). "Publish My Data: the design and implementation of a loosely-coupled data 'publishing' service". Proceedings of VALA 2010 http://www.vala.org.au/vala2010/papers2010/VALA2010_123_Burton_Final.pdf

[79] William K. de la Mare (1997), Abrupt mid-twentieth-century decline in Antarctic sea-ice extent from whaling records, Nature 389, 57-60 doi:10.1038/37956. The article contains a defunct URL (which contained a typographical error, antidiv instead of antdiv to begin with)

2.8.20 Karen Wiltshire (Alfred Wegener Institute for Polar and Marine Research - AWI)

Helgoland is known as Germany's only high-sea island. Indeed it's tall cliffs of red sandstone seem to rise from the middle of the North Sea. The closest place on the mainland is 65 km away, at the German coast. This is far enough to allow oceanographic and biological observations at true high sea conditions – and at the same time not so far as to make a permanent research station completely unrealistic. Thus BAH, the Biological Station Helgoland, was founded in 1892.

The first oceanographic data was made in 1873. Since 1962, data capture is continuous at the "Helgoland Roads" site off the island. Prof. Dr. Karen Wiltshire, biologist, head of the Helgoland station and vice- director of the Alfred-Wegener-Institute for Polar and Marine Research says: "This data is our gold!" This short sentence carries a lot of connotations with it.

The most obvious is the appreciation of the value of data – very much scientific insight can be gained from it. As an example, Wiltshire refers to one of her publications[80], based on this data, which gained 73 citations. She notes that it was based on a subset of the Helgoland data which came about since a single scientist at Helgoland began counting algae … and found seasonality. Only now that this dataset has evolved into a long term dataset, the dependence of the seasonality on warming can be extracted.

The second import of "gold" is purity, nobility – or, in scientific terms: quality. Meanwhile, much of the reputation of the Helgoland station and the institute as a whole rides on the quality of its data, according to Wiltshire. With respect to this reputation she is, however, worried about re-use by "anybody" (if the data were openly accessible): The data are quite complex and thus, not easy to analyse and interpretation may easily go astray if inexperienced scientists to not closely observe the "metadata" – which has already happened a number of times, even within the institute. Wiltshire is therefore contemplating an access model employed by a partner institute in the UK, which holds similar data from other regions: Access to data is granted to scientists only if they have absolved a (three weeks) training at this institute, before.

A similar questions shows up at a quite different instrument, the so called "ferry box", installed at the Helgoland roads: It provides data streams, for example on CO2 - concentration, for near-realtime monitoring. Technically, this could be made openly accessible. Wiltshire considers to do this only as the institute can provide a parallel measurement with another instrument, for calibration and more meaningful metadata. She says "Excellent datasets and knowledge about their limitations is a distinctive feature of Alfred Wegener Institute. Therefore, quality control is one of our paramount concerns."

On the other hand, access to (easily interpreted) quality controlled physical oceanography data is open, in accordance with regulations by ICES (Int. Council for the Exploration of the Sea), after an embargo of 3 years.

---

[80] Wiltshire, KH; Manly, BFJ; The warming trend at Helgoland Roads, North Sea: Phytoplankton response ; HELGOLAND MARINE RESEARCH (2004) Vol. 58 Iss. 4 pp. 269-273 DOI: 10.1007/s10152-004-0196-0 (cited 73 times)

Wiltshire Karen Helen; Malzahn Arne Michael; Wirtz Kai; et al.; Resilience of North Sea phytoplankton spring bloom dynamics: An analysis of long-term data at Helgoland Roads; LIMNOLOGY AND OCEANOGRAPHY (2008) Vol. 53 Iss. 4 pp. 1294-1302 DOI: 10.4319/lo.2008.53.4.1294 (cited 39 times)

The third – perhaps most difficult to resolve – meaning of "gold" is, of course, the more than just metaphorical value of data. It has cost more than 4 years to "sort through" and quality control the raw and primary data on the abundance and size of algae. In her opinion, giving the data "away" does require a compensation of the scientific reputation not gained otherwise by AWI. Wiltshire has had a very bad experience in this context: Part of the data could be accessed through a third party repository. At a partner (!) institute a paper was published which relied heavily on it, but gave no credit to AWI, BAH or their scientists. The partners refused to amend their paper in this respect.

Therefore, Wiltshire limits access to the algae data at this time and is even thinking about requiring co-authorships. In this case, the role of BAH might include production of data extracts "ready to interpret" for the article in question – which would reduce the risk of misinterpretation.

2.8.21 Stefan Winkler-Nees (German Research Foundation - DFG)

Stefan Winkler-Nees works as program officer at the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation). The DFG is the central, self-governing research funding organization in Germany.[81] As part of the unit "Scientific Library Services and Information Systems" (LIS), Winkler-Nees is responsible for the DFGs strategic activities in the context of permanent access to research data. His background is in marine geosciences and climate research. Having completed a number of post-doc projects in Europe and overseas he was employed by a software company. After six years he returned to science with a position at the DFG.

In 2006, the DFG Committee on Scientific Library Services and Information Systems released a position paper on future activities in digital information. One of the main topics of this paper within the field of "information management" is the "further development of the structures of the provision of primary research data". Based on this paper, the DFG started to develop a strategy for means and measures to improve the management of research data in the future.

Challenges and opportunities were discussed in a series of discipline-specific round-table discussions. "The feedback from representatives of the different disciplines varies to a large degree. Interesting for me was the fact, that not only the science, technology and medicine disciplines (STM) see the relevance of data sharing", says Winkler-Nees. A main result of the communication with the disciplines was that the funding of substantial infrastructures must grow in conjunction with the implementation of policy activities. "Without an infrastructure that assists scientists in a convenient and efficient way managing their data, no culture of data sharing will evolve", says Winkler-Nees. The consultation and the communication with scientists and infrastructure representatives are described as of prime importance. "As a research funding organization we have to promote the dialog between the actors especially in this regard." Winkler-Nees points out that the permanent access to scientific data is a challenge across all disciplines without exception. The DFG started also to improve the dialog with other funders. Hence the DFG cooperates with the partners in the European initiative "Knowledge Exchange" on joint strategies in the field of research data management.[82] Winkler-Nees explained: "In disciplines working internationally, we have to develop common strategies on data sharing." The DFG sub-committee on "Information Management" published in 2009 "Recommendations for the secure storage and availability of digital primary research data". In this context, a recommendation is given: "Every scientist shall make his primary research data freely available beyond his institution whenever possible." This paradigm, with respect to the disciplinary particularities, is guiding the activities of the DFG.

Since 2008, the DFG has also been part of a national initiative called "Digital Information" of the "Alliance of German Science Organizations".[83] With this leading initiative, German science organizations have agreed to coordinate their activities and to ensure the long-term availability of digital information and its integration into virtual research environments. The partner organizations agreed to align their funding programs in the area of research data and, when necessary and appropriate, to merge or harmonize them. "The promotion of sharing data is such an important task that

[81] http://www.dfg.de

[82] http://www.knowledge-exchange.info

[83] http://www.allianz-initiative.de

cooperation and communication between all stakeholders plays a major role", says Winkler-Nees. The Alliance of German Science Organizations released in 2010 "Principles for the Handling of Research Data". In this document, the partner organizations declare their support for "open access to data from publicly funded research." "However this kind of policy paper needs to be followed-up by the development of an appropriate infrastructure to support the implementation of a culture of data sharing", says Winkler-Nees. For this, the working cooperation between scientists, librarians and IT- and information management specialists is essential. All research data management services must in principle be adapted to scientific requirements, but additionally need to be operated with the necessary information management expertise.

The approach of linking scientists with infrastructure professionals is also implemented in the DFG strategy: all activities in the field of research data management were designed in a communication process between the LIS unit and the different disciplinary units. "The close cooperation between LIS and different disciplinary units helps us to promote dialogue on the opportunities and challenges of data sharing", says Winkler-Nees. This communication also enhances the positive awareness effect on this topic. The internal discussions increased attention to the significance of data sharing.

In 2010 the DFG established this topic in their "Guidelines for Proposals". Hence applicants must indicate what measures they plan to secure the collected data as well as to facilitate re-use. This requirement is intended to encourage applicants to share their data. "Due to the diversity in disciplines, we have decided to take small, but effective steps. Some disciplines, such as the geosciences, are already demanding further steps, like mandatory data management plans. But we must take into account the needs of all disciplines." Winkler-Nees emphasises that data bureaucracy must be avoided. "We shouldn't forget that there are disciplines where data sharing is not possible or difficult due to legal aspects."

In order to encourage the development of infrastructures, the DFG, in 2010, released a call for proposals on "Information infrastructures for research data". This call generated an enormous interest in a large variety of disciplines. "The large number of applications shows the importance of that topic." All proposals were also reviewed by scientists, to ensure their relevance to the related discipline. "With a funding of 9.9 million Euros for 27 infrastructure projects, we hope to facilitate the sharing of research data and to set a foundation stone for the future scientific information infrastructure", says Winkler-Nees.

In future, the DFG will promote the sharing of research data in international frameworks. A joint statement of a group of major international funders of public health research may serve as an example. True to the motto "sharing research data to improve public health", 17 signatories, including major public funding agencies, charitable foundations and international organizations, have committed to cooperation to increase the availability of data emerging from funded research in January 2011. "With a view on internationally working scientists and in particular considering the European level, we have to set the course for a culture of data sharing. We have to adjust our activities with other significant stakeholders", says Winkler-Nees. "The financial and legal frameworks for national activities have to be developed together with all relevant partners and organizations. While funding research data infrastructures such as repositories we have to develop sustainable funding solutions. This is a challenge."

3. CONCLUSION

The collected interview stories from relevant stakeholders in the previous chapter provide evidence relating to data sharing in terms of success stories and lessons to be learned. They form the information baseline for *status quo* of data sharing and re-use today.

This chapter will summarise the implications of these stories in the form of elaborated hypotheses that will be used to analyse drivers and barriers to data sharing, to be explored in the subsequent phases of the project.

Based on the ODE-objectives, relevant hypotheses were drawn from all interview stories. Firstly, for each interview, up to five hypotheses were extracted directly by the interviewer. Secondly, these raw hypotheses were placed into the ODE-evidence base, along with the corresponding interview story, to be commented on or added to by all partners. These hypotheses were circulated and discussed by all involved project partners. Hence, these hypotheses rely on key points that became evident and were stressed in the interviews, in relation to the questions addressed by the ODE-project. Through this process all project partners could participate in extracting or altering relevant hypotheses from the corresponding interview story in order to carve out the potential for innovation and impact of attitudes, policies, e-Infrastructures initiatives, and drivers and barriers for data sharing. In a subsequent step these hypotheses were summarized and generalized into a broader and more understandable format allowing all involved partners to categorize the individual hypotheses via a poll.

3.1 Different perspectives of data sharing

Fourteen different categories are elaborated, addressing all the raw hypotheses, comprising significant perspectives on drivers and barriers for data sharing through a European eco-system of data repositories.

1. Education (training, especially within disciplines)
2. Legislation (national and EU-wide basic legal requirements)
3. Financing (sustainable funding of infrastructures)
4. Culture and attitude (incentives & exclusion, education)
5. Quality (more R&D, manifold aspects of quality)
6. Policies (feasibility & concreting, especially in terms of different disciplines)
7. Cooperation (enhancement of international dialogue & networking, multi-disciplinary, extension to USA and Australia)
8. Infrastructure (promotion/enforcement of stakeholders' cooperation, availability of infrastructure, scaling problems, deluge, technical & practical support)
9. Publishing & visibility (coupling of textual publication and data, development of new forms of publications)
10. Data flow improvements (data management & data publishing & data re-usability)
11. Disciplines (same challenges, but different solutions; practical bottom-up solutions)
12. Accreditation & certification (trust, reputation, quality assurance, peer-review, sustainability)

13. New career paths for "data scientists"
14. Efficiency (no duplication of work, financial cost)

We are aware that not all perspectives have been captured through this procedure as the hypotheses are based on 21 selective interviews of stakeholders, who have been chosen out of the ODE-partners expertise and cooperation's sphere. In particular, the views of libraries and publishers, as well as preservation aspects, are not represented thoroughly in this approach.

3.2 Hypotheses of data sharing from different perspectives

This summary of all hypotheses was presented and discussed with all partners in a follow-up ODE face-to-face meeting in 11 July 2011. The final elaborated hypotheses presented in this baseline report help to identify and suggest engagement in conference publicity as well as to re-factor the results into a set of questions and statements, which can be used as the straw-man conceptual outline in the next project phase of WP5, "Drivers and barriers: questions and answers".

1. Education
    - 1.1 Successful data sharing needs skills
        - 1.1.1 Specific personnel need training for specific data preparation (harmonizing, standardizing) and data quality checks
        - 1.1.2 Data users need training and consultation on data finding and data usage by specific personnel
        - 1.1.3 Specialized data centres should train scientists in proper data management
        - 1.1.4 To avoid misuse and lack of acknowledgement of very special data, access should be restricted to skilled persons trained by the data creator
    - 1.2 Behaviour
        - 1.2.1 It must be practical, pragmatic and easy to share data
        - 1.2.2 Despite existing infrastructures, researchers' hesitation to share data is one of the most prominent barriers for data sharing
        - 1.2.3 Change of attitude: premature data releases should not be enforced, but the mere possibility of data misinterpretation is no reason for not sharing data
    - 1.3 Incentives
        - 1.3.1 Proper and qualitative data management and data sharing will enhance scientific reputation of institutions and scientists
        - 1.3.2 Proper citing and acknowledgement of shared/re-used data will enhance reputation of scientists as well as diminish hesitation in data sharing

1.3.3 Data sharing and data reusability will enhance interdisciplinary research

1.4 Appreciation and Recognition

1.4.1 Tribute has to be given to the effort of scientists who manage and prepare data for storage and sharing

1.4.2 Establishment of new job profiles and careers: data scientists with official acknowledgement for their data management work and without the commitment to own research profiles and publishing

2. Legislation

2.1 Amongst researchers there is a certain wariness of, but not outright resistance towards, the Freedom of Information Act (FoIA). This is reflected by the interplay of hesitation due to either possible data misuse or loss of exclusive data exploitation and the effective added value by sharing and reusing data

2.2 Discrepancies in local, regional, national and international legislation is a severe barrier for effective data sharing and data utilization

2.3 Data restrictions via federal, national and institutional confinements are often caused by administrational barriers due to strategic interests

2.4 Legal issues of restricted access to data on local and national levels can only be straitened by international legislation

2.5 Legislation should also take into account terms of use, licensing and reconditioning of older but nevertheless useful data

2.6 Compatibility and standards of data and international infrastructures should be supported by proper international legislation and directives

2.7 Data policy and data sharing are not legally clear and sound on a national as well as international level thus integration between science and politics is needed

3. Funding

3.1 Data archiving with public assignments (the backbone of data sharing) cannot work under pure economic perceptions. Cooperation with publishers and additional e-infrastructures funding is necessary.

3.2 The exponentially increasing flood and complexity of scientific data needs additional financial acknowledgments to develop, maintain and guarantee future integrative e-infrastructures

3.3 Data intensive science in the context of ever growing international cooperative research networks needs extra funding for developing integrative infrastructures

3.4 If no data preservation is taking place huge and costly effort is needed to recover "old" data.

3.5 Establishing a culture of data sharing and data reusing needs extra funding

3.5.1 New business models need to be tested

3.5.2 Preparing data for reuse and publication requires additional personal and infrastructural investments beyond pure research funding

3.5.3 Often the full potential of data cannot be exploited during a projects lifetime hence the continuity of research data management needs to be guaranteed

3.6 Research funding agencies should enforce funding requirements for publishing the data behind the scientific publication

3.7 Data sharing should be prominently promoted through a common system where research projects and researchers receive funding credits for sharing data

3.8 Not sharing data should be considered generally as an intellectual and financial loss. Basic essential needs for sharing data or drastic damage caused by not sharing data has to be the driver of gathering and sharing data worldwide (e.g. weather forecast)

3.9 The financial as well as legal framework for national activities needs to be built up jointly with all relevant stakeholders. Hence the financing of international infrastructures such as research data repositories is a big challenge

4. Culture/Attitude

4.1 Disciplines

4.1.1 Data sharing is subject to strong disciplinary aspects, e.g. in SSH data often needs special preparation before publication due to ethical constraints

4.1.2 Data sharing and the integration of research data with scholarly communication varies a lot across disciplines and needs individual approaches

4.1.3 While some communities are pioneers in open access to scientific articles the sharing of data still brings many technical challenges concerning data quality, data standardisation and data reusability

4.2 Behaviour

4.2.1 Data sharing has to be practical, pragmatic and overcome listlessness, inconsideration, egoism and untidiness

4.2.2 Since data storing and sharing is not a naturally understood commitment for scientists we need fundamental changes in the incentives for data sharing. If not scientists will perpetuate bad habits

4.2.3 Premature data releases should not be enforced, but the mere possibility of data misinterpretation is no reason for not sharing data

4.2.4 Research data management needs to be considered as a continuous effort throughout the full life cycle of data up to sustainable long-term archiving that goes far beyond mere project funding

4.2.5 Research data should be considered as part of the world's cultural heritage. The long term preservation of this data is necessary to grant researchers access to this data now and in the future

4.3 Education

4.3.1 Specialized data centres should train scientists in proper data management

4.3.2 Scientific supervisors should undertake to educate and urge young academics to data management and data sharing so that it will become self-evident

4.3.3 Data management as a logical prerequisite to data sharing should become a fixed integral part of academic education

4.4 Hesitation

4.4.1 Despite existing infrastructures hesitation by researchers is one of the most prominent barriers for data sharing since it could be a laborious task that is not yet compensated in research evaluation

4.4.2 The model of enhanced publication could be considered a step to establish data sharing in the different disciplines to overcome hesitation

4.4.3 To improve one's own scholarly record research data should be considered as an independent digital object that is shared and cited independently of any other research object

4.4.4 Data acquisition by scientists should be driven from small to large scale implying that data from local studies are the base of global knowledge making cooperative data sharing self-evident in research

4.5 Incentives

4.5.1 The life cycle of data is neither part of the scientific evaluation procedure nor is the act of data sharing a beneficial part of a scientist's personal carrier.

4.5.2 Explicit additional funding and incentives are necessary to compensate for time consuming data delivery in dedicated formats and defined time horizons

4.5.3 Further reasons to publish data could be to avoid mistakes as well as suspicion of fraud. Sometimes raw data need to be available to enable reproduction of derived data and results

4.5.4 The added value of sharing data to enhance interdisciplinary research and global understanding is not yet fully recognized

4.5.5 Sharing of research data could be linked with higher citation rates

5. Quality

5.1 Data quality checks by scientists must be assured. In cooperation with scientists, high standards must be defined as the basis of data sharing. This guides the way to a reliable, publishable, and citable data set

5.2 Funding is crucial for the delivery and sharing of high quality data in a dedicated format and defined time horizon

5.3 New incentives should be created giving reflecting the term “data are our gold” and that data needs to be stored and shared in an appropriate way to exploit and sustain its quality

5.4 Mechanism of quality assurance (automated as well as manual) should be improved as an integral part of data management and a basic prerequisite for data sharing to enhance trust and avoid misinterpretation of data

5.5 High data quality defines the reputation of an institute/scientific community. Proper documentation of data and understanding of metadata must be a key concern for scientists

5.6 Data quality should be assured by embedding the data management process in the original scientific project. Hence, only direct cooperation with the scientists ensures quality and top scientific standards

5.7 Specific personnel and new job profiles emerge for data preparation and quality assurance of data

6. Policies

6.1 There are strong disciplinary differences in data sharing, especially regarding data privacy in the context of confidential, personal or ethical aspects

6.2 Data policy and data sharing are not legally clear and sound on a national as well as international level thus integration between science and politics is needed

6.3 Communication between science and politics should be improved by the installation of advisory expert groups and interdisciplinary commissions. The end users must be engaged

6.4 Open Access to data? There are different constraints for sharing data publicly. Open whenever - possibly closed when needed.

6.5 The work behind data collection should be recognized properly. Assurance should be given to accredit the intellectual personal work that makes data originally fit for scientific usage

6.6 Amongst researchers there is a certain wariness of, but not outright resistance, towards the Freedom of Information Act (FoIA). This is reflected by the interplay of hesitation due to either possible data misuse or loss of exclusive data exploitation and the effective added value by sharing and reusing data

6.7 Should data publishing be enforced to ensure good scientific practice as well as enabling data review (control) and to avoid suspicion of fraud?

6.8 A collective will for improving the availability and utilization of data resources is needed as well as a data policy to realize this

6.9 To avoid misuse and lack of acknowledgement of very special data, access should be restricted to skilled persons trained by the data specialists.

6.10 Data management plans are necessary

6.10.1 to firmly allocate resources for data availability

6.10.2 to explicitly determine retention times and policies (considering raw/primary data and potentially including: data deletion!)

6.10.3 to make data auditable if necessary

7. Cooperation

7.1 National as well as international data policy and data sharing practice must be clear and sound. This depends on good interaction between science and politics.

7.2 Communication between science and politics should be improved by the installation of advisory expert groups and interdisciplinary commissions. The end users must be engaged.

7.3 The financial as well as legal framework for national activities needs to be built up jointly with all relevant stakeholders. Hence, the financing of international infrastructures such as research data repositories is still a big challenge

7.4 Seamless integration of data repositories with research is needed. Strong collaborations between research communities and journals must be advised

7.5 Continuous interdisciplinary training of the diverse scientific community via guiding and implementing optimal use of data generation and data preservation must be envisioned

7.6 Cooperation creates synergies in and between disciplines. Interdisciplinary efforts lead to data standardisation and methodologies for data preservation and sharing

7.7 Data acquisition by scientists should be driven from small to large scale implying that data from local studies are the base of global knowledge making cooperative data sharing self-evident in research

7.8 A cross-cutting international forum for research data management to create strong cooperation should be implemented

8. Infrastructure

8.1 An international research community needs an international data infrastructure and international support

8.1.1 Without an infrastructure that assists scientists in managing their data no culture of data sharing will be created

8.1.2 As stated by the interdisciplinary "International Polar Year" community: "After decades of reports with data in their titles the community found inadequate services almost no international support and few solutions."

8.1.3 Specific and well recognized needs drive transnational, disciplinary infrastructures for data sharing e.g. meteorological data sharing benefits society

8.1.4 Data acquisition should be driven from small to large scale implying that data from local studies are the base of global knowledge making cooperative data sharing (implies cooperative infrastructures) self-evident in research

8.1.5 A misbalance between scientific expertise and political decision making results in lack of cross-border information exchange and data sharing infrastructures. Communication between science and politics should be improved in order to engage the end user

8.1.6. A cross-cutting international forum for research data management in e-Science and e-Infrastructure is needed for the creation of general-purpose European e-Infrastructure.

8.1.7 International coordination is needed to create a common strategy to obtain synergy or savings between states or projects

8.2 Data archiving with public assignments (the back bone of data sharing) cannot work under pure economic perceptions. Cooperation with publishers and additional e-infrastructures funding is necessary

8.3 Disciplines

8.3.1 In some disciplines the amounts of data grow faster than financial and technical means to share it causing scaling problems and data deluge

8.3.2 There are disciplines with high (theoretical) willingness to share data, but no solution is available to complex challenges

8.3.3 A limited number of specialized and certified data centers should be assigned by scientific disciplines to enhance discoverability. These data centres must provide guidelines on what and where to deliver

8.3.4 Some disciplines view data-sharing positively but lack (crucial) infrastructure; some discipline hesitate to share data despite existing infrastructure

8.3.5 Disciplinary efforts are needed to join forces in regards to data standardisation and methodologies

8.3.6 In many cases, data are preserved but difficult to discover or too difficult to extract/compile; education must be linked to technical infrastructure

8.4 Ensuring Quality, Practicality and Sustainability

8.4.1 For the establishment of a data sharing and data reuse culture sustainable infrastructures and services are needed

8.4.2 The processing of data to high-quality information products is very complex. Data centres need to support the preparation of information products for particular needs

8.4.3 Data archiving and data reuse must be organized by certified and scientifically close-by institutions ensuring high standards and quality, which are indispensable for trust and the reputation of scientific work

8.4.4 The end users must be engaged in the planning of the infrastructures

8.4.5 Scientific data are scattered widely. Better research would be possible with better visibility and discoverability of data. Hence, unification and simplification of data access is necessary.

9. Publishing/Visibility

9.1 Data sharing and data publishing needs to be recognized in the evaluation of scientist's achievements while recognizing

9.1.1 The work in preparing data

9.1.2 The intellectual and scientific achievements

9.2 Datasets need to be citable entities. Hence citation conventions need to be developed

9.3 As a principle, data need to be openly available, and only restricted where necessary

9.4 The link between articles and data will play a major role

9.4.1 Providing datasets with articles increases citation rates

9.4.2 Cooperation between data centres and publishers will enhance recognition of data

9.4.3 If data is more visible and discoverable via publishing, research will be more efficient

10. Data Flow Improvements

    10.1 Research data management needs to be considered as a continuous effort before, during and after projects

    10.2 Data management applies to raw, primary and derived data

    10.3 Data repositories need to collaborate closely with research communities and journals. Seamless integration of data repositories with research is needed

    10.4 Data management is performed by

        10.4.1 Data scientists within research groups

        10.4.2 Professionals at certified data centres

    10.5 Researchers need guidance on appropriate data management plans

    10.6 Data management will be based on disciplinary standards with considerable interdisciplinary interoperability

11. Disciplines

    11.1 Disciplines need to define how data management is integrated into good scientific practise and publishing habits. Likewise, the disciplines need to define what is worth archiving and what is worth publishing/sharing

    11.2 Specific and well recognized needs drive transnational as well as disciplinary infrastructures for data sharing e.g. meteorological data sharing benefits society, shared astronomy data benefits science

    11.3 There are "overwhelming" financial & technical reasons to build big data infrastructures along gene sequencers

    11.4 Some disciplines view data-sharing positively but lack (crucial) infrastructure. On the other hand some disciplines hesitate to share data despite existing infrastructure.

    11.5 There are huge differences between disciplines in what constitutes the major challenges, barriers and drivers for data sharing: e.g., volume of data, ethical constraints, habits, (perceived) value to science

12. Accreditation and Certification (trust, quality, sustainability)

    12.1 There is still no common agreement upon "good practice" in sharing and to re-using data in a responsible and fair way

    12.2 Apart from the use of 'bad data', misuse of good data can damage the reputation of its originator or its originating institution.

12.3 Education and new practices, such as coupling the publication of articles and data, may lead to improved quality and, thus, reputation

12.4 Certified repositories will guarantee future data re-use

13. New career paths for "data scientists"

13.1 Data need preparation and quality assurance by professionals to become re-usable. This needs to be factored into research budgets

13.2 Depending on the discipline, professional preparation of data may be beyond the competence or beneath the recognized scope of the role of a researcher

13.2.1 "Data scientists" must be "embedded" in research groups

13.2.2 Support to researchers by data centres is most helpful

14. Efficiency

14.1 Well preserved, more discoverable and accessible data are enhancing research and perception

14.2 Global and interdisciplinary "compilations" and coverage of shared data enable better and more research

14.3 Better stewardship of data is less costly than re-creating data or recovering data from unprofessional storage

4 OUTLOOK

Through interviews with some relevant stakeholders, views and opinions on challenges and opportunities for data exchange have been collected in 21 stories. Relevant hypotheses were drawn from all the interview stories and elaborated into 14 different categories or perspectives.

On the basis of this report, and an upcoming report on the handling of research data in scholarly communication, the ODE partners will enter the next project phase. In this phase drivers and barriers for data sharing, reuse and preservation will be refined and validated through consultation with a broad variety of experts in order to reach a better understanding of the challenges and opportunities in this complex field.

In particular, the hypotheses will be presented to, and discussed with, some of the interviewees and other stakeholders during the APA 2011 conference in London.

WP5 will apply a methodology which will subject the hypotheses generated by WP3 to a representative test.

5. ANNEX

5.1 GLOSSARY

ANDS	Australian National Data Service
APA	Alliance for Permanent Access
AWI	Alfred Wegener Institute for Polar and Marine Research
BSH	Federal Maritime and Hydrographic Agency
CERN	European Organization for Nuclear Research
CSC	Finnish IT Centre For Science
DFD	German Remote Sensing Data Centre
DFG	German Research Foundation
DIMS	Data and Information Management System
DLR	German Aerospace Centre
DNB	Deutsche Nationalbibliothek
DOI	Digital Object Identifier
DoW	Description of Work
DPHEP	Study Group for Data Preservation and Long Term Analysis in High Energy Physics
D-SDA	German Satellite Data Archive
EB	Evidence Base
EBI	European Bioinformatics Institute
EC	European Commission
e-IRG	e-Infrastructure Reflection Group
ELIXIR	European life science infrastructure for biological information
ERA	European Research Area
ESFRI	European Strategy Forum on Research Infrastructures
EU	European Union
FP7	Seventh Framework Programme
GSI	Helmholtz Centre for Heavy Ion Research
GTS	GTS Global Telecommunication System
HA	Helmholtz Association of German Research Centres
HELCOM	Helsinki Commission
ICES	International Council for the Exploration of the Sea
ICSU	International Council for Science
IMF	Remote Sensing Technology Institute
IMO	International Meteorological Organization
IOC	Intergovernmental Oceanographic Commission
IPCC	Intergovernmental Panel on Climate Change
IPO	International Programme Office

IPY	International Polar Year
JISC	Joint Information Systems Committee
LHC	Large Hadron Collider
LIBER	The Stichting LIBER Foundation
LIS	Scientific Library Services and Information System
MODEG	Marine Observation and Data Expert Group
NESCen	National Evolutionary Synthesis Centre
NSIDC	National Snow and Ice Data Centre
ODE	Opportunities for Data Exchange
OECD	Organisation for Economic Co-operation and Development
OKF	Open Knowledge Foundation
PANGAEA	Publishing Network for Geoscientific and Environmental Data
Parse. Insight	Permanent Access to the Records of Science in Europe
R&D	Research and development
REKLIM	Helmholtz Climate Initiative Regional Climate Change
SCOAP3	Sponsoring Consortium for Open Access Publishing in Particle Physics
SOAP	Study of Open Access Publishing
STFC	Science and Technology Facilities Council
STM	The International Association of STM Publishers
UNEP	United Nations Environment Program
UNESCO	United Nations Educational, Scientific and Cultural Organization
VO	Virtual Observatories
WCP	World Climate Program
WCRP	World Climate Research Program
WDC	World Data Centre
WDC-MARE	World Data Centre for Marine Environmental Sciences
WDC-RSAT	World Data Centre for Remote Sensing of the Atmospheric
WDS	World Data System
WLCG	Worldwide LHC Computing Grid
WMO	World Meteorological Organization
WOCE	World Ocean Circulation Experiment
WP	Work Package
ZKI	Centre for Satellite Based Crisis Information