Studies on Subject-Specific Requirements for Open Access Infrastructure

Edited by Christian Meier zu Verl and Wolfram Horstmann

First published 2011

Universitätsbibliothek Bielefeld
Universitätsstraße 25
D-33615 Bielefeld

A catalogue record for this book is available from the
Deutsche Nationalbibliothek

ISBN: 978-3-943363-00-5
DOI: 10.2390/PUB-2011-1

Typeset: Da-TeX Gerd Blumenstein, Leipzig (Germany)
Proofreading: Katharine Timberlake, Portknockie Banffshire (UK)

Funded by the European Commission

SEVENTH FRAMEWORK PROGRAMME

Open Access Infrastructure for Research in Europe
OpenAIRE

Contents

Executive Summary

This study addresses subject-specific requirements for research infrastructure with a focus on the influences of Open Access (OA). OA is treated in a broad sense covering open access to literature, open data and open science. Considering the wide variety of aspects to be analysed and the early stages of developing a general account of OA infrastructure, the study took a case-based approach and deliberately did not attempt to provide a representative account of research. In the pragmatic approach taken, six partners (institutions and organisations) were chosen to provide their subjective view on OA infrastructure. These partners are considered as exemplars of research and infrastructure institutions in a given subject area.

When comparing the chapters, the most obvious observation can be summarised in one word: “diversity”. On the first view, this subject-specific diversity may appear to be the natural enemy of infrastructure, since infrastructure is about commonalities in terms of global standards, joint facilities and shared resources rather than about differences between diverse subject-specific requirements. Simultaneously, it is obvious that research must be extremely diverse in terms of thematic and methodological specialisation in order to tackle the ever more specific research challenges of the world. **Thus, any roadmap for OA infrastructure must address this natural tension field between diversity and infrastructure.** This study chose the approach of addressing this natural tension field directly by first providing an account of diversity by the subject-specific chapters. Then this diversity is reflected on specific aspects of OA infrastructure such as OA to literature and OA to data. It is not expected that the study will provide a complete picture and detailed plan for the next decades; rather it is expected that the reader will gather impressions of diversity and develop a (maybe sometimes tacit) understanding of how diversity can be managed within research infrastructure development in a way that leaves research with sufficient degrees of freedom for self-organised developments while supporting the emergence of synergies between those self-organised developments through shared resources that apply principles of openness.

OA and infrastructure are two completely different phenomena: OA is a mode of communication while infrastructure refers to facilities. However, the benefits of both can be characterised by referring to the same aspects of research: i.e. cost considerations, the enabling of research otherwise not possible, transparency and comparability as well as synergies. The reason for this mutual relation between infrastructure and OA is most obviously that **both infrastructure and OA imply a notion of sharing**.

The comparative analysis elucidates characteristics of the research lifecycles in general and its constituent aspects of literature management and data management. Research lifecycles show common steps: (i) data collection, (ii) processing, (iii) enriching, (iv) archiving, and (v) re-using. However, the variance in the descriptions appears stronger than these rather abstract commonalities. Literature management shows strong commonalities in tooling but strong differences is publishing practices. Data management shows a large variety in both tooling and data management practice. The comparative analysis shows that OA to literature is a growing or established practice in the subject areas but not yet fully developed. OA to data is considered an important future activity. This indicates that is OA infrastructure can be built immediately and in a rather generic sense for literature and has to be built with more patience and consideration for subject-specific requirements in the future.

As mentioned above, infrastructure is an opponent to diversity since infrastructure is not only an essential prerequisite but also a collection of rigid conditions or constraints: it is an inherent property and explicit objective of infrastructure to make research uniform. Openness, however, is a way to maximise the permeability of research resources (literature and data) within research infrastructure so that the collaborative, interdisciplinary and international research activities that are needed to tackle a given next challenge can emerge. **The key challenge is developing research infrastructure that operates in an open mode and, thereby, supports the diversity of research practices through increased information flows between subjects.**

Measures to support infrastructure developments (e.g. funding programmes) should therefore take into account the following observations, which can be interpreted on the basis of the subject-specific requirement descriptions throughout this volume.

i. Digital literature and data resources are an essential precondition of research. The provision of digital literature and data resources through infrastructural services are perceived as a matter of course (or implicitness) and are not questioned unless they are obviously missing. Thus, “knowledge infrastructure”, as the entirety of resources and processes related to digital literature and data resources used in research, is not conceived as an explicit facility but rather as an invisible capacity.

ii. OA is described as a *modus operandi* for working with digital literature and data resources rather than as an end in itself or an ethical principle.

iii. OA to literature and OA to data refer to very different parts of the research process. While literature shows universally generic characteristics, data is much more related to subject-specific methodologies and facilities. Even though the benefits are the same for literature and data, the obstacles vary broadly and require that OA to literature and OA to data are differentiated in policy and infrastructure development.

iv. Due to the universally generic role of text-based resources in research, OA to literature can be regarded as a general prerequisite for efficient and effective as well as innovative research and should be mandated uniformly over all subject areas – even if the specific implementation of OA to literature is left at the discretion of the subject areas (e.g. through subject-specific repositories).

v. OA to data has (yet) to be reflected in a fully subject-specific way in policy and research infrastructure development. The emerging practice of mandatory project-specific *data management plans* that address the question of OA to data could be sharpened by asking the question: “Are data open and if not, why not?” OA in data management plans could be supported by providing a generic Open Data policy with subject-specific *addendi* to such a generic policy. Such a subject-specific addendum to a generic Open Data policy may well be mandatory in a given subject area.

vi. The difference between OA to literature and OA to data may be transient as more and more systematic connections between literature and data can be observed. Explorations towards infrastructural linkage between literature and data (e.g. enhanced publications) should be intensified.

vii. The provision of research infrastructure services by institutions and organisations is requirement driven and situated in a given research context – even *within* a smaller subject area – and supports collaboration among researchers from various disciplines. The development scheme in practice tends to be incremental and evolutionary and based on prototypes and working solutions rather than applying theoretic frameworks and capacious facilities.

viii. The layer cake model of research infrastructure – from the generic information and communications technology (ICT) infrastructure to the next level of information/data/knowledge infrastructure to the level of subject-specific applications – does not reflect the complex organisation of research infrastructure. The distinction between “horizontal” developments based on generic research processes and ICT standards and “vertical” developments based on subject-specific research questions is helpful since it breaks up the layer cake model and suggests a hierarchical matrix model. However, a network model of research infrastructure consisting of a multitude of subject-specific nodes that apply common local design principles (e.g. metadata standards, exchange protocols) in order to communicate with one another and share resources amongst other nodes reflects best the descriptions of research infrastructure in this study and is assumed to be the most promising approach for future research infrastructure developments.

As a summary, future research infrastructure developments should consider the following principles in order to reflect the diversity of research as the key challenge.

i. Support subject-specific developments that are research driven, incremental and evolutionary in order to match and adapt to the established situated practices.

ii. In a separate strand, support the development of generic infrastructural services and standards applicable in local subject-specific nodes. Services and standards should obviously be maintained by institutions and organisations with long-term commitment.

iii. Provide systematic cross-talk between the subject-specific and generic developments by:

   a. providing research and development programmes that explicitly address the question of how to link subject-specific and generic developments. Examples for activities are science and technology studies, networking events or focused infrastructure projects,

   b. installing advisory boards or oversight groups for projects and funding programmes that have representations of both subject specialists and infrastructure specialists, and

   c. enforcing mutual participation of subject specialists and infrastructure specialists in assessments and reviews.

iv. Apply OA as a *modus operandi* in all activities – mandatory for literature and recommended for data, with an appropriate consideration of subject-specific exceptions.

A Introduction

Christian Meier zu Verl and Wolfram Horstmann

As the internet continuously catalyses the development of novel methods to perform research, elementary questions about future forms of research communication are being posed. One of these questions is how *openness* of research can be optimally exploited through the internet, in order to tackle research problems previously impossible to analyse and also in order to increase time effectiveness and cost efficiency. Hence, research is transforming constantly by capabilities of new technologies: “Collaboration is growing for a variety of reasons. Developments in communication technologies and cheaper travel make it easier than ever before for researchers to work together, the scale of research questions, and the equipment required to study demands that researcher are mobile and responsive” (Royal Society, 2011). Openness in the internet shall ease the collaboration of researchers around the globe and the sharing of resources. This is often referred to as Open Access (OA).

Originally, OA activities were referring predominantly to text-based publications. More recently, topics such as Open Data or Open Science were entering the discussion. In order to adopt a neutral stance in this study, it should be noted that OA is not pre-supposed as an imperative requirement for research. Specific aspects of research may require access restrictions, among them quality considerations, competition, privacy and security. The question posed in this study is rather, in which parts of research is OA beneficial for research itself and in which parts could OA even being regarded as a restriction for the function of research?

OA to literature is a universal issue. Not only the distribution of knowledge is faster and easier but also the development of reputations and the system of publication (e.g. editors, publishers, libraries) is affected by OA. OA to literature varies between different research disciplines. For example, OA is accepted in parts of the natural sciences, while OA in the humanities or social sciences is not equivalently established (Harley, Krzys Acord, Earl-Novell, Lawrence and King, 2010; Theodorou, 2010; Taubert and Weingart, 2010).

The shift in the OA activities from text to data to all research resources has deep implications: while there is at least some kind of common sense across research disciplines of "text", the understanding of "data" massively varies across disciplines. Obvious reasons for this variety can be seen in the dependency of data on the context, in which they are appearing. While text publications are often an end-product of research, data can appear anywhere in the research lifecycle. While texts require rather simple means to be communicated and utilised, such as print or electronic display, data often depend critically on a specific instrument, software or expert knowledge, which is only to be found in one specific discipline. As a consequence, the scope of benefits and restrictions of OA to data depends on subject-specific forms of research (RIN and NESTA, 2010). In other words: "[It depends all on] who shares what, with whom, and at what stage of research" (Borgman, 2010).

Extending the scope of the OA discussion from text, to data, to all research resources, also inevitably introduces the question: "Which subject-specific requirements on research infrastructure exist?" Answering this question may lead to the conclusion that a wide-scoped implementation of OA principles is only possible by a subject-specific approach. It may also lead to the conclusion that a strong generic infrastructure is the appropriate perspective. However, these big and essential questions for research infrastructure development in the next decades must be addressed. In order to tackle these questions in ways that will be accepted by subject communities, this study analyses subject-specific requirements on research infrastructure, especially with respect to OA.

Definitions

We refer to the scope of OA in terms of the Berlin Declaration:

> *Establishing open access as a worthwhile procedure ideally requires the active commitment of each and every individual producer of scientific knowledge and holder of cultural heritage. Open access contributions include original scientific research results, raw data and meta data, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material.*

We refer to the definition of OA in slightly modified terms of the Budapest Declaration:

> *By open access, we mean its immediate, free availability on the public internet, permitting any users to read, download, copy, distribute, [export], search or link to the [materials], crawl them for indexing, pass them as data to software or use them for any other lawful purpose.*

It should be mentioned that OA is not seen in this study as an end in itself. It is acknowledged that parts of research infrastructure need careful consideration of privacy and security. Rather, the idea is to identify those parts of research infrastructure to which is OA beneficial. In order to analyse the implications of widening the OA discussion from text to data, we will focus on implications for research infrastructure.

By research infrastructure, we mean the entirety of production and services, which includes instruments like large sensors, satellites, laboratories and many more facilities, such as digital services and virtual research environments. The research process within that refers to all facilitating processes: the researcher and his or her behaviour is not part of the infrastructure.

There are several approaches that focus on other parts of research infrastructure but that are not covered here in this description (for example, the human factor of research infrastructure; Lee, Dourish and Mark, 2006).

The question how OA infrastructure can be defined – as opposed to the more generic concept of research infrastructure – shall deliberately be left open in this introduction, not to pre-suppose subject-specific definitions of each case study.

1 Context

What makes this study unique? While other studies point out issues like communication, archival publication or data sharing, curation and re-use, our study addresses the interplay between subject specificity, OA and infrastructure. The combination of case studies provided by highly specific and renowned institutes and authored by subject experts shall shed light on the diversity of research cultures. Five different research disciplines will be thoroughly described in order to show principles of existing research infrastructures and draw conclusions for a roadmap.[1]

This report is related mainly to three current studies.

- Harley, Krzys Acord, Earl-Novell, Lawrence and King (2010) "Assessing the future landscape of scholarly communication": This report focuses on researchers' perspectives on different aspects of (i) tenure and promotion, (ii) publication practices, (iii) sharing, and (iv) public engagement. Researchers mostly count their record of publications to develop their tenure. Therefore, the management of own publication matters

[1] The context of this study is the European project 'Open Access Infrastructure for Research in Europe' (OpenAIRE), funded by the European Commission (EC) under the Seventh Research Framework Programme (FP7). OpenAIRE develops OA infrastructure to support and implement the OA policy of the European Commission. Our study within OpenAIRE evaluates subject-specific requirements on future OA infrastructures. It is produced to provide an understanding of research communication in different disciplines in order to elucidate necessary steps to develop new technical systems for OA infrastructures.

much more than every data practice. The practice of publication is a key driver within research communication. Each discipline weighs some factors of publication in different ways, such as speed of publication, target audience, peer review, new publication models, to name but a few. Data sharing is divided into four dimensions: (i) personal communication, (ii) informal exchange, (iii) the wider circle of colleagues, and (iv) the public. Along these dimensions, researchers organise their data-sharing practices in general. Another influencing factor is the disciplinary arrangement about or attitude towards data sharing. This may differ from discipline to discipline.

- Lyon et al. (2010) "Disciplinary approaches to sharing, curation, re-use and preservation": This report focuses on seven case studies along four fields of research (life sciences, social science, architecture and climate) and aims to investigate researchers' perspectives and practices on data, methods and (software) tools. One result of this study is that institutional repositories have to develop domain-specific strategies because a generic approach will not cover the needs of researchers which are different by each discipline. However, three main points are located to establish good practices on data curation within each research discipline: (i) to change attitudes towards data management, (ii) to build up an infrastructure, and (iii) to train expertise in data curation.
- RIN and NESTA (2010) "Open to all? Case studies of openness in research": This report focuses on six different disciplines of research. Two key dimensions of openness are located: (i) What will be shared and (ii) with whom? The scope of openness or restriction depends on the disciplinary organisation of research. There are many advantages of data sharing such as (i) improved efficiency, (ii) improvements in research quality, (iii) enhanced visibility, (iv) ability to ask new questions, and (v) easier (inter-)disciplinary communication. But today, there are disadvantages as well, such as (i) a possible lack of credit, (ii) lack of time, (iii) competitive advantage, and (iv) ethical, legal and other restrictions.

The current discussion about OA is also based on many other studies, some of which should be noted. They focus on specific aspects such as on data storage, sharing and re-use. These practices have to deal with different questions. While data storage tends to address technical problems, data sharing has also to handle cultural aspects. If researchers re-use the shared data, common questions will be asked: (i) how can I understand shared data? (ii) how can I trust shared data? (iii) are there tools to assess data quality? (Faniel and Jacobsen, 2010). These questions relay directly to the importance of documentation of data as metadata.

This also renews questions about how to ensure the integrity, accessibility and stewardship of such data. Documentation of data will be one main part to ensure their integrity. High standards for openness and transparency are a primary prerequisite for integrity. Data sharing is most powerful if the generated data is part of an open flow of information and freely accessible. Stewardship has to handle problems like selecting preservable data (not all data can be preserved), documenting, referencing and indexing data as well to ensure the wealth of data sharing for research (Carlson and Anderson, 2007; National Academy of Sciences, 2009).

Probably one of the most important drivers to develop an improved research infrastructure is the upcoming deluge of data that cannot be handled without new research tools. These tools should easily operate the mass of data. Therefore, we need those standards for data and metadata which will allow sharing and access to information in general (Hey and Trefethen, 2005).

The benefits of shared data can be enormous for research (e.g. Hey, Tansley and Tolle, 2009): (i) reproducibility of research results will be simpler, (ii) it will advance research in general, (iii) new questions can be asked, and (iv) a public good will be returned to the public (Borgman, 2010). By now, in some research fields re-use and reanalysis are already integral part of research processes (e.g. life sciences, climate science and information and communications technology (ICT)) but in other fields data sharing, and the benefit of re-use and reanalysis is not put into practice due to specific requirements for sharing and storage (e.g. social science, Gläser and Laudel, 2008) Thus, the practice of data sharing is organised in research disciplines in different ways. But each discipline currently requires own standards for data formats to share and store their data. Beyond the problem of standardisation, further problems have to be worked out, such as the possibility of citing data (Nelson, 2009).

Therefore, it is helpful to take a look at the widest developed research area. Biology and medicine are two representative examples. There is a common sense to publish digitally and to improve the sharing of knowledge by using joint infrastructure. Many communities have started to build such infrastructures. But there are many seen and unseen problems by building these, which have to be solved, above all the fragmentation of infrastructural services. One possible solution could be that existing institutional and disciplinary silos are replaced by cybersilos (Buetow, 2005).

“What researchers want” is one of the great questions when designing research infrastructure. Two main issues are perceivable if you ask researchers about data: (i) they distinguish between data storage, and access during the project and after publication of results, and (ii) they also distinguish between raw, processed and annotated data. “The bottom line is that a researcher does not wish to be interrupted in what he wants to do most: his research" (Feijen, 2011).

2 Scope

This study will highlight the subject-specific requirements to get an in-depth understanding of today's research infrastructures and future needs:[2]

- Life Sciences and Health
- Information and Communication Technology (ICT)
- Social Science and the Humanities
- Research Infrastructure and e-Infrastructure
- Environment

This study is divided into three parts: (1) an introduction, (2) case studies about five research institutes as exemplars of research disciplines with structured descriptions, and (3) a comparative conclusion which synthesises our results.

The cases studies are at the heart of the whole study and each case study will elaborate the following four subjects:

- an overview of existing relevant information services and e-infrastructures in the respective subject area. It contains a description and analysis of diversity (e.g. publication behaviour, subject classification, research workflows, infrastructures, data types consider aspects such as: tools to generate data, measures of quality of the data), requirements for the publication deposit process and requirements for future infrastructures,
- a conceptual proposal of how subject-specific information services for OA infrastructure should look like,
- a vision for the next-generation information services exploiting OA principles from a disciplinary perspective and practical outputs as well as advices to future directions for funding agencies like the EC and others,
- an answer to the question: how can subject specificity be represented in OA infrastructure?

According to this, every case study will be structured as: (i) case narrative(s) to provide practical or specific insights into "researcher behaviour", (ii) current status of research infrastructure, workflows and research lifecycle

[2] The European Commission decided that its OA policy shall be first implemented as a OA pilot project within six of ten of the funding areas in the Seventh Framework Programme (FP7). The analysed research areas in this study of the OpenAIRE project correspond to the FP7 by the EC. For a detailed look at FP7 and the ten funded areas visit the following web page: http://ec.europa.eu/research/fp7/understanding/fp7inbrief/structure_en.html (consulted 9 August 2010).

focusing on specific aspects of the data management lifecycle, (iii) current status of OA to literature, (iv) current status of OA to data, (v) challenges, and (vi) future directions and summary. A detailed catalogue of research questions is given below.

3 Disciplines and institutions

Even though each different institution stands for a discipline, it should be noted that their subject-specific approach is not meant to represent the whole discipline. Other institutions in the same discipline might well have a different approach to perform research or to provide infrastructure. Thus, each institution is representing only itself as a case. This should give the reader an indication of how one particular subject-specific approach to research infrastructure looks like. All participating institutions were carefully selected to provide a rich and insightful analysis from their disciplinary areas. Two disciplinary areas (Environment and Research Infrastructure/e-Infrastructure) are represented by two institutions. All institutions will be characterised here briefly (alphabetical order) before they give their detailed account in the next chapters:

- **The Cognitive Interaction Technology – Center of Excellence (CITEC)** at Bielefeld University is an exemplar within the area of ICT with a highly interdisciplinary approach, including informatics, engineering, computing, linguistics, sports, biology, psychology and social science. It is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) as part of the German Excellence Initiative. CITEC describes itself in the following way: "The vision of the CITEC scientists are technical systems that can be operated easily and intuitively, ranging from everyday objects to fully-blown humanoid robots. The future technology should adapt itself to its human users instead of forcing us humans to adjust to the often cumbersome operation of the current equipment" (www.cit-ec.de, consulted 2 August 2010).
- **Consiglio Nazionale delle Ricerche – Istituto di Scienza e Tecnologie dell'Informazione (CNR-ISTI)** is an exemplar within the area of research infrastructures, e-infrastructures and computer science. This Italian institution stresses the importance of digital information providers as costumers. On their homepage the ISTI points out that "[t]he Institute is committed to producing scientific excellence and to playing an active role in technology transfer. The domain of competence covers Information Science, related technologies and a wide range of applications. The activity of the Institute aims

at increasing knowledge, developing and testing new ideas and widening the application areas." Specifically, the team collaborating to this report belongs to the Multimedia Networked Information System Laboratory, which consists of 48 researchers and technicians conducting research and development activities on algorithms, techniques and methods for information modelling, access and handling, with special focus on the design, development and production of middleware and services for dynamic and autonomic service-oriented infrastructures (SOA, Grid-based) capable of supporting the construction and sustainable maintenance of very-large networked multimedia information systems (http://galileo.isti.cnr.it/AboutISTI, consulted 2 August 2010).

- **The Department of Informatics and Telecommunications of the National Kapodistrian University of Athens** is also an exemplar within the area of research infrastructures in building and supporting e-infrastructures for scientific and health data management, digital libraries, cultural heritage interconnections, communication networks (www.di.uoa.gr).
- **The Italian Consultative Group on International Agricultural Research and Bioversity International (CGIAR/Bioversity International)** is an exemplar within the area of environment and agriculture. The main aim of CGIAR is to "reduce poverty and hunger, improve human health and nutrition, and enhance ecosystem resilience through high-quality international agricultural research, partnership and leadership" (http://www.cgiar.org/who/index.html, consulted 2 August 2010).
- **The Data Archiving and Networked Services (DANS)** is an exemplar within the area of social science and the humanities. The institute is under the auspices of Royal Netherlands Academy of Arts and Sciences (KNAW) which is also supported by the Netherlands Organization for Scientific Research (NWO). DANS characterises itself as follows: "DANS has been storing and making research data in the arts and humanities and social sciences permanently accessible. To this end DANS itself develops permanent archiving services, stimulates others to follow suit, works closely with data managers to ensure as much data as possible is made freely available for use in scientific research" (http://www.dans.knaw.nl/en/content/about-dans, consulted 2 August 2010).
- **The European Molecular Biology Laboratory/European Bioinformatics Institute (EMBL-EBI)** is an exemplar within the area of health and life science like genome research and bioinformat-

ics. The EBI branch in Cambridge (UK) points out that "[t]echnologies such as genome-sequencing, microarrays, proteomics and structural genomics have provided 'parts lists' for many living organisms, and researchers are now focusing on how the individual components fit together to build systems. The hope is that scientists will be able to translate their new insights into improving the quality of life for everyone. However, the high-throughput revolution also threatens to drown us in data. There is an ongoing, and growing, need to collect, store and curate all this information in ways that allow its efficient retrieval and exploitation. The European Bioinformatics Institute (EMBL-EBI), which is part of the European Molecular Biology Laboratory (EMBL), is one of the few places in the world that has the resources and expertise to fulfil this important task" (http://www.ebi.ac.uk/Information/About_EBI/about_ebi.html, consulted 2 August 2010).

- **The World Data Center for Climate/Deutsches Klima Rechenzentrum (WDCC/DKRZ)** is also an exemplar within the area of environment/climate. WDCC is part of the World Data Center System in earth sciences. WDCC is maintained by the Data Management division of the German Climate Computing Centre (DKRZ) located in Hamburg, Germany. The WDCC is aimed at collecting, scrutinising, and disseminating data related to climate change on all time scales. Emphasis is on data products from climate modelling and related observational data. The WDCC focuses on geo-referenced data using the operational CERA data and information system. Input is accepted in electronic form. At the WDCC, a visiting scientist programme exists. Facilities and services include data processing, copying and analysis. Data are available on most media including CD-ROM, via internet, and other media on request. On-line access exists via the World Wide Web, and FTP access is possible on request. A special project of WDCC is running the climate model part of the IPCC Data Distribution Center (DDC). The DCC of the Intergovernmental Panel on Climate Change (IPCC) facilitates the timely distribution of a consistent set of up-todate scenarios of changes in climate and related environmental and socioeconomic factors (http://www.mad.zmaw.de/wdc-for-climate).

4 Methods

Our study is designed as a comparative case study for the following reason. Subject-specific requirements may differ from institution to institution or even from laboratory to laboratory. Thus, an in-depth look into very specific institutional solutions is essential to describe such subject-specific requirements. A comprehensive analysis is therefore practically impossible due to the number of research institutions worldwide. Furthermore, averaging across different institutions has the risk of losing exactly the capability to observe the phenomena that are under scrutiny in this study, namely fine-grained differences of handling a specific research problem. On the other hand, a case study has the capabilities to provide detailed analyses and to detect even subtle differences. At the same time, the *comparative* approach allows findings to be generalised across subjects by elucidating coincidences and congruencies.

When different research institutes are compared, it is conceivable that similar research institutes with similar subjects use similar infrastructures while others use totally different infrastructures. This implies that there exists not only one solution that supports research in general and we have to accommodate different kinds of OA infrastructures to support as best as we can and explore OA all over science. A good and practical way to study these subject-specific requirements on infrastructures is to study single cases and compare them finally as a multiple case study. Each case can be based on different methods but all cases will answer nearly the same detailed questions.

Three methodological instruments – which are applied differently in each case study – are used:

- literature/document analysis
- interview
- observation.

Reviewing literature is the most obvious method to approach the subject-specific requirements. Collecting and analysing documents is a way to get an understanding of subject-specific infrastructure, their organisation, workflows, for example. Analysis – as opposed to literature analysis – uses scripts to explain workflows, data storing or the like. Most information can be extracted by analysing institutional papers about their infrastructure. If literature and document analysis leaves unanswered questions, interviews could be conducted. These could be semi-structured, recorded or transcribed. Observations are needed to get access to real internal meetings and workflows. All observations are recorded by video cameras and transcribed, too.

The depth of research methodology that is applied in case studies is left open and decided by the subject specialists who author the case study. In some cases, literature analysis is sufficient; in other cases, advanced methods, such as interview and observation, are required.

In sum, the first two methods (literature/document analysis and interview) are needed to get a theoretical understanding of the subject-specific infrastructure. The last method (observation) can help us to understand the practical value of infrastructure. By triangulation of these methods, we can draw a comprehensive picture of the current (OA) infrastructure, their design and their usage. Thus, we get a highly credible description and analysis of publication behaviour, subject classification, data types, research workflows and infrastructures, to name but a few facets of OA intrastructure.

5 Research questions

The ensuing catalogue of questions re-formulates the conceptual questions of the previous section (*Scope*) to put research in concrete terms, and makes our research work itself co-inciding and comparable. They shall help to give each case study a common thematic scope. Each case also has additional lists of research questions.

I. Literature

1. *Literature management*
   a. How is literature produced and managed?
   b. Which tools support these practices?
2. *Publication services and policies*
   a. Which forms of publication are common at your institute?
   b. Is OA already established as an equal alternative to commercial publishers?
   c. Which new forms of publication services are on horizon?

II. Data

1. *Storage, preservation and curation*
   a. What tools are followed regarding data storage?
   b. What tools are there for people to follow good practice with respect curating and preserving their research outputs?
2. *Processing and manipulation*
   a. What tools enhance data by processing and manipulation?
   b. What value (e.g. metadata) is added to data as they pass through different stages of processing?
3. *Policies, access and sharing*
   a. What policies (formal/informal) exist and how do tools reflect these policies?
   b. What practices are followed for sharing outputs and which tools are used?
   c. What limitations are there on access to research outputs?
4. *Quality assurance*

a. What practices exist in your field for controlling quality in research outputs (similar to the procedure of peer review)? And which tools support these controlling practices?

6 Bibliography

Angrosino, M. *Doing Ethnographic and Observational Research*. Sage Publications, London, 2007.

Bohlin, I. Communication regimes in competition. *Social Studies of Science* 2004, 34, 365–391.

Borgman, L. Research Data. Who will share what, with whom, when, and why? China-North America Library Conference, Beijing. Available at: http://works.bepress.com/borgman/238. 2010.

Buetow, KH. Cyberinfrastructure: Empowering a "Third Way" in Biomedical Research. *Science* 2005, 308, 821–824.

Carlson, S. & Anderson, B. What *are* data? The many kinds of data and their implications for data re-use. *Journal of Computer-Mediated Communication*, *12*(2), 15. Available at: http://jcmc.indiana.edu/vol12/issue2/carlson.html. 2007.

Faniel, IM & Jacobsen, TE. Reusing scientific data: how earthquake engineering researchers assess the reusability of colleagues' data. *Computer Supported Cooperative Work* 2010, 19, 3–4.
Feijen, M. (SURF) What researchers want: A literature study of researchers' requirements with respect to storage and access to research data. Available at: http://www.surffoundation.nl/nl/publicaties/Documents/What_researchers_want.pdf. 2011.

Gläser, J. What internet use does and does not change in scientific communities. *Science Studies* 2003, 16, 1.

Gläser, J & Laudel, G. Creating competing constructions by reanalyzing qualitative data. *Historical Social Research* 2008, 33, 3.

Gomm, R. *Key Concepts in Social Research Methods*. Palgrave Macmillian, Hampshire, 2009.

Greyson, D, Vézina, K, Morrison, H, Taylor, D, & Black, C. University supports for Open Access: a Canadian national survey. *Canadian Journal of Higher Education* 2009, 39, 1–32.

Harley, D, Krzys Acord, S, Earl-Novell, S, Lawrence, S, & King, CJ. *Assessing the Future Landscape of Scholarly Communication.* Centre for Studies in Higher Education, University of California Press, Los Angeles, London, 2010.

Harley, D, Earl-Novell, S, Arter, J, Lawrence, S, & King, CJ. et al. *The Influence of Academic Value on Scholarly Publication and Communication Practices.* Centre for Studies in Higher Education, University of California, Berkeley, 2006.

Hey, T, Tansley S, & Tolle, K, eds. *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Microsoft Research, Redmond, Washington, 2009.

Hey, T & Trefethen, AE. Cyberinfrastructure for e-Science. *Science* 2005, 308, 817–821.

Lee, CP, Dourish, P, & Mark, G. The human infrastructure of cyberinfrastructure. ACM, New York, 2006.

Lyon, L, et al. (DCC-SCARP) Disciplinary approach to sharing, curation, reuse and preservation. Available at: http://www.dcc.ac.uk/sites/default/files/documents/scarp/SCARP-FinalReport-Final-SENT.pdf. 2010.

National Academy of Science. *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age.* National Academy of Science, Washington, DC, 2009.

Nelson, B. Empty archives. *Nature* 2009, 46, 160–163.

Research Information Network (RIN) and National Endowment for Science Technology and the Arts (NESTA). Open to all? Case studies of openness in research. Available at: http://www.rin.ac.uk/system/files/attachments/NESTA-RIN_Open_Science_V01_0.pdf. 2010.

Royal Society. *Knowledge, Networks and Nations.* Royal Society, London, 2011.

Taubert, NC, Weingart, P. “Open Access”. Wandel des wissenschaftlichen Publikationssystems. In: Sutter, T & Mehler, A (eds.) *Medienwandel als Wandel von Interaktionsformen*. VS-Verlag, Wiesbaden, 2010.

Theodorou, R. OA repositories: the researchers’ point of view. *The Journal of Electronic Publishing* 2010, 13. Available at http://dx.doi.org/10.3998/3336451.0013.304.

B Agricultural Research

Hugo Besemer, Chris Addison, Francesca Pelloni, Enrica M. Porcari and Nadia Manning-Thomas

1 Introduction

Agricultural science combines amongst others applied socioeconomic disciplines, applied plant animal physiology and environmental sciences (soil science, hydrology, erosion/geomorphology).

Research workflows, like for other applied sciences, depend on the disciplines and methods that are applied, as well as on the way that the organisation that does the research is embedded in the agricultural sector. This chapter was written from the perspective of the Consultative Group on International Agricultural Research (CGIAR), a global partnership that unites organisations engaged in research for sustainable development with funders, including governments, foundations and international and regional organisations. CGIAR's mission implies working for international development, but many of the processes apply to national agricultural research organisations as well. As it impossible to give a general framework for research workflows in our field, we will present case studies from the CIAGR to illustrate the diversity. Typically these workflows include processes such as:

- problem identification and analysis,
- observations/acquisition of data,
- analysis of results,
- possibly design and testing of a remedy (e.g. pest or erosion control measures),
- validation (e.g. checking that a proposed solution work at farm level),
- publishing and dissemination.

In our field, problem identification includes consultation with organisations of beneficiaries and it may be done in the framework of applications for funding. In the case of the CGIAR, these are usually organisations for international development. Such organisations may have specific requirements for the disclosure of research results and some are in the process of formulating their own policies with regard to storage and accessibility of data sets.

Data may be acquired directly through observations but may also be acquired from other parties. For example, satellite images or digital maps are used for research with a spatial component. These images are often purchased from commercial firms. Observations on farm level require collaboration with farmers and may often be done in collaboration with local institutes or for example extension organisations. The latter do the initial data collection, while the data is further processed at research centres. Both in the case of satellite images and in the case of on-farm surveys, the question arises who owns the data, as data is collected and value is added in a chain.

The validation step is crucial for agricultural research as its results need to be communicated with potential beneficiaries, like rural communities. Some decades ago there was in many countries a clear division of labour between research bodies and extension agencies that were often publicly funded. In recent years, agricultural extension has changed and it is beyond the scope to describe the very diverse way that the agricultural knowledge systems have developed in different countries and different parts of the world. But as this picture is getting less straightforward, the communication of research results beyond the circle of scientific peers is becoming a direct concern for agricultural research organisations. In this context, in 6.5 Knowledge sharing, we will be drawing some lessons from CGIAR's efforts with regard to knowledge sharing. Data sets may be used to communicate with and collaborate with scientific peers, but they may also be knowledge products to communicate the results of research with the potential beneficiaries and the general public. The case studies were selected to illustrate these issues and the diversity of scientific methods in agricultural science. The CGIAR is described as an organisation in more detail in section 2.

From the socioeconomic angle, there are two case studies:

Socioeconomic surveys: which concentrate on the level of individual households or farms. Agriculture is an activity that is often carried out by such smallholders. For these surveys, the data curation/data repository approach appears to be appropriate.

Ongoing agricultural research and development capacity survey: which concentrates on the level of national agricultural knowledge systems. While agriculture is often an activity of smallholders, research and development is often an activity that is not done within the individual enterprises, but in national or regional organisations. Raw data is acquired in a variety of forms, and is stored in a central database.

From the genetic and environmental angles, there are also two case studies:

Multisite agricultural trial database for climate change analysis:

which describes an effort where traditional outputs of agricultural research (field trials) are combined in a model with existing environmental (climate) and geographical data.

Plant genetics resources – the Singer system and further: which describes research activities where the primary output is not a collection of data (in a collection of sets or a database) but a collection of certified and documented seeds. The data collection activities are aimed at making the collections of seeds accessible for experiments where genetically uniform plant material is required and taking stock of the certification and documentation process. Technically this is a central database importing on an ad hoc basis updates from local databases.

The acquisition management, publishing and dissemination of these sets are illustrated in the example case studies chosen here.

Lessons for OpenAIRE

- Models and other integrated knowledge products may be an alternative to repositories to bring together data sets from different sources.
- Research results may be captured in databases rather than static data sets. A data preservation and re-use policy should take databases into account.

1.1 Case study: socioeconomic surveys: International Food Policy Research Institute (IFPRI)

The IFPRI is an international agricultural research centre working on informing national agricultural and food policies to find sustainable solutions for ending hunger and poverty. Much of the Institute's research work relies on data collected through socioeconomic surveys and experiments. This case study describes the steps involved and the issues affecting the acquisition, storage and dissemination of these data (Figure B.1).

The story begins with the collection of raw data in the field. This has changed recently with the adoption of new technologies for the recording of information and new approaches to capture data.

The IFPRI Mobile Experimental Economics Laboratory (IMEEL) was established in 2007 by the Markets, Trade, and Institutions Division (MTID) of the IFPRI. Its primary objective is to collect data through economics experiments in the field to better understand the behaviour of smallholders and the poor in rural areas, especially in Africa, Central America and the Caribbean, Latin America and south-east Asia (Vargas, 2010; Vargas and Viceisza, 2010). These experimental data are usually combined with survey data to understand farmers' decisions on the adoption of new technologies,

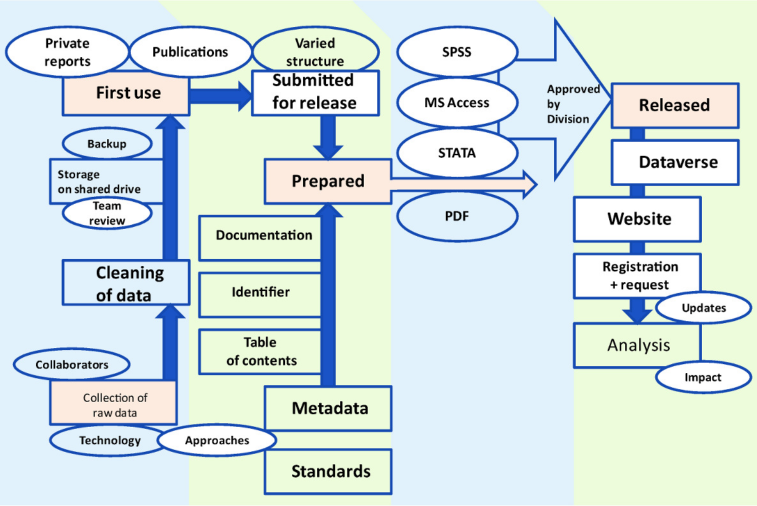


Figure 1: Workflow of data from collection by researcher through dissemination to user to analysis of use

Figure B.1 IFPRI workflow of data from collection by researcher through dissemination to user to analysis of use

participation in marketing activities; contracting arrangements and farmer groups.

A number of methods are used for collecting data including a variety of personal digital assistants, cell phones and tablets. Whilst there may be different risks in digital collection, the advantages of software to improve data collection provide increased efficiency in the collection and reduce the need for processing. For example, the software includes controlled responses and range checking, thus reducing errors in collection. The main software used for the surveys includes mQuest, Satellite Forms and Lime Survey. The output in each case is a rectangular data file readable into statistics packages or Microsoft Excel. The choices of handheld devices for data capture is based on their battery life, ease of use and their durability.

The capture of raw data involves a number of collaborators from other organisations who are often given the opportunity to use the raw data for their own studies. The capture of data in digital media in the field means that adequate backup needs to be in place in the field environment.

The data captured is then cleaned by the research team and will then be stored in a shared area for review and validation. Whilst the data is held on the shared drive it is regularly backed up from the Institute's servers.

The data will then be used within the organisation either for the production of a donor report or limited distribution report or for a publication. The software used to analyse the data during this stage is SPSS, Stata, Excel or Access. Any models produced or developed during this stage are held on the researcher's machine or the shared drives. Several of these models will be worked into a knowledge product and shared with the public through the institutions website.

The data is not released until the derived research is published. Once used for a publication, the publications review committee will require the author to submit the supporting data set. This may submitted in several forms: STATA, SPSS, Excel, Access and PDF. It is then tidied, documented and packaged by the Library and Knowledge Sharing Unit in discussion with the researcher. A table of contents will be produced to indicate the various supporting components of the data set which comprises original questionnaires and resultant data sets. Attention will be paid to ensuring anonymity of survey participants, standard formats for files where applicable and the addition of appropriate metadata.

Once approved by the Division, the resultant files and records are then published both on the internet site and in and external repository: Dataverse.[1] Dataverse is a data repository run by Harvard which provides metadata storage, file format conversion, collection management and customisation of display. Users coming through the website are asked to register and record how they will be using the data, so that analysis of use is possible later and users can be informed when there are updates of the data set or similar data sets are available.

Models and tools developed during the analysis may be similarly packaged but are normally provided online as knowledge products through the institute's website. Increasingly tools are being provided online through the site itself and through portals. For example, the welfare simulator embedded on the food security portal allows users to use their own data and run the simulator online.

There has also been a move not just to provide data online for download but to provide access to the data through application programming interfaces and visualisations of the data through interactive maps and graphs.

1.2 Case study: ongoing agricultural research and development capacity survey

Agricultural research capacity can only be developed if it can be measured. This is the premise of the formation of the Agricultural Science and Tech-

[1] http://dvn.iq.harvard.edu/dvn/dv/IFPRI.

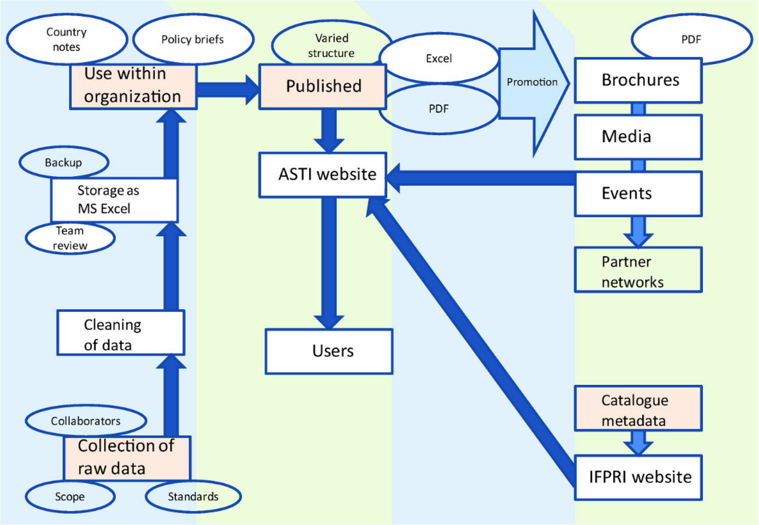


Figure B.2 ASTI data collection, management, dissemination and promotion

nology Indicators (ASTI) project to capture data on the current state of agriculture research in a selection of countries (Figure B.2). The site http://www.asti.cgiar.org hosts this data and the country notes and policy briefs which result from their analysis.

Raw data is collected in collaboration with partners using the OECD Frascati manual with some adjustment for the collection of data. This allows the data collected to be compared with other data sets more readily. Standard definitions are used to define scope. The data is collected with collaborators and consultants and the national partners coauthor and copublish the data in the form of country notes. These notes are produced from the raw data set but will share different levels of detail depending on the Intellectual Property Rights agreed.

The form of questionnaire for the collection differs according to the source of information. There are currently three types of questionnaire: one for NGOs and government departments, one for higher education and one for the private sector. These are constantly improved and revised as necessary.

ASTI manages a portfolio of data, from time series data across country, regional and global level covering agricultural research and development investments, institutional arrangements, funding sources, degree qualifications and female participation in agricultural research and development (ASTI, 2011; Norton, 2010).

The ASTI project has recognised the importance of promoting the data sets, realising that although the sets are valuable and there is a ready demand, active steps need to be taken to reach the potential audience. With this in mind, they have an active communication strategy and have produced a number of specific promotional products and held a number of media events and policy seminars.

The three levels of output have been the country briefs and notes; the data sets themselves and the website to allow the user to investigate the data themselves. These outputs have been complimented by promotional activities such as the ASTI blog, brochures, flyers and posters. ASTI seminars and outreach events to reach the variety of stakeholders and media and working through partners' own workshops and capacity-building programmes to raise awareness of the data sets. One of the major challenges has been to communicate with such a diverse range of stakeholders and make ASTI data known.

1.3 Case study: multisite agricultural trial database for climate change analysis

The online database developed at agtrials.org is the development platform for the CGIAR research programme on Climate Change and Food Security (CCAFS) Global Trial Sites Initiative. It shows the result of discussions between plant breeders running the agricultural trials and the geographers from a spatial data background (Figure B.3).

Agtrials.org is a development organised through the community working within the CCAFS and emphasises a pragmatic approach to the collection of metadata and data which reflects the realities of the diverse research environments involved. A series of trials were identified which could be easily incorporated into the database with emphasis on what was possible within existing time and resource constraints. The application development focussed on providing a data repository application where users could easily load historical trial metadata and information on current trials within the CCAFS programme. It needed to provide both private and public access. It built on experience on previous systems which were purely location based and incorporates the requirements of the plant breeders.

Data is provided in a variety of formats and development of the application is continuing to accommodate the design of the database and metadatabase, which can cope with the different types of user. Researchers also provide, where available, information on weather conditions during the trial and soil characteristics. There was no off-the-shelf solution to this requirement and the project develops with the contracts between the programme and researchers

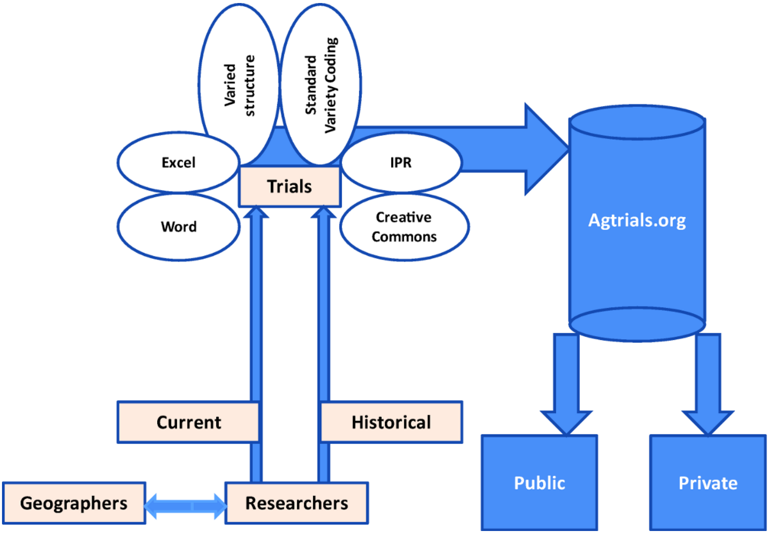


Figure B.3 Trials data storage and sharing using agtrials.org

developing hand in hand with this reference system. Most data is in Microsoft Excel file formats or Microsoft Word with no standardisation on format, but all sets follow a standard nomenclature for varieties. The technical format of the trial results has been kept open at this stage to encourage registration of a variety of sets.

Whilst each data set will have a statement on intellectual property rights, the same rights are not used across the site. Guidance is provided on the use of Creative Commons Licences but as many different organisations are involved there is no blanket statement. A series of user guidelines are available to explain the use of data from the site. In addition to current and historical trials, there is now an option to add "simulated" trials from crop simulation models. More models are planned to be developed within the group.

The interesting part of this system is not the database alone but the process by which the community is developing a data reference point for the CCAFS programme, with the dual approach of developing the research relationships with the programme and consolidating the reference index for the trials involved. The subsequent phases of the project will include models that can identify analogue environments so that the result of one set of trials can provide information on the performance of a variety in a similar environment elsewhere.

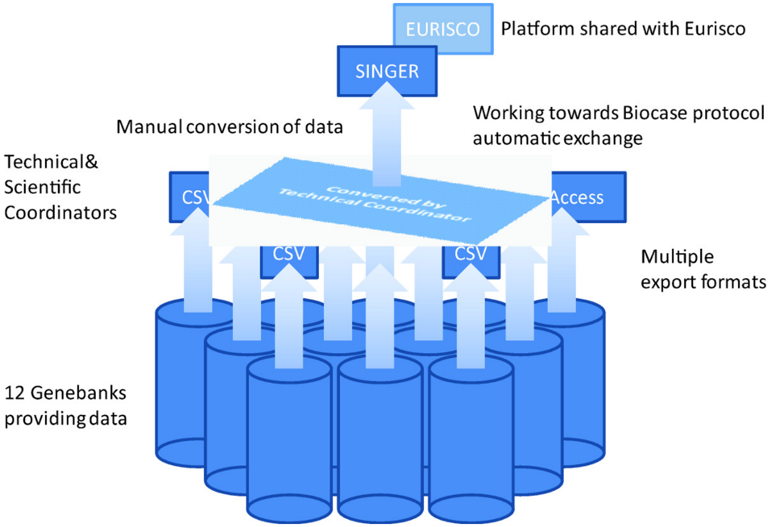

EURISCO
Platform shared with Eurisco
SINGER
Manual conversion of data
Working towards Biocase protocol automatic exchange
Technical& Scientific Coordinators
CSV
Converted by Technical Coordinator
Access
CSV
CSV
Multiple export formats
12 Genebanks providing data

Figure B.4 The Singer system providing access to the CGIAR gene banks

1.4 Case study: plant genetics resources: the Singer system and further

There are a variety of data-oriented systems maintained by CGIAR centres and partners in the field of plant genetic resources. We will now concentrate on the oldest and most mature system. Singer gives access to the collections of the gene banks of different centres (Figure B.4). Of more recent data is the Crop Genebank Knowledge Base where instructional materials are collected on best practices for gene banks. For the "Generation Challenge Program", a facility is under development to collect data sets from this programme on molecular plant breeding and there is a portal to collect field reports on different crops.

The basic currency that is dealt with in gene banks and the Singer system are accessions, certified plant propagation, such as bags of seeds. Certain fields in agronomic research cannot do without them and that is probably the reason why this is one of the oldest and most mature cross-centre data-oriented operations within the CGIAR. The system is managed through an informal working group of technical and scientific representatives from the 12 participating gene banks. There is a scientific coordinator at Bioversity International as well as a technical manager who performs all database-oriented operations.

Data from the different gene banks are sent in by the different gene banks in a wide variety of technical formats, like MS/Access databases, CSV files, etc. Conversion to the central database is performed on case-by-case bases by the technical coordinator at Bioversity. Around 2005, attempts were made to develop a more standardised updating method using the WSDL/UDDI web services technology that was developing at that time. The Biocase[2] protocol was developed in conjunction with Singer, and was deployed at six centres. However, there were two bottlenecks: the performance (speed) of the system that implemented the protocol and the relative difficulty to produce the flat files that were required by the system from the various database implementations with which the participating gene banks are managed. The Biocase protocol had been implemented successfully elsewhere, for the Global Biodiversity Information Facility (GBIF). In the Singer network, it is recognised that data-transfer methods and an upload facility should be developed in the years to come. But the Singer case shows that if there is a perceived need for exchange a cooperative system can be kept alive without a sophisticated technical backbone. Until now, the most important investments have been done in the intellectual foundations of the system.

The minimal data elements to describe an accession have been laid down as the FAO/IPGRI Multicrop-passport Descriptors[3] that were agreed in 1997 and updated in 2001 . . . (Alercia, Diulgheroff, & Metz, 2001). For specific needs, there are extensions like the guidelines for developers of crop descriptor lists.[4] The data model of the Singer database is documented on the Singer website, thus giving further guidance for data harmonisation.

With regard to data quality, there is an ongoing capacity development effort aiming at improving the management of gene banks. Instruction materials can be found at the Crop Genebank Knowledge Base.[5]

There are internal agreements on how the data that is entered in the system is used. The purpose of the Singer system is the discovery of accessions and it should lead to a transaction whereby a scientist requests seeds or other plant propagation materials for further research. These transactions, the documentation to come with the material and the obligation to share results with the originators of the material are governed by the "Standard Material Transfer

[2] http://www.biocase.org/products/protocols/index.shtml.

[3] http://www.bioversityinternational.org/nc/publications/publication/issue/faoipgri_multi_crop_passport_descriptors.html.

[4] http://www.bioversityinternational.org/index.php?id=19&user_bioversitypublications_pi1[showUid]=3070.

[5] http://cropgenebank.sgrp.cgiar.org.

Agreement (SMTA)".[6] Singer data is shared with the European network of gene banks, Eurisco, that is using the same database facilities at Bioversity.[7]

The most important lesson from this case is the need to invest in the intellectual infrastructure of a network for data exchange.

2 Current status of research infrastructure workflows and research life cycle

2.1 Introduction to the research infrastructure

The CGIAR is a global partnership that unites organisations engaged in research for sustainable development with the funders of this work. The funders include developing and industrialised country governments, foundations and international and regional organisations. CGIAR research is dedicated to reducing poverty and hunger, improving human health and nutrition and enhancing ecosystem resilience. It is carried out by 15 members of the Consortium of International Agricultural Research Centers in close collaboration with hundreds of partner organisations, including national and regional research institutes, civil organisations, academic institutions and the private sector.

A research organisation like the CGIAR that studies agriculture from all different angles should be compared to a university as a whole rather than to individual research groups and institutes. For the CGIAR, an extra complication is that the organisation operates in centres on different continents. The centres are partly organised on a disciplinary basis (e.g. rice, tropical crops, genetics) but increasingly on a more multidisciplinary basis (e.g. arid environments, agroforestry) and along new interdisciplinary programmatic axes, the CGIAR research programmes.

The Consortium of International Agricultural Research Centers[8] was established in April 2010, as part of a major reform of the CGIAR, this year celebrating its 40th year. The Consortium was formed to ensure closer alignment with the needs of partners and beneficiaries and to lead, coordinate and support the 15 member centres that make up the Consortium, some of which have been carrying out agricultural research with resource-poor farmers and their communities, for over 50 years. The Consortium supports and facilitates system-level approaches and interactions and has responsibility for

[6] http://singer.cgiar.org/index.jsp?page=smta.

[7] http://eurisco.ecpgr.org/static/about_eurisco.html.

[8] http://consortium.cgiar.org.

formulating strategy[9] and for developing multiyear and multicentre research programmes that implement on that strategy.[10] The Consortium employs over 9000 staff operating in over 150 locations (Figure B.5).

2.2 Scientists, centres and system-wide programmes

Whilst individual scientists are employed by one of the CGIAR centres, they are increasingly outposted on other campuses of other centres or partner organisations. They therefore use the research infrastructure of their host, their employer and the programme for which they work. A few of these examples were drawn upon in the case studies already presented. The first example showed the situation in a centre, the next the situation for a CGIAR research programme and the others for existing system-wide programmes. The solutions vary for different work, but groups sharing data-platform requirements across the CGIAR centres (e.g. the Consortium for Spatial Information; see 4.5.2) are increasingly using shared data platforms and carrying out joint developments.

2.3 Knowledge sharing

The approach to the role of knowledge sharing in the CGIAR has changed significantly during the last 5 years. The system has developed a dedicated group to encourage knowledge sharing across the system and many of the centres now have job titles including knowledge sharing.

Research organisations like the CGIAR cannot be satisfied just knowing they have produced high-quality science. It is essential that the outputs of their research are communicated and put to use, in the village, on the ground, in the lab or across the negotiating table.

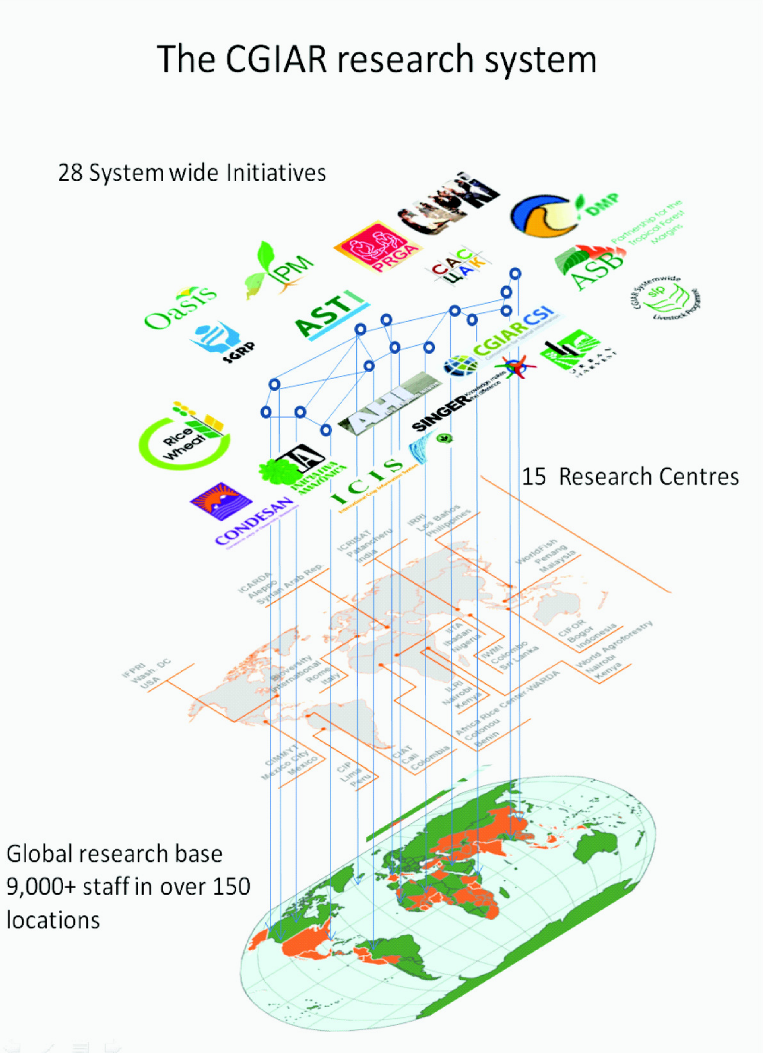
Therefore, the Triple-A Framework[11] was developed by the CGIAR ICT-KM programme looking at the availability, accessibility and applicability of research outputs and knowledge from the CGIAR. According to the framework, the three As are:

- **availability:** "can I find it?" – the need to assemble and store outputs so they will be permanently accessible and to describe them in systems so others know, and can find, what has been produced.

[9] See the Strategy and Results Framework (SRF): http://consortium.cgiar.org/our-strategic-research-framework.

[10] See the CGIAR Research Programs (CRPs): http://consortium.cgiar.org/our-strategic-research-framework/cgiar-research-programs-crps.

[11] http://ictkm.cgiar.org/document_library/program_docs/ICT-KM%20AAA_complete.pdf.



The CGIAR research system

28 System wide Initiatives

15 Research Centres

Global research base
9,000+ staff in over 150
locations

Figure B.5 Research structure of the CGIAR

- **accessibility:** “can I put my hands on it?” – the need to make outputs as easy to find and share and as open as possible, in the sense that others are free to use, re-use and redistribute them.
- **applicability:** “can I make use of it?” – the need to ensure that research and innovation processes are open to different sources of knowledge and outputs that are easy to adapt, transform, apply and re-use.

The framework is aimed at managers, researchers and information professionals to help them better understand the current AAA status of their research knowledge, how to identify, choose and develop pathways to improved accessibility for their outputs and eventually to improve chances that they will be put to use.

The first part of the framework is a benchmarking exercise which seeks to evaluate and measure the availability and accessibility of research outputs at a given time. This then helps CGIAR centres and programmes and their scientists decide on the level of AAA they want for their research outputs and also the pathways with which to turn these outputs into international public goods.

The Triple A approach has been developed and promoted to encourage sharing of international public goods produced by research. There has been more adoption of action-oriented research, more knowledge sharing during projects, changes in the peer-review process and more interim results are made available. There are new requirements from external stakeholders, such as journalists requesting access to data. There are new ways of working with fellow researchers outside the organisation, such as platforms like Basecamp and wider use of social media (Figure B.6).

3 Current status of Open Access in agriculture

As there is no clear indicator of how to measure the state of Open Access in a domain, we have approached it in two ways: (i) assessing how Open Access journals are indexed in major indexes, i.e. Web of Science (ISI) and Scopus; and (ii) overviewing Open Access document repositories within the CGIAR.

3.1 Coverage of agricultural Open Access journals in scientific journal metrics indexes

Our main question address the success of the Golden route to Open Access for agriculture, compared with two other subject domains.

[12] http://ictkm.cgiar.org/document_library/program_docs/ICT-KM%20AAA_complete.pdf.

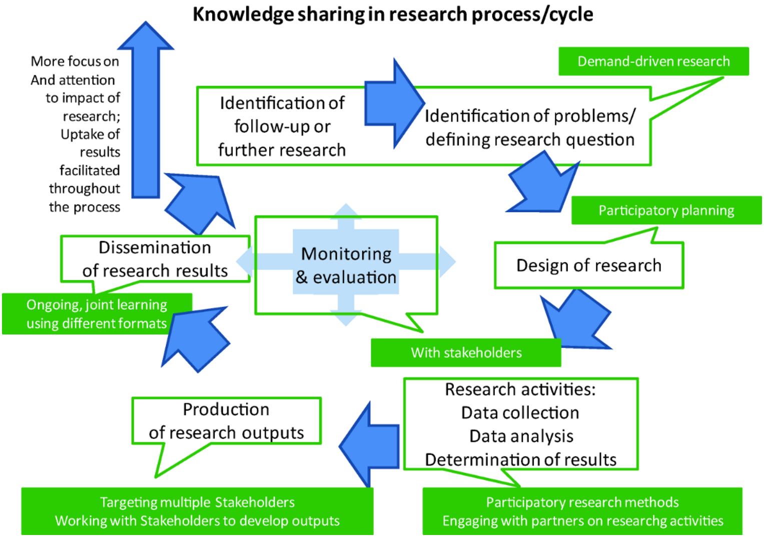

Knowledge sharing in research process/cycle
More focus on And attention to impact of research; Uptake of results facilitated throughout the process
Identification of follow-up or further research
Identification of problems/ defining research question
Demand-driven research
Participatory planning
Design of research
Monitoring & evaluation
With stakeholders
Dissemination of research results
Ongoing, joint learning using different formats
Research activities: Data collection Data analysis Determination of results
Participatory research methods Engaging with partners on researchg activities
Production of research outputs
Targeting multiple Stakeholders Working with Stakeholders to develop outputs

Figure B.6 Knowledge sharing in research processes and cycles (from Nadia Manning-Thomas from background note[12] for the Consortium of CGIAR Centers)

The most comprehensive list of Open Access journals is the Directory of Open Access Journals (DOAJ). We have matched (using title, ISSN and e-ISSN) the DOAJ journal list with the list of the most important lists for scientific journal metrics, i.e. Scopus, that calculates Scimago Journal Rankings – SJR and SNIP values) and the Journal Citation Report (ISI) that calculates the journal impact factor (IF).

For a cross comparison, we kept the DOAJ subject classification for each journal. The coverage of Open Access journals in different subfields of agriculture is given in Table B.1 and Figure B.7. In short, an Open Access journal in agriculture has a chance of 38% to be included in Scopus and 27% to be included in JCR (ISI).

In Table B.2 we compare these figures with the field of biology. The table shows that Open Access publishing in the field of biology is more successful than in the field of agriculture.

Finally we did the same for the field of medicine and health sciences (Table B.3). These results indicate that Open Access journals are less successful in medicine than in agriculture and in biology.

The question remains what the percentage of Open Access journals is of the total of journals in Scopus and JCR. We can make this comparison partially

Table B.1 Publications in Open Access journals: agriculture

DOAJ category	Scopus	JCR (ISI)	DOAJ
Agriculture (general)	33	24	104
Animal sciences	30	24	74
Aquaculture and fisheries	6	3	9
Biotechnology	16	10	25
Forestry	7	7	28
Nutrition and food sciences	4	5	26
Plant Sciences	19	10	33
Total	115	83	299
Percentage of DOAJ	38	27	100

Table B.2 Publications in Open Access journals: biology and agriculture

DOAJ category	Scopus	WOS (ISI)	DOAJ
Biochemistry	15	12	36
Biology	63	53	65
Botany	23	16	63
Cytology	6	2	7
Ecology	11	5	35
Genetics	20	17	37
Microbiology	14	11	38
Total	152	116	281
Percentage of DOAJ	54	41	100

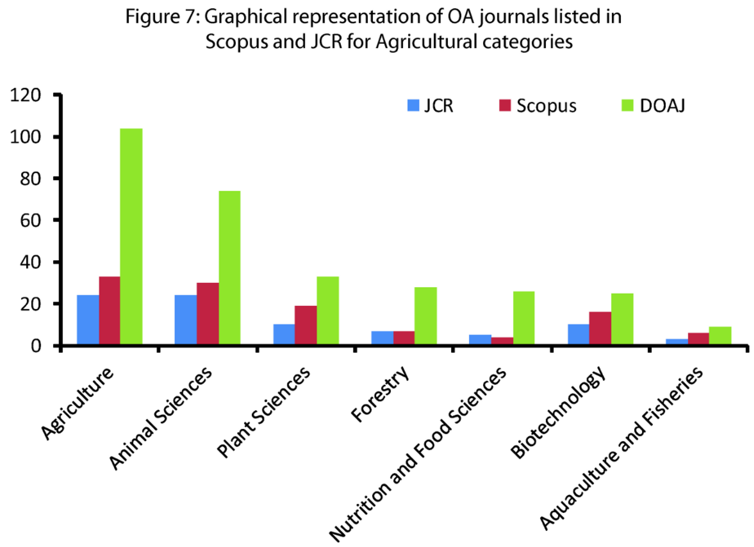


Figure B.7 Graphical representation of Open Access journals listed in Scopus and JCR for agricultural categories

for Scopus, and not at all for JCR(ISI): in both cases, the subject categorisation is different. For JCR a meaningful comparison is not possible; Scopus lists the Agricultural and Biological sciences as one category. In that category, 13.4% of the journals are Open Access, against 5.9% of the journals in the medical field.

The overall conclusion of this exercise is that agriculture is not behind the other research fields studied. One caveat is we have only looked at the total number of journals listed, not the relative importance of the journals (as expressed through IF, SJR and SNIP values).

3.2 Open Access repositories in the CGIAR

CGIAR's activities to make its publications available and accessible has resulted in publication databases/institutional repositories at all centres (Table B.4).

These repositories have been analysed in an article by Arivananthan, Ballantyne and Porcari, (2010). The content of the repositories from six centres was assessed. It appeared that there is a huge variation with regard to the availability of full text for publications (from 19% to 100% of the document descriptions). Especially articles from peer-reviewed journals were missing.

Table B.3 Publications in Open Access journals: medicine and health sciences

DOAJ category	Scopus	WOS (ISI)	DOAJ
Medicine (general)	167	69	378
Allergy and immunology	12	3	19
Anatomy	3	1	10
Anaesthesiology	5	0	11
Cardiovascular	22	11	59
Dentistry	14	3	67
Dermatology	9	1	20
Gastroenterology	9	3	26
Gynaecology and obstetrics	9	1	32
Internal medicine	104	36	237
Neurology	34	15	80
Nursing	6	2	31
Oncology	31	9	66
Ophthalmology	7	1	26
Otorhinolaryngology	3	0	17
Pathology	15	2	32
Pediatrics	14	3	44
Pharmacy and materia medica	29	7	65
Physiology	13	9	25
Psychiatry	16	7	40
Public health	50	17	47
Sports medicine	3	1	17
Surgery	24	9	69
Therapeutics	29	12	64
Urology	9	0	15
Total	637	222	1497
Percentage of DOAJ	42	14	100

Around 40% of peer-reviewed journal articles and 54% contributions to peer-reviewed books could be found in Google Scholar. However, it should be remarked that, since the study, many centres have collaborated in the Google Books publishers programme and this may likely have improved the coverage of books and book chapters.

To improve the accessibility of CGIAR's publications, there are more opportunities such as collecting pre-prints of articles in peer-reviewed journals (Green route to Open Access). It is suggested that the coverage in search engines can be improved, for example by uploading sitemap files. In the Opendoar[13] registry of Open Access repositories, seven CGIAR centres can

[13] http://www.opendoar.org.

be found, while 11 centres participate in FAO's AGRIS system.[14] Through AGRIS, these publications are also indexed in Google Scholar, but material from centres that do not participate in AGRIS can also be found back there. It is beyond the scope of this study to check systematically the coverage of CGIAR publications in these external systems, but it would be interesting to see how it develops in view of these efforts.

Table B.4 Open Access repositories in the CGIAR

Institute/programme	Start year of systematic collection of full texts	Earliest publication
Bioversity[15]	2004	1977
CGIAR secretariat[16]	"latest titles"	
Center for International Forestry Research (CIFOR)[17]	2001	1993
International Center for Agricultural Research in the Dry Areas (ICARDA)[18]	2005	1977
International Center for Tropical Agriculture (CIAT)[19]	2001	
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)[20]	2001	
International Food Policy Research Institute (IFPRI)[21]	2000	
International Institute of Tropical Agriculture (IITA)[22]	2005	1990
International Livestock Research Institute (ILRI)[23]	2008	1977
International Maize and Wheat Improvement Center (CIMMYT)[24]		
International Potato Center (CIP)[25]		
International Rice Research Institute (IRRI)[26]	2007	1999

[14] http://agris.fao.org.

[15] http://www.bioversityinternational.org/publications/search.html.

[16] http://www.cgiar.org/publications/secretariat.html.

[17] http://www.cifor.cgiar.org/online-library/browse.html.

[18] http://icarda.catalog.cgiar.org/textbase/search.htm.

[19] http://ciat.catalog.cgiar.org/ciat_bibliography.html.

[20] http://dspace.icrisat.ac.in.

[21] http://www.ifpri.org/pubs/pubs_menu.asp.

[22] http://biblio.iita.org/index.php?page=pubyear&kind=year&type=iita.

[23] http://mahider.ilri.org.

[24] http://www.cimmyt.org/en/services/library/recent-publications.

[25] http://cip.catalog.cgiar.org/cat-cip.asp.

[26] http://ricelib.irri.cgiar.org:81/screens/opacmenu.html.

International Water Management Institute (IWMI)[27]	1984	
World Agroforestry Centre (ICRAF)[28]	2004	1978

In 2006, CGIAR launched the CGIAR Virtual Library providing access to research on agriculture, hunger, poverty and the environment. This is a shared, integrated service that allows users to tap into leading agricultural information databases, including the online libraries of all 15 CGIAR centres.

3.3 Open Access mandates

Advocacy is an important component of the CGIAR's Open Access policy. In 2004, the idea that research outputs should be treated as global public goods was introduced. Four years later, the Triple-A framework of availability, accessibility and applicability was introduced to encourage specific activities to communicate CGIAR knowledge to potential beneficiaries.

To create more awareness of Open Access to publications and related issues, a workshop was held at Bioversity in 2010. In the same year, a discussion about deposit mandates arose. Two centres (CIAT and ICRISAT) made their respective mandates public. Both statements require that scientists deposit their version of an article as soon as it is accepted by a journal. Neither statement includes a clause prohibiting publication in non-Open Access journals, but the CIAT statement requires scientists to consult the intellectual property rights officer before they transfer their copyrights.

In 2010, there was also a discussion about whether there should be a deposit mandate covering the entire CGIAR. This discussion was instigated by a letter sent by a number of science writers to CGIAR management.

The CGIAR has, however, recently undergone a reform process resulting in the establishment of a new legal entity, the Consortium of Agricultural Centers, supported by a Consortium Office. The Consortium was established to lead, coordinate and support centre research and cross-centre activities through the new CGIAR research programmes. While the centres may be developing individual Open Access policies, it is recognised that a system-wide strategy and supporting mechanisms would improve and speed up the Open Access process. The development of such a strategy falls under the Consortium's mandate and is included in its agenda as part of the Strategy Results Framework.[29]

[27] http://iwmi.catalog.cgiar.org/qryscr/catalogbs.htm.

[28] http://www.worldagroforestry.org/our_products/publications.

[29] http://consortium.cgiar.org/wp-content/uploads/2011/08/CGIAR-SRF-Feb_20_2011.pdf.

4 Open Access to data: overview of CGIAR data sets

4.1 Introduction

We do not have major publishers' systems to examine the position of Open Access for data across the sector. We can, however, look across the CGIAR as a data publisher at the various services provided. There is not an accepted measure for the "openness" of data sets like the "Romeo-Sherpa colours" that are commonly used to indicate how "open" a publication is. To test such a measure we have attempted to classify the data sets that are made available according to the 5 star scheme developed by Tim Berners Lee to assess the degree to which data is made available openly online:[30]

⋆ Make your data available on the web (any format)
⋆ ⋆ Make data available as structured data (e.g. Excel instead of image scan of a table)
⋆ ⋆ ⋆ Use a non-proprietary format (e.g. csv instead of Excel)
⋆ ⋆ ⋆ ⋆ Use URLs to identify things, so that people can point at your data
⋆ ⋆ ⋆ ⋆ ⋆ Link your data to other people's data to provide context

It is not straightforward to classify a system with one star as the data within that system may be varied, the services below have been grouped to reflect the majority of their data content. Examples range from data only available by email request which falls outside the star system through to linked data sets which are fully marked up to comply with linked data requirements.[31] An assessment based on current descriptions (Figure B.8) shows the breakdown of services offering data in the various star categories.

In our view, the classification exercise using the 5 star system goes some way to indicating the degree of opening access to data, as discussed below. Many systems in the study, however, appear to be outside the system or in more than one category. The 5 star system cannot be used alone as a measure of openness as it is designed to apply to the web and software retrieval rather than to the end user directly.

4.2 Outside the star system

No star Data available only on request (any format)

The system listed below shows the data available from centres only by mail, fax or email request.

[30] http://lab.linkeddata.deri.ie/2010/star-scheme-by-example.
[31] http://data.ifpri.org/rdf/ghi.

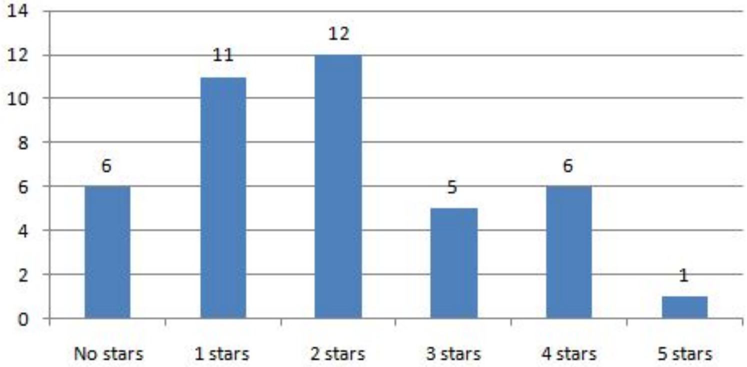


Figure B.8 Approximate number of services offering data in each star category on basis of descriptions

4.2.1 Center for International Forestry Research (CIFOR)[32]

Bogor Barat, Indonesia.

Data policy: access to the data below is possible only via mail, fax or email requests.

- **DOMAIN:**[33] the DOMAIN software uses Geographical Information System layers of environmental factors, such as climate, soil and land use, to construct an environmental habitat envelope or domain on the basis of points for the known distribution points of a species. The application then generates a map showing similarities across areas within the target region.
- **Criteria and Indicators:**[34] the Criteria and Indicators Toolbox Series comprises nine tools that were developed during the CIFOR project on Testing Criteria and Indicators for Sustainable Forest Management. The tools are aimed to help users develop and assess criteria and indicators of sustainable and equitable forest management.
- **FLORES:**[35] the Forest Land Oriented Resource Envisioning System is a model to help explore the consequences at landscape scale of policies and other initiatives intended to influence land use in tropical development. It seeks to provide an accessible platform to foster interdisciplinary collaboration between researchers and resource managers and to facilitate empirical tests of hypotheses and other propositions.

[32] http://www.cifor.cgiar.org

[33] http://www.cifor.cgiar.org/online-library/research-tools/domain.html.

[34] http://www.cifor.cgiar.org/acm/pub/toolbox.html.

[35] http://www.cifor.cgiar.org/online-library/research-tools/flores.html.

- **TROPIS:**[36] the Tree Growth and Permanent Plot Information System promotes more effective use of existing data and knowledge about tree growth in both planted and natural forests throughout the world.
- **VegClass:**[37] is a rapid, cost-effective method of surveying and classifying vegetation in forested landscape mosaics, developed using a minimum combination of variables including vegetation structure, plant species and plant functional types.

4.3 Data available online in publications

⋆ Make your data available on the web (any format)

All centres make their publications available online (see 3.2 Open Access repositories in the CGIAR). A number of data sets are represented within the documents without necessarily being available as separate data files. These collections date back to the 1970s; more recent reports will, of course, have data sets available separately as data files. Some examples are given below:

4.3.1 Africa Rice Center (West Africa Rice Development Association)[38]

Cotonou, Benin.

- **Technical Reports:**[39] which include some data on rice statistics and the genetic evaluation of rice in Africa.

4.3.2 Bioversity International (formally known as the IPGRI)[40]

Rome, Italy. English; documents also available in Chinese, French, Italian, Spanish, Portuguese, and Russian.

- **Bioversity International Publications:**[41] a web-based institutional repository providing access to publications that have been published or sponsored by Bioversity.

4.3.3 Center for International Forestry Research (CIFOR)[42]

Bogor Barat, Indonesia

[36] http://www.cifor.cgiar.org/online-library/research-tools/tropis.html.

[37] http://www.cifor.cgiar.org/online-library/research-tools/vegclass.html.

[38] http://www.warda.cgiar.org.

[39] http://www.warda.cgiar.org/warda/techreport.asp.

[40] http://www.bioversityinternational.org.

[41] http://www.bioversityinternational.org/publications.

[42] http://www.cifor.cgiar.org.

- **Publication repository**:[43] catalogue of CIFOR publications searchable by author, title, publication year, language and type of publication.

4.3.4 International Potato Center (CIP)[44]

Lima, Peru.

- **Publications:**[45] catalogue of CIP Publications (books, manuals, reports, working papers and training materials distributed for sale by CIP) on potato, sweet potato and Andean roots and tubers. Database searchable by author, title, publication year, keyword and language.

4.3.5 World Agroforestry Centre[46]

Nairobi, Kenya.

- **Publications:**[47] a wide range of the World Agroforestry Centre publications are searchable and available online.

4.3.6 International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)[48]

Andhra Pradesh, India.

- **Owned services:** AGROPEDIA,[49] ICRISAT Open Access repository,[50] ICRISAT institutional repository.[51]

4.3.7 International Livestock Research Institute (ILRI)[52]

Nairobi, Kenya.

- **Mahider:**[53] institutional repository of ILRI. Mahider is a complete index of research outputs produced by. Where available, the repository gives access to the full content of the output. The repository is built using Dspace; most of the outputs listed are Open Access.

[43] http://www.cifor.org/online-library/browse.html.
[44] http://www.cipotato.org.
[45] http://cip.catalog.cgiar.org/catalogs_menu.asp.
[46] http://www.worldagroforestrycentre.org.
[47] http://www.worldagroforestry.org/our_products/publications.
[48] http://www.icrisat.org.
[49] http://ring.ciard.net/agropedia.
[50] http://ring.ciard.net/icrisat-open-access-repository.
[51] http://ring.ciard.net/icrisat-institutional-repository.
[52] http://www.ilri.org.
[53] http://mahider.ilri.org.

4.3.8 International Rice Research Institute (IRRI)[54]

Los Banos, The Philippines.

- **Publications repository:**[55] owned services: DSpace at IRRI.[56]

4.3.9 WorldFish Center[57]

Penang, Malaysia. English and native languages.

- **Publications repository:**[58] covers WorldFish publications as well as works in other publications by WorldFish scientists and researchers

4.3.10 International Water Management Institute (IWMI)[59]

Colombo, Sri Lanka.

- **Publications repository:**[60] provides access to scientific documents published or jointly sponsored by IWMI. All IWMI publications are global public goods and are available for free from their online database.

4.4 Structured data sets

⋆ ⋆ Make data available as structured data (e.g. Excel instead of image scan of a table)

⋆ ⋆ ⋆ Use a non-proprietary format (e.g. csv instead of Excel)

Several of these services provide data of both types and so have been grouped together.
In many cases, the data is predominantly in proprietary forms as Excel, and statistical programs such as Stata and SPSS are used for the subsequent analysis.

4.4.1 International Center for Tropical Agriculture (CIAT)[61]

Cali, Columbia. English and Spanish.

Data policy: The plant genetic resources conserved by CIAT are a component of the world "designate collection" of the UN Food and Agriculture

[54] http://irri.org.

[55] http://dspace.irri.org:8080/dspace.

[56] http://ring.ciard.net/dspace-irri.

[57] http://www.worldfishcenter.org/wfcms/HQ/Default.aspx.

[58] http://www.worldfishcenter.org/wfcms/HQ/article.aspx?ID=118.

[59] http://www.iwmi.cgiar.org.

[60] http://www.iwmi.cgiar.org/Publications/index.aspx.

[61] http://www.ciat.cgiar.org.

Organization (FAO). Under a 1994 agreement with FAO, CIAT makes its germplasm available free of charge to farmers, farmer associations, breeders, agronomists, extension agencies, universities and Bioversity institutes with a clearly articulated need.

- **Database on plant genetic resources**
- **Product catalogue**
- **Online methods and query tools**

4.4.2 International Maize and Wheat Improvement Center (CIMMYT)[62]

Mexico City, Mexico. English and Spanish.

- **CIMMYT's Natural Resources Group (NRG) and Maize Program:** produced the *Africa Maize Research Atlas*, *Asia Maize Research Atlas*, and the *Latin American Maize Research Atlas*[63] a stand-alone, CD-ROM software that combines powerful and flexible GIS tools with preloaded data on climate, soils, elevation, population, land use and maize mega-environments for sub-Saharan Africa, southern Asia and Central and South America.
- **CIMMYT Socioeconomics Program (SEP):**[64] provides core data on agricultural prices and production through Open Access databases. Data are gathered from important global sources (World Bank, USDA, FAO, etc.) as well as from CIMMYT metadata projects. Information generated by CIMMYT includes descriptions of SEP projects and those from CIMMYT's former Economics Program and the Impacts, Targeting and Assessment Unit.

4.4.3 International Potato Center (CIP)[65]

Lima, Peru.

CIP databases:[66] are available online, which includes the following:

- **SINGER:**[67] (genetic resource collections) CGIAR genetic resources databases, including information on CIP's collection of potato, sweet potato and Andean root and tuber crops.

[62] http://www.cimmyt.org.

[63] http://www.cimmyt.org/en/services/geographic-information-systems/resources/maize-research-atlas.

[64] http://apps.cimmyt.org/agricdb/default.aspx.

[65] http://www.cipotato.org.

[66] http://cipotato.org/resources/databases.

[67] http://singer.cgiar.org.

- **World Potato Atlas**[68] and **World Sweetpotato Atlas:**[69] information about world potato production with emphasis on developing countries.
- **Inter-genebank Potato Database (IPD):**[70] a global database of potato germplasm available in the member gene banks.
- **World Potato Species Atlas:**[71] distribution maps of all currently recognised wild potato species.
- **DIVA GIS:**[72] tools (downloadable software), georeferenced databases and thematic maps related to genetic resource management.

4.4.4 International Center for Agricultural Research in the Dry Areas (ICARDA)[73]

Aleppo, Syrian Arab Republic.

- **Arid climate cereal and legume varieties:**[74] online data available on barley, bread and durum wheat, kabuli chickpea, lentil, faba bean, peas and forage legumes.

4.4.5 WorldFish Center[75]

Penang, Malaysia. English and native languages.

- **FishBase:**[76] online relational database with information on 28,500 species. Also available in CD-ROM format.
- **ReefBase:**[77] online free and easy access to data and information on the location, status, threats, monitoring and management of coral reef resources in over 100 countries and territories. Includes online GIS maps.
- **TrawlBase:**[78] a system for organising, storing, retrieving and exchanging a huge amount of data from past trawls in the seas of Asia.

[68] https://research.cip.cgiar.org/confluence/display/wpa/Home.

[69] https://research.cip.cgiar.org/confluence/display/wsa/Home.

[70] https://research2.cip.cgiar.org/confluence/setup/setupdbchoice-start.action.

[71] https://research.cip.cgiar.org/genebankdb/auto_2list.php?cmd=resetall&id=5.

[72] https://research.cip.cgiar.org/confluence/display/divagis/Home.

[73] http://www.icarda.cgiar.org.

[74] http://www.icarda.cgiar.org/Crops_Varieties.htm.

[75] http://www.worldfishcenter.org/wfcms/HQ/Default.aspx.

[76] http://www.fishbase.org/search.php.

[77] http://www.reefbase.org.

[78] http://www.worldfishcenter.org/trawl/index.asp.

4.4.6 World Agroforestry Centre[79]

Nairobi, Kenya.

- **Agroforestree Database:**[80] a species reference and selection guide for agroforestry trees.
- **Useful Tree Species in Africa:**[81] this tool enables users to select useful tree species for planting anywhere in Africa using Google Earth.
- **Botanic Nomenclature:**[82] a compilation of the taxonomic status of over 8000 woody and herbaceous taxa found in agroforest ecosystems including synonyms and common names.
- **Tree Slides Database:**[83] allows to search the collection of agroforestry images.
- **Tree Seed Suppliers Database:**[84] lists suppliers of seeds and microsymbionts for over 5939 tree species. Also available on CD-ROM and in a book version.

4.4.7 International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)[85]

Andhra Pradesh, India.

- **SAT Electronic Library:**[86] an online service to CGIAR's scientific community and the partners from National Agricultural Research Systems (NARS). The SAT Electronic Library consolidates various resources and services available both in-house and on the internet. The various sections are SATSource Database, SRLS Database, SCIRUS Search, SWETSWise Searcher, agricultural sites on the web and full-text publications.
- **infoSAT:** electronic repository of reprints collected and preserved through the project SATCRIS (Semi-Arid Tropical Crops Information Service). While the full-text access to documents in the repository is restricted to ICRISAT researchers/partners, the access to metadata is open to all.

[79] http://www.worldagroforestrycentre.org.

[80] http://www.worldagroforestry.org/resources/databases/agroforestree.

[81] http://www.worldagroforestry.org/our_products/databases/useful-tree-species-africa.

[82] http://www.worldagroforestry.org/Sites-old/TreeDBS/Botanic.asp.

[83] http://www.worldagroforestry.org/Sites-old/TreeDBS/slides.asp.

[84] http://www.worldagroforestry.org/Sites-old/TreeDBS/tssd/treessd.htm.

[85] http://www.icrisat.org.

[86] http://www.elibrary.icrisat.org/welcome.htm.

4.4.8 International Water Management Institute (IWMI)[87]

English; documents also available in German, French, Dutch, Norwegian, Spanish, Portuguese, Danish, Swedish and Russian.

- **Climate Atlas Web Query (CAWQuer):**[88] service creates online climate summaries for user-specified locations. User registration required.
- **World Water and Climate Atlas:**[89] gives irrigation and agricultural planners rapid access to accurate data on climate and moisture availability for agriculture.
- **Eco-Hydrological Database:**[90] focuses on management of specific information pertaining to various aspects of freshwater ecosystems' functioning, requirements and management. User registration required.
- **Water Data Portal:**[91] an integrated portal providing a one-stop access to all data stored in IWMI's archive.
- **Global Irrigated Area Mapping (GIAM)** and **Global Map on Rainfed Cropland Areas (GMRCA)**[92]
- **African Transboundary Water Law:**[93] contains a searchable database of more than 150 different treaties, amendments and protocols which have been signed to manage the use of Africa's transboundary waters.

4.4.9 International Food Policy Research Institute (IFPRI)[94]

Washington, DC, USA.

Data policy: data is available online per request. Requested data sets are for the use of the requestor only and cannot be used by others without the permission of IFPRI. Proper citation is required; citation information is included with the documentation of each data set. IFPRI encourages the use of these data sets, but emphasises that many of them contain "raw" data files.

- **IFPRI data sets:**[95] browsing through the data sets is available on household, community and institution-level surveys and social accounting matrices, geospatial data, agricultural investment and expenditure

[87] http://www.iwmi.cgiar.org.
[88] http://www.iwmi.cgiar.org/WAtlas/AtlasQuery.htm.
[89] http://www.iwmi.cgiar.org/WAtlas/Default.aspx.
[90] http://dw.iwmi.org/ehdb/wetland/index.asp.
[91] http://waterdata.iwmi.org.
[92] http://www.iwmigiam.org/info/main/index.asp.
[93] http://www.africanwaterlaw.org.
[94] http://www.ifpri.org.
[95] http://ifpri.catalog.cgiar.org/datasetquery.htm.

and regional data. Data are indexed and available online per request (request form).[96]

- **IFPRI Knowledge Products:**[97] research tools, best practices and services which IFPRI shares as international public goods.
- **Agricultural Science and Technology Indicators (ASTI):**[98] initially managed by the International Service for National Agricultural Research (ISNAR) which has since then been taken over by IFPRI. The ASTI time series database[99] provides access to agricultural research and development indicators for developing countries in tabular format.

4.4.10 International Institute of Tropical Agriculture (IITA)[100]

Ibadan, Nigeria.

- **IITA Catalogues and Databases Directory:**[101] access to publications related to IITA's research.
- **Statistical Database:**[102] access to authoritative agriculture websites based on research related to that of IITA.
- **Genetic Resources Center:**[103] holds plant material (germplasm) of major food crops of Africa. Distributed without restriction for use in research for food and agriculture.

4.4.11 International Livestock Research Institute (ILRI)[104]

Nairobi, Kenya.

- **Library database:**[105] presents all publications which have been published by ILRI staff members, ILRI's corporate documents and extracted records from CABI and AGRIS based on ILRI's mandate.
- **Research outputs:**[106] ILRI institutional repository (Mahider) contains metadata and the link to the full content on an increasing proportion of ILRI's research outputs.
- **GIS data and services:**[107] all spatial data layers generated by ILRI are searchable and downloadable.

[96] http://www.ifpri.org/data/dataform.htm.
[97] http://www.ifpri.org/knowledge-products.
[98] http://www.asti.cgiar.org.
[99] http://www.asti.cgiar.org/timeseries.aspx.
[100] http://www.iita.org.
[101] http://www.iita.org/catalogsanddatabases.
[102] http://www.iita.org/web/iita/statistical-databases.
[103] http://www.iita.org/genetic-resources-center.
[104] http://www.ilri.org.
[105] http://ilri.catalog.cgiar.org/ilribsrc.htm.
[106] http://www.ilri.org/ResearchOutputs.
[107] http://192.156.137.110/gis/default.asp.

4.4.12 Bioversity International[108](formally known as the IPGRI)

Rome, Italy. English; documents also available in Chinese, French, Italian, Spanish, Portuguese and Russian.

- **New World Fruits Database:**[109] provides easier access to some basic, but often difficult to obtain, information on fruits from the New World. Links are provided to additional information, such as experts working on the different species, references and URLs, making the database a useful starting point in a search for more information on the selected species.
- **Species Compendium Database:**[110] a searchable database providing information at taxon level about seed survival during storage, germination requirements and dormancy, reproductive biology, pests and diseases.

4.4.13 International Rice Research Institute (IRRI)[111]

Los Banos, The Philippines.

Data policy: nonexclusive, nontransferable, limited license is granted to view and use information retrieved from their website site, provided solely for personal, informational, noncommercial purposes and provided that copyright notice or other notices are not removed or obscured. In no event shall materials from the website be stored in any information storage and retrieval system without prior written permission from IRRI.

- **Rice Knowledge Bank:**[112] the first comprehensive, digital rice-production library containing an ever-increasing wealth of information on training and rice production.
- **Rice bibliography:**[113] online search for all rice and rice-related articles.

4.4.14 Africa Rice Center (West Africa Rice Development Association)[114]

Cotonou, Benin.

[108] http://www.bioversityinternational.org.

[109] http://www.bioversityinternational.org/databases/new_world_fruits_database/search.html.

[110] http://www.bioversityinternational.org/databases/species_compendium_database/index.html.

[111] http://irri.org.

[112] http://irri.org/knowledge/irri-training/knowledge-bank.

[113] http://ricelib.irri.cgiar.org:81/screens/opacmenu.html.

[114] http://www.warda.cgiar.org.

- **WAGIS:**[115] contains information on germplasm conserved in AfricaRice's gene bank, procedure to obtain seeds from AfricaRice using the Standard Material Transfer Agreement (SMTA) and many other data sets.
- **WAIVIS:**[116] The West Africa Inland Valley Information System contains databases on the agro-ecosystems of inland valleys in West Africa.
- **Additional data sets:** can be obtained upon request by email on:
  - breeding,
  - INGER Africa (varietal evaluation),
  - physiology, grain quality, drought, iron toxicity, photosynthesis,
  - soil fertility,
  - impact assessment.

4.5 Data portals

⋆ ⋆ ⋆ ⋆ Use URLs to identify things, so that people can point at your data

⋆ ⋆ ⋆ ⋆ ⋆ Link your data to other people's data to provide context

4.5.1 Global Bioversity Information Facility (GBIF)[117]

The GBIF makes digital biodiversity data openly and freely available on the internet for everyone and endorses both open source software and open data access. GBIF provides scientific biodiversity data for decision-making, research endeavours and public use. GBIF is a network of data publishers who retain ownership and control of the data they share. Linked data sets provide a more robust representation of biodiversity than any single data set. GBIF provides access to primary biodiversity data held in institutions in developed and developing countries. Data shared through GBIF are repatriated data. GBIF is a dynamic, growing partnership of countries, organisations, institutions and individuals working together to mobilise scientific biodiversity data.

Data use agreement The goals and principles of making biodiversity data openly and universally available have been defined in the Memorandum of Understanding on GBIF. The Participants who have signed the Memorandum of Understanding have expressed their willingness to make biodiversity data available through their nodes to foster scientific research development internationally and to support the public use of these data.

[115] http://africarice.org/wagis/default.asp.

[116] http://africarice.org/waivis/index.htm.

[117] http://data.gbif.org/welcome.htm.

GBIF data sharing should take place within a framework of due attribution. Therefore, using data available through the GBIF network requires agreeing with the following:

I Data use agreements
   (a) The quality and completeness of data cannot be guaranteed. Users employ these data at their own risk.
   (b) Users shall respect restrictions of access to sensitive data.
   (c) In order to make attribution of use for owners of the data possible, the identifier of ownership of data must be retained with every data record.
   (d) Users must publicly acknowledge, in conjunction with the use of the data, the data publishers whose biodiversity data they have used. Data publishers may require additional attribution of specific collections within their institution.
   (e) Users must comply with additional terms and conditions of use set by the data publisher. Where these exist they will be available through the metadata associated with the data.

II Citing data Use the following format to cite data retrieved from the GBIF network:
   Biodiversity occurrence data published by: (Accessed through GBIF Data Portal, data.gbif.org, YYYY-MM-DD)

III Definitions A series of definitions of data types and approaches are in the full agreement.[118]

4.5.2 Consortium for Spatial Information (CGIAR-CSI) data portal[119]

The CGIAR-CSI is a community of the many geospatial scientists within the CGIAR, linking the efforts of CGIAR scientists, national and international partners and others working to apply and advance geospatial science for international sustainable agriculture development, natural resource management, biodiversity conservation and poverty alleviation in developing countries. The CGIAR-CSI works to facilitate collaboration and capacity building for data sharing, data dissemination and geospatial analysis amongst the 15 CGIAR centers and the broader global research and development communities.

A number of data sets are made available by the CSI on climate, elevation, soil, poverty and others. A few samples:

– **WorldClim:**[120] a set of global climate layers (climate grids) with a spatial resolution of a square kilometer.

[118] http://data.gbif.org/terms.htm?forwardUrl=http3A2F2Fdata.gbif.org%2Fdatasets%2F.

[119] http://www.cgiar-csi.org/data.

[120] http://www.cgiar-csi.org/data/links-to-datasets/56-datasets/7-worldclim.

- **FutureClim:**[121] spatially downscaled climate projection data by Peter Jones (CIAT), Philip Thornton (ILRI) and Jens Heinke (PIK).
- **GADM:**[122] spatial database of the location of the world's administrative boundaries.

4.5.3 International Rice Research Institute (IRRI)[123]

- **World Rice Statistics (WRS):**[124] presents comprehensive time series information related to rice. Data on rice production, trade, consumption, inputs, prices and other related information are compiled from international and national statistical sources, personal communications and responses to questionnaires sent by IRRI's Social Sciences Division.
- **International Rice Information System:**[125] IRIS is the rice implementation of the International Crop Information System (ICIS) which is a database system that provides integrated management of global information on genetic resources and crop cultivars. This includes germplasm pedigrees, field evaluations, structural and functional genomic data (including links to external plant databases) and environmental (GIS) data.[126]

5 Challenges and opportunities

5.1 Challenges

Based on the review of data management within the CGIAR system in the previous sections, a number of challenges to providing data still remain and must be addressed.

Facilities and capacity The term "data curation" refers to a means of collecting, organising, validating and preserving data in such a way that researchers and stakeholders can make best use of the data over time. Data curation and data exchange facilities can be difficult – both technically and organisationally – to integrate into the workflows of research groups. For any system to be successful, however, it must focus on the users' needs, so these hurdles must be overcome and adjustments made.

[121] http://www.cgiar-csi.org/data/links-to-datasets/56-datasets/8-futureclim.
[122] http://www.cgiar-csi.org/data/links-to-datasets/56-datasets/9-gadm.
[123] http://irri.org.
[124] http://irri.org/world-rice-statistics.
[125] http://irri.org/knowledge/tools/international-rice-information-system.
[126] http://www.gosic.org/gtos/cgiar-data-access.htm.

To create a more widely used system of data curation, incentives (e.g. organisational credits or financial incentives) for data-curation efforts should be available to participating researchers. There should also be public recognition and commendation of those who openly publish and share data.

It is also important to keep in mind that, while data sharing is encouraged, there will always be instances where privacy is still warranted and, in fact, essential (e.g. the identities of survey participants must remain confidential).

Ownership and attribution One of the biggest challenges to making data publicly available is the management of intellectual property rights. The fear that there is a lack of control once data are released and a tendency to attribute sources incorrectly (or not at all) can often prevent data sharing entirely.

Similarly, there are complications when data-sharing systems include derived data sets and other components provided by third parties (e.g. satellite images or meteorological data). Derived data sets may be possible to share when source materials are not available.

In addition to clearly defining the property rights of a data-sharing system, an opportunity for those accessing the data to directly correspond with the researcher(s) who developed the data must be arranged. Since it is often not feasible to provide standalone data, there will be some need to set up correspondence with the original researcher(s).

Cost and duration The financial decision on expenditure in this area is complex. A cost–benefit analysis for data curation is not straightforward and a number of questions regarding what to keep and what to document must be answered. These financial decisions become increasingly complicated upon considering the finite nature of solutions to data-management problems. This brief lifespan is reduced even further because the data produced by CGIAR research changes rapidly. This means that work on updating data-management systems needs to be continuous, which is both time- and cost-consuming.

5.2 Opportunities

On the flipside of challenges, of course, are opportunities, and data sharing offers numerous opportunities to improve the way research findings are spread – and therefore used – throughout the world, which should motivate CGIAR researchers to participate in such activities. There are a variety of benefits to making research more available, accessible and applicable and researchers should be made aware of this.

Increased visibility Data curation and exchange can enhance the visibility of research and, thereby, the renown of its researcher. Thus, one of the benefits of data sharing is that the international community will come to associate a particular researcher (and his or her organisation) as an expert in a particular field. This incentive can boost interest in and resources for a particular project.

Improved access If data and information are more readily available and accessible to others in the field, it follows that studies will be more rigorously scrutinised and evaluated. This will result in more solid findings based on more reliable data.

Greater access to data can also improve partnerships and collaborative efforts by allowing all parties ownership of the material.

Greater impact By sharing data, information, findings and knowledge with more people – namely stakeholders – researchers and their organisations can make a greater contribution to the lives and livelihoods of the people they aim to serve. Often, drawing conclusions from a particular study and publishing within the academic and scientific communities does not translate into making a difference in the lives of farmers or families in developing countries. In order for research to be relevant, it needs to make its way to stakeholders in a format they can use. Data sharing and other interactive methods can help achieve this successful follow through.

6 Future directions and summary

The internet is an ever-evolving medium; as it changes, so too does the potential for its use as a research tool for and within developing countries.

6.1 Provision of data sets with publications

The demand for “the data behind the research” is strong and growing. Thus, there is a clear trend to systematically supply data sets alongside publications. While a certain lag-time still exists between document publication and data set release (in order to accommodate the need to adequately document the data sets), we are working to close that gap. But concerns still exist regarding “squeezing the last drop of usefulness” out of a data set before making it available and accessible and finding the resources to curate and maintain these data sets. Overall, however, the practice of including data sets with print publications serves to enrich the literature and enhance its value. It is important that we start considering data sets as an integral part of the global public goods we provide.

6.2 More work on interactive data visualisations

As software, server performance and bandwidths have improved, so has the ability to provide, within a website, the tools to allow a user to visualise data sets more readily and analyse a data set. The trailblazing work done by Gapminder – a nonprofit organisation that uses animated, interactive graphics to demonstrate statistical time series – has been adopted by some of the CGIAR centres that now use Google charts for visualisations and develop Flash-based visualisations to present data.

6.3 Publishing data in more interactive formats

Metadata is the foundation for information infrastructure and it is found throughout information technology systems: in service registries and repositories, web semantics, configuration management databases (CMDB), business service registries and application development. As the Semantic Web develops, more research organisations are producing linked data (e.g. linked data serialisations) that can be incorporated into linked applications.

In the past, the CGIAR had a strong commitment to metadata and indexed collections, so the challenge for the future is to open up these collections and integrate them in a way that allows their interoperability not only across the agricultural research community but also across the internet community as a whole. For this reason, linked data approaches, harvestable metadata and applications that use them both are clearly a necessity for the CGIAR in the future. We need to match this ambition with the skills sets and resources to deliver these systems.[127]

There are plans to expose data collections from the CGIAR as Linked Open Data and first experiments have been done. The Global Hunger Index (GHI) is available as Linked Open Data[128] and has been integrated into FAO's country profiles.[129] From this experience, a number of lessons have been learned that have been laid down in an online guide:[130]

[127] An example of an effort towards this is described in the blog post "Open Access Agriculture: opening the gates" at http://ictkm.cgiar.org/2010/10/27/open-access-agriculture-opening-the-gates

[128] http://data.ifpri.org/rdf/ghi.

[129] http://aims.fao.org/news/integrating-ifpris-linked-open-data-fao-country-profiles.

[130] http://linkedinfo.ikmemergent.net/content/global-hunger-index.

- Data that makes sense to the human reader is not necessarily sufficiently harmonised to be processed by other computer applications. A number of “cleaning steps” are required (steps 1, 2 and 3 in the online guide).
- Choose an appropriate data model and corresponding ontology for the data collection. The recipient applications that consume the Linked Open Data may have different requirements and the GHI is exposed using as ontologies two alternatives: “Scovo” and “Datacubes”. The important thing is that data provider (GHI) and data consumer (FAO country profile application) refer in the same way to the same things (steps 4 and 5 in the online guide).
- The Linked Open Data needs to be transferred to a server that can expose the data according to a protocol that the data consumer can handle. To encourage to take up the data it is essential to provide good documentation and examples how to use it (steps 6 and 7 in the online guide).

These steps are to degree technical but they require insight in the subject matter and in the way that others may want to use the data. These processes offer opportunities for researchers to communicate more closely with potential beneficiaries of their work beyond the circle of their direct peers.

6.4 Application programming interfaces

Data need to be machine readable so that they can be processed remotely by applications, combined and analysed online. This would be particularly advantageous for time series data in which an initial setup for reading information could be repeated when the data are refreshed.

6.5 Knowledge sharing

As the CGIAR moves ahead with its change process, it is continuously being told that it needs to do a better job at sharing its vast wealth of research-generated knowledge, so that this knowledge can be applied to solving real problems. While written publications – the major output of most projects – are a key source of high-quality information, they are not often widely available or accessible. In fact, for the majority of stakeholders working in agricultural development, they are not even applicable to them. Therefore we need to do a better job sharing our agricultural data, information and knowledge in ways that make them available, accessible and applicable.

Since we have already invested in knowledge sharing and innovative approaches to achieve it, the challenge becomes keeping the momentum going, which begs the question of how to maintain that energy. Universally, there are two types of motivators: positive reinforcement for a job well done and negative reinforcement for the opposite. The anecdotal story of dangling a carrot in front of donkey to entice it to move forward or using a stick to coerce him provides an interesting framework for understanding what motivates behavioural change. Based on the story, the carrot – an appealing treat for the donkey if it moves forward – represents incentives or rewards; the stick – an undesirable repercussion should the donkey refuse to move – represents mandates, policies, enforcements and punishments.

So, which works better when it comes to facilitating better knowledge sharing between researchers and research institutes? The carrot-like incentives or the stick-like mandates and enforcements?

The answer, at this point, is that most people are quick to respond to the question of how to share; they start throwing around resources, tools and methods that can help capture, store and provide access to knowledge. However, despite these active responses, many of these knowledge-sharing approaches are not actually being used widely or comprehensively. So perhaps we need to go back to some more fundamental questions: Why is knowledge sharing important in the first place? What does it aim to achieve?

Why should we share our knowledge? Is it not, after all, the vital capital of a researcher or research institution? Is it our role to share it? Do we have the capacity to do this sharing? What do we (and our institutes) gain by engaging in this time- and resource-consuming work? These are the remarks we often hear on the subject of knowledge sharing and Open Access. And the last question – What do we gain? – is particularly important. In order for researchers and research institutes to carry out knowledge sharing activities, there must be some benefit. This can range from greater visibility, improved fundraising potential, enhanced partnerships or better contribution to impact.

The carrot: incentives and rewards Much has been discussed on the subject of incentivising knowledge sharing; in fact, it has often been proposed that the CGIAR's performance evaluation mechanisms be redesigned to reward staff for going beyond publication requirements to get the research information out to various stakeholders. If researchers go the extra mile to organise workshops, build capacity or disseminate information via radio programmes, for example, how should they be recognised or compensated? In the CGIAR, we are continuously exploring, testing, documenting and celebrating new ways of sharing knowledge so that it can keep growing.

The stick: mandates or enforcements In some cases, organisations and institutions have developed policies and mandates to enforce certain actions amongst staff. These policies and mandates require staff to do particular things in a certain way and staff are evaluated accordingly and can be rewarded or punished if those activities are not carried out. This way of bringing about change has been found by some to be more effective in providing a wider stimulus for change. It also brings a certain consistency to the change desired. For example, CGIAR research staff are required to publish a certain number of journal articles per year, especially in ISI-ranked journals. There is also a reward system enforced by the overall CG system for these centres to increase their publication-per-researcher ratio. Which works better?

6.6 Promotion

The ASTI case study highlighted in this report (see 1.2 Case study) points to the importance of not only collecting and curating data, and subsequently making it available online, but also the necessity of promoting it through different channels. In particular, the study highlights the project's success in attracting media interest through events and press briefings. Therefore, in addition to using online communications, it is key to develop and use the variety of media available, such as print, visual design and face-to-face products, to reach targeted audiences.

While conventional methods for sharing data, information and knowledge, such as conferences, seminars, journal articles and reports, along with the more recent use of institutional repositories, play an important role in the communication of research and development, the way people search for information has been changing, especially in certain countries and amongst certain demographics. Social media has been growing in importance and steadily breaking down barriers to communication, allowing people to connect, engage and share in a more informal way. Social media tools currently used by the CGIAR include blogs, wikis and podcasts, and services such as Facebook, LinkedIn, Twitter and Flickr.

"Social media" has two main components: "social" and "media". Media tools can bring content to a much wider audience and at a much faster rate than previously thought possible. Since the way people search for information is evolving, the tools used must be adapted accordingly. Information overload is now a major concern, with many people not wanting to spend time visiting a website, blog, database or any other resource unless someone they trust points them in that direction. Through social media channels, it is possible to seek out recommendations and suggestions from colleagues, peers and experts.

This has implications for the way research and development organisations communicate. While opening access to information is widely regarded as important, simply pushing it out to target audiences does not guarantee that it will be read and used. Information is useful only when it is received and read by the right person at the right time. Social media tools can help get an organisation's messages to the right people.

As new media tools are based on being social, users can build a community around their content or particular channel, and this is where social media's true value lies. These communities, or social networks as they are called, are formed around a common interest or shared purpose, and often result in an environment based on trust that facilitates effective collaboration and sharing. This enables a natural flow of peer review and feedback and also enhances transparency. But it is imperative that social media also be used for much more than forming communities or reaching out to those with shared interests. Social media needs to be explored, understood and harnessed to make knowledge available, accessible and applicable to the wide array of audiences participating in the social media arena.

Social media allows for and, in many cases, insists on much more continuous and less formal communication activities. The communication of research usually takes place at a project's conclusion, but social media can facilitate the sharing of ideas, experiences and knowledge throughout the whole research cycle. Social media can give access to the inner workings of research activities, enabling a variety of audiences to learn from them. For example, audiences can read about a project's progress on a project blog, see actual evidence of research activities in the form of Flickr pictures and hear and see testimonials in YouTube videos. Having access to the knowledge being generated throughout the research cycle is not only important to donors, but it can also benefit other target audiences of the research.

Moreover, social media allows audiences to participate and provide almost real-time feedback. By considering an audience's needs and questions, it is possible to keep the research grounded and facilitate better partnerships (as opposed to just having research recipients), both of which enhance the sharing of knowledge, in terms of its use and impact. As such, social media can increase stakeholder involvement and enhance a project's accountability and the overall achievement of its objectives.

Social media can help organisations to reach and involve a wide range of necessary stakeholders such as donors, other research communities, implementing partners, the general public and even the intended beneficiaries, such as farmers. Since social media tools are freely available and internet access is becoming more available throughout developing countries, even in rural areas, research knowledge is now being more readily picked up and demanded directly by farmers. For example, three posts on the new CGIAR Consortium blog received comments from farmers located in rural areas of developing countries asking for more information about getting access to the seeds that one of the posts talked about, among other things. Social media channels can also play a brokering role by connecting people to others who have the knowledge they need, such as linking farmers with extension agents.

Social media, therefore, has the potential to make research outputs much more available, accessible and applicable. It can increase the visibility of, participation in and adaptation of research knowledge. Compatibility among different social media tools provides an added dimension of connectivity so that social networks can be interlinked, creating an audience base that has the potential to expand rapidly.

6.7 Collaborative efforts

The CGIAR system, like many organisations, is adept at generating information. However, the challenge is in knowing how to extract and manage the knowledge buried within the volume of information being produced and then being able to apply that knowledge to emerging needs.

Knowledge sharing is a way of putting information, communities, processes, and tools together to allow the CG to collaborate more effectively and make better decisions. Tools and technologies by themselves cannot ensure successful partnerships, collaboration or teamwork, nor make the CGIAR work as a system; they are necessary but not sufficient. And, while the tools and technologies can contribute to improvements in personal and organisational performance, significant gains require changes in organisational culture and individual behaviour.

People, and the tacit knowledge they have, are central. It is through greater understanding and support to the cross-functional communities that organisational culture can shift towards one of ongoing learning and collaborative sharing of knowledge and expertise – a CGIAR without boundaries. It is now realised that the best knowledge-transfer technique is face-to-face interaction and that the best knowledge repository is a community or group of people, supported by a technology solution.

The CGIAR is committed to strengthening incentive systems that promote knowledge-sharing practices and for communities of scientists to improve the way they work.

6.8 Ubiquitous telecommunications infrastructure

Thanks to the falling costs of all things digital, there has been a steady flow of investment into communications infrastructure around the world. Cell phone networks carrying voice and internet data are being deployed in even the poorest countries and with time will expand to cover most rural areas. These wireless networks are sophisticated and easily managed. Multipurpose public networks will be offered by private telecomm companies and governmental agencies, while self-organising device networks (such as ZigBee, a low-cost, low-power, wireless mesh networking standard) can be installed with minimal planning or oversight. Agriculture and agricultural research can increasingly take communications capacity for granted in the years ahead.

This new infrastructure will enable new applications of communications to both the gathering and dissemination of information by agricultural researchers and practitioners. First, for gathering information, the historical and remotely sensed data that has been gathered to date can be complemented by near-real time, ground-based data. Sensors can transmit the information they detect through increasingly ubiquitous wireless data networks into internet-based servers. Radio-frequency identification (RFID) tags can be attached to vehicles, buildings and selected goods; combined with Geographic Positioning System (GPS) information, objects can be automatically tracked and even audited in real time. The result will be both real-time interpretation of current conditions and longitudinal analyses that reflect up-to-date information.

The costs of the sensors, tags, GPS and RFID devices and the communications between them are dropping so rapidly that new data-gathering applications can be expected to proliferate in the near term. Here are some relevant examples:

- sensors and cameras in fields or on farm equipment,
- sensors of water levels in irrigation or in soils,
- sensors in food storage,
- early detection of pests,
- emissions sensors,
- tagging of livestock,
- tagging of other natural resources,
- tagging trucks and shipping containers,
- market, banking and distribution data.

Like satellite imagery, these new types of data will require considerable processing to ensure their quality and consistency and to make them comparable from one location to the next. The research community will need to establish processes for validation and distribution of these data, as they have with other public information goods.

The same networks that collect and carry sensor data will also be used to disseminate information into rural areas. Cell phones are already being used at an increasing rate by rural residents. For them, the value of communication is high, and there are many ways to effectively share the fixed costs of phone devices and electric power among numbers of users. As phones get larger screens, touch interfaces and voice recognition, and as new classes of inexpensive and rugged “netbooks” are developed, many new opportunities for agricultural extension will arise. It will start by providing today's information to new audiences. It will grow into provision of new services that are more localised and more up-to-date, building on the data gathering that is enabled by the networks that feed these devices. With new audiences and new services will come new requirements for assuring the quality of the information provided.

6.9 Cloud computing

The combination of progress in system software, computing hardware and internet communications has now enabled the construction of general-purpose data centres that can be reconfigured by command to support any software application in minutes ("virtualisation" software was the key innovation).

There are already data services that allow a user to have many hundreds of computers at their command, and yet pay for them by the hour or minute, without owning or operating the hardware themselves. The costs are far less than even falling hardware prices would suggest, since the cost of the data centre can be shared among many "bursty" users. In effect, the data centre acts like a utility, providing as much computing as requested at just the times when needed. Since these data centres are invariably shared over the internet, they are sometimes called computing "in the Cloud," giving rise to the common term "Cloud computing".

Many observers believe that Cloud computing will soon be the lowest-cost option for nearly all types of data centre computing. Cloud providers are already more cost-effective for "bursty" high-performance computing, like video and image processing, bioinformatics and most types of scientific data analysis. We can expect research centres in agriculture to have accounts on several Cloud providers and to select them at different times for different purposes.

The shift to Cloud computing is a good thing for today's researchers, by cutting the total cost of scientific computing. But it also brings two new opportunities for international agriculture. First, it completely separates the utilisation from the operation of computing facilities. In other words, users of data centres no longer need the capacity to procure and operate them. As long as one has a browser on the internet, one can "order up" essentially any computer software at any scale and pay only for what is used. As a result, many more organisations will be able to take advantage of large-scale advanced computing.

A second implication of Cloud computing is an increased impetus to share data among researchers. It is a common pattern today to move large data sets, such as satellite images or longitudinal data sets, from one data centre to another for use in different projects. The transfers add delay and can be error prone. By contrast, Cloud data centres are a natural repository for public information goods such as shared data sets, so that users in any location or institution can instantly access, analyse and interpret data without the need to move it to their own facilities. This reduces the need for high-speed or high-capacity network connections, since much less data moves between the users and the source of the data. A researcher with a moderate-speed connection to the internet can work with data as well as other researchers regardless of location. In addition, researchers will normally leave the results of a Cloud analysis at the Cloud data centre, allowing potential re-use by others. Properly managed, this can enable new kinds of collaboration and project organisation.

Implications Leading institutions in agricultural research have an opportunity to flesh out these possibilities today and thereby create templates for future models of progress. Here is one illustration.

A research centre that works with a crop could choose a group of similar varieties that have been cultivated for a long period at one of their facilities. The centre will already have basic long-term data across many seasons, along with much bioinformatic data. These data could be stored in one or more Cloud computing facilities and could be supplemented from now on by extensive sensor data, collected and made available in near real-time from the fields where the varieties are cultivated. In effect, these fields become a “bio-observatory” for those varieties. In addition, one or more regions where those varieties are currently cultivated by farmers could also be instrumented with some sensors, and the markets in those regions could employ tags or other methods for continuous data collection. Once a data collection like this is available in a Cloud data centre, a series of analytical studies could be commissioned at various developing country institutions around the world.

These institutions would be chosen for having familiarity with the varieties but currently lacking the facilities to do their own extensive data analysis or interpretation. In addition, adaptation and extension projects could be commissioned at additional national organisations to produce materials for delivery into the areas where these varieties are grown.

Like any collaborative research, this kind of project would have to confront issues of data harmonisation, accessibility and ownership. Part of the value of this project would be the demonstration of solutions to these issues, as a pattern for future projects to follow.

Naturally, this entire scenario could be adapted in many ways to the other agricultural research topics. For example, a project could treat livestock instead of crops or could extend a system like Fishbase[131] for a class of fish. Genetic studies of crop pathogens, patterns of water supply and utilisation in a watershed, forest growth and production patterns: all lend themselves to this sort of project. There will be limitations to the effectiveness of any single project, but the first projects are likely to provide key lessons to light the way for the research community in utilising the next wave of technological changes.

6.10 Increased use of spatial analysis and GIS

As more data sets become available, the opportunities to compare agricultural data across similar agro ecological zones leads to an increased need to accurately geo-locate data. This has lead to an increasing use of geographical information systems and spatial analysis. One example of this is presented in the case studies. We see this trend developing in the future with increased interest in spatial analysis and the development of new models and tools in this area. Agriculture is inextricably tied to the physical environment and the unpredictability of nature. Factors such as climate, soil and water availability play more of a defining role in agriculture than in any other economic sector. And nowhere is this more evident than in Africa. If smallholder farmers are to be consistently successful, from one season to another, and from one year to another, they need to have access to essential geospatial (location-specific) information.

[131] An online database with information on 28,500 species developed by the WorldFish Center http://www.fishbase.org.

7 List of figures

8 List of tables

9 Bibliography

ASTI. *ASTI Toolkit. Monitoring Agricultural R&D. Capacity and Investment Indicators: A Practitioner's Guide to ASTI's Methodologies & Data Collection Standards.* Washington, DC, International Food Policy Research Institute, 2011.

Alercia, A, Diulgheroff, S, & Metz, T. FAO/IPGRI Multi-crop Passport Descriptors. Rome, Bioversity, 2001. http://www.bioversityinternational.org/fileadmin/bioversity/publications/pdfs/124.pdf?cache=1318460131.

Arivananthan, M, Ballantyne, P, & Porcari, E. Benchmarking CGIAR outputs for availability and accessibility. Agricultural Information Worldwide 2010, 3(1), 17–22. http://journals.sfu.ca/iaald/index.php/aginfo/article/view/35/51.

Norton, G. W. (2010). Impact Assessment of the IFPRI Agricultural Science and Technology Indicators (ASTI) Project. Washington, DC, International Food Policy Research Institute, 2010. http://www.ifpri.org/sites/default/files/publications/ia32.pdf.

Vargas, R. (2010). Breaking the Norm: An Empirical Investigation into the Unravelling of Good Behaviour (no. 00948). Washington, DC, International Food Policy Research Institute, 2010. http://www.ifpri.org/publication/breaking-norm.

Vargas, R, & Viceisza, A. An Experiment on the Impact of Weather Shocks and Insurance on Risky Investment (no. 01100). Washington, DC, International Food Policy Research Institute, 2010. http://www.ifpri.org/publication/experiment-impact-weather-shocks-and-insurance-risky-investment.

10 Further reading

CIAGR blog The CGIAR has promoted a number of knowledge-sharing approaches through a series of blog articles, which it sees as a useful vehicle to continue these developments in the future.

Create awareness by raising the profile of your organisation on social networking sites. Cultivate long-term support for your organisation by creating your own network of scientists, research partners and interested individuals. http://ictkm.wordpress.com/2009/05/06/social-networks-friend-or-foe. Use social media tools to promote your projects, events and activities. Announce time-sensitive, newsworthy items by microblogging. Microblogging involves posting short sentences (max. 140 characters) that can be used to promote your journal article or a useful website, act as a reminder for an activity or even ask questions. http://ictkm.wordpress.com/2009/04/02/microblogging. Promote your name. Use social media to establish your reputation in the research and development arena. Blogging is a good way for researchers to share their research ideas with others and gain feedback from a wider, online audience. Well-thought-out blogs attract people with similar thoughts and queries, people who can validate your ideas and also challenge you by sharing varying opinions. http://ictkm.wordpress.

com/2009/04/23/blogging-for-impact. There are many ways you can engage with others and share knowledge using social networking sites.

Engaging people

- Promote issues that resonate with people to encourage involvement and gather support for your cause.
- Form strategic alliances with influential people and institutions that help boost your organisation's profile.
- Bring together expertise and talent, whether potential research partners, service providers or other experts.

Sharine knowledge Social media transcends geographic boundaries. Test your research ideas by sharing them with your colleagues globally. You can collaborate at a fraction of the time and cost associated with face-to-face meetings. Collaborative sharing sites also come with security options that allow secure knowledge sharing. http://ictkm.cgiar.org/2009/05/29/wikis-sites-docs-and-pads-the-many-flavours-of-collaborative-writing. Create an environment where people recognise your expertise and you can establish your organisation as the expert in your field of research. Whether you are a researcher who is new to a field and eager to learn more, or the resident expert, share your knowledge and experiences by contributing to insightful blogs.

- Communicate your research outputs better by adjusting your content to fit different social media tools. Think of social media as strategic communication lines that branch outward to several different networks, which in turn branch into other networks.
- Reach out to interested people outside your regular circle and gain valuable ideas/feedback from your pool of social networks. Pay attention to conversations that are already ongoing on social media sites. Sharing is a two-way process, and you should take the time to interact with others in a similar fashion.

Share resources within interested communities. Social bookmarks and news feeds are great online organisation tools that keep track of what's being published on useful websites and blogs you frequent. Share this with others and then see the favour being returned. http://ictkm.wordpress.com/2009/05/18/social-bookmarking-storm-brewing. http://ictkm.wordpress.com/2009/06/19/newsfeeds-delivering-the-latest-news-to-your-virtual-doorstep.

11 Glossary

Cloud computing

A shared Cloud data centre will typically have over 1000 computers, which can support at least 100,000 user “virtual” computers. This is super-computer scale by any standard, so most research centres will not own one but rather will share one with hundreds of other customers. Commercial Cloud providers, such as Amazon, Google and Microsoft, already offer services and some government-run research Clouds exist. Shared by many thousands of customers, these are extremely cost-efficient. They employ a relatively small staff of system managers, keep a low budget for electric power, can survive routine equipment failure without service interruption and adopt continuous modular upgrades of new types of hardware. There are choices in many countries, which allow for flexibility where there are legal restrictions.

Social media

The following definition is given by Wikipedia (http://en.wikipedia.org/wiki/Social_media): “Social media is online content created by people using highly accessible and scalable publishing technologies. Social media is a shift in how people discover, read and share news, information and content; it supports the human need for social interaction with technology, transforming broadcast media monologues (one to many) into social media dialogues (many to many). It supports the democratisation of knowledge and information, transforming people from content consumers into content producers. Social media has become extremely popular because it allows people to connect in the online world to form relationships for personal, political and business use.”

C Information and Communication Technology

Dennis Spohr and Philipp Cimiano

This chapter describes the case study carried out at the Centre of Excellence Cognitive Interaction Technology (CITEC) at Universität Bielefeld. The aim is to provide a representative example of a research institution in the wider field of information and communications technology (ICT), with a specific focus on cognitive interaction and robotics engineering. After a brief introduction to the general structure and mission of CITEC, we will discuss the general scope of the case study, as well as how the methods presented in the introduction to this book have been applied in detail.

1 History, structure and mission

CITEC is a research institution founded at Universität Bielefeld as part of the Excellence Initiative of the German federal government and the state governments in 2007. According to its statutes,[1] CITEC is a competence centre for fundamental research and technology transfer and cultivates an international network of cooperation with industrial and scientific institutions. This includes industrial partners like Miele & Cie KG and Honda Research Institute Europe GmbH, as well as members from internal and external institutions, such as the Research Institute for Cognition and Robotics (CoR-Lab) and the Centre for Interdisciplinary Research (ZIF) at Universität Bielefeld. The network of external researchers is integrated into the so-called *Virtual Faculty*, which includes renowned experts in research fields related to cognitive interaction technology from all over the globe. Finally, CITEC maintains an international and multidisciplinary graduate school offering scholarships for PhD students.

[1] http://www.cit-ec.de/sites/www.cit-ec.de/files/CITEC_Satzung_0.pdf.

CITEC consists of 37 research groups[2] – 11 of which newly funded with financial support from the excellence cluster, 24 that were part of different departments of the university before the foundation of CITEC and two senior professorships – comprising overall more than 250 scientific staff. The groups come from a variety of scientific backgrounds, such as neurobiology, linguistics or computer science. What all of these groups share as part of their mission within CITEC, however, is the common goal to obtain a better understanding of cognitive interaction, as well as its implementation in technical systems. Within CITEC, they are organised in four major research areas, namely:

1. **Motion intelligence:** This area investigates how perception and action can be combined in a way that allows robots to operate autonomously in unpredictable environments and situations. This is approached by investigating animals and humans performing different cognitive tasks, from various perspectives (such as biological, psychological or physical), in order to arrive at a comparable level of sensorimotor capabilities in robotic systems.
2. **Attentive systems:** The primary aim of this area is to combine experimental and empirical methods as well as engineering approaches in order to identify the mechanisms that enable artificial systems to understand and actively focus on what is important and align their processing resources with their human partner accordingly.
3. **Situated communication:** This research area focuses on how language, perception and action can be coordinated in a way that enables efficient cooperation between humans and technical systems. As such, this involves research of linguistic and psychological phenomena in communication, as well as the computational aspects of their implementation in artificial systems.
4. **Memory and learning:** The focus of this area is to find technical solutions to cognitive issues like memorising in order to arrive at architectures which enable an artificial system to acquire, store and retrieve knowledge, as well as to improve its capabilities by learning. This is achieved by combining experimental research into biological brains with the development of new algorithms for learning and memorising knowledge.

As can be derived from the above description, each of these research areas combines aspects from engineering with aspects of other scientific disciplines, such as life sciences and computer sciences, indicating a high degree of in-

[2] Being a very young institution, the number of research groups at CITEC is still constantly changing. At the time this study was carried out, it consisted of only 32 research groups.

terdisciplinary cooperation. In fact, the CITEC website[3] states the overall vision and goals as follows:

*The vision of the CITEC scientists are interactive tools that can be operated easily and intuitively, ranging from everyday objects to fully-blown humanoid robots. The future technology should adapt itself to its human users instead of forcing us humans to adjust to the often cumbersome operation of the current equipment. Just as every human being automatically adapts his speech and actions to the addressee in order to be understood, technological systems should adjust their behaviour to their interaction partner. In order to interact naturally with humans and to flexibly adapt to changing conditions, a system needs to be endowed with the corresponding cognitive abilities. Consequently, the study of the fundamental architectural principles of cognitive interaction – be it between humans or human–machine interaction – is the necessary pioneering work. It is supplemented by new possibilities of technological application, which need to be designed, constructed, and tested. We believe that this dual goal of combining basic research with technological application in order to significantly advance our understanding of cognition itself can only be realized through **intense interdisciplinary cooperations**.*

As the previous paragraphs have shown, CITEC is a very young research institution that is heterogeneous along several dimensions. On the one hand, it comprises newly created research groups as well as groups that existed at Universität Bielefeld long before CITEC was founded – some of which that had not been in cooperations before. On the other hand, cognitive interaction technology is in itself a very heterogeneous field of research, requiring a high degree of interaction between researchers from various disciplines. Moreover, many research questions require empirical and experimental insights into biological and behavioural aspects of cognition in animals and humans (e.g. when interacting with artificial devices) as well as investigations from engineering and computational perspectives. Finally, CITEC is very internationally networked, including academic as well as industrial partners all over the globe (such as Finland, Ireland, Japan, Spain and the USA).

This heterogeneity poses high demands both on the infrastructure that needs to be in place in order to support such interconnectivity, as well as on the specification of common policies for the management and exchange of both data and literature, which may differ considerably between disciplines, due to very different traditions and historical backgrounds, as well as between academic and industrial partners, in particular with respect to legal issues. These aspects will be discussed later in this chapter, after an introduction to the methodology and the case narratives.

[3] http://www.cit-ec.de.

2 Methodology

In this section, we specify how the methods explained in the introduction to this book have been applied in this case study, as well as the scope of each method in terms of the number of research groups which participated. As was further mentioned in the introduction to this book, this case study had been preceded by a preliminary study on a small subset of research groups, covering, however, all of the research branches introduced above (see 1 History, structure and mission). This study put different methods to the test and thus helped to estimate the qualitative and quantitative impact of each of the methods applied. In addition to this, the results of each method were then integrated into the actual case study itself, thereby obtaining a detailed and representative picture of research infrastructure, literature and data management at CITEC. The methods that have been applied in this case study are described below.

2.1 Introductory interview

Interviews were held with the leader(s) of a research group in order to get a general understanding of the research topics dealt with in a group and to determine the applicability of the other methods. These interviews, which lasted between 10 and 40 minutes each, were first recorded on audio and later protocoled and analysed.

2.2 Observation

Observations of experiments were carried out as part of the typical research agenda of a group (i.e. the experiments were scheduled independently of the observation). An experiment was observed only if it was ensured that the observation would not interfere with the workflow of the experiment. Each observation lasted between 45 and 60 minutes and was recorded on audio and video and later protocoled and analysed.

2.3 Questionnaire

Based on the introductory interview and observation, a questionnaire was developed in cooperation with the Universitätsbibliothek Bielefeld, containing questions believed to cover the most important aspects of research infrastructure. This questionnaire was used to guide the semi-structured interview and was circulated among all research groups which had not participated in a semi-structured interview.

2.4 Semi-structured interview

The questionnaire was used to guide semi-structured interviews with a selection of groups and lasted between 30 and 60 minutes. As with the introductory interview, the semi-structured interview was audio recorded, protocoled and analysed.

2.5 Website analysis of publication behaviour

The above methods were complemented by an empirical analysis of a selection of group websites in order to investigate publication behaviour. If such websites were available, the publication section of a group's website was inspected by applying the heuristics in Figure C.1 to each of the publications listed there.

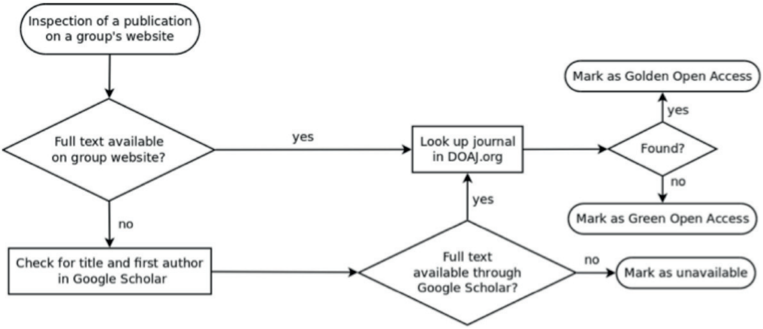

Inspection of a publication on a group's website
Full text available on group website?
yes
no
Look up journal in DOAJ.org
Found?
yes
Mark as Golden Open Access
no
Mark as Green Open Access
Check for title and first author in Google Scholar
Full text available through Google Scholar?
yes
no
Mark as unavailable

Figure C.1 Flowchart of the literature analysis process

The analysis was limited to the first 100 publications listed on the website, starting with the latest ones. In addition, some pre-processing was applied before consulting Google Scholar (e.g. in case the bibliographic information on the website contained typographical errors). More often than not, several variations of titles were queried (e.g. subphrases enclosed within double quotes), as Google Scholar frequently returned inaccurate results for titles containing hyphens. Moreover, because groups frequently publish articles in cooperation, the chance of counting a particular publication more than once is quite high. In order to exclude such cases, at least to some extent, the analysis was carried out without emptying the cache memory of the web browser. In other words, when analysing a particular publication, it was immediately visible to the investigator whether the publication had been accessed at a previous point in time. If this was the case, it was not counted a second time. Finally, it should be noted that this analysis did not consider the university-wide publication repository PUB available at Universität Bielefeld because it was still under development at the time the analysis had been carried out and was still in a transition state at the time of writing.

This analysis included only groups that have a group website on which they provide bibliographic information about their publications, since an exhaustive investigation – which would have required identifying the researchers of a particular group and then navigating to their (potentially external) homepages in order to apply the aforementioned process – did not seem feasible. Therefore, in order to cover such cases as well, the questionnaire contained a series of questions dealing with literature management and publication behaviour, in particular with respect to Open Access. Moreover, as will be discussed below (see 3.4 Robotics and engineering (RobEng) and 6.2 Results of empirical website analysis), a representative collection of the aforementioned branches is covered by this analysis.

3 Case narratives

In order to be able to interpret the results of this case study beyond the level of individual groups, some form of classification is needed. The research areas of CITEC presented above, however, are not suitable for such a classification, as they do not partition the set of research groups into disjoint sets. As a result, it would not be feasible to attribute the findings within a particular group to one particular research area. In addition to this, CITEC is a highly interdisciplinary research centre and even the boundaries between groups are not always clear-cut. For instance, some researchers are affiliated with more than one research group. Moreover, since almost all groups have a cognitive and a computational component, a strict classification in terms of "traditional" scientific disciplines can at best be approximated.

In order to arrive at a meaningful classification nevertheless, we present below four case narratives that try to approximate a classification of groups in terms of traditional scientific disciplines. Each case narrative gives a brief description of the groups associated with the respective disciplines, as well as the common practices observed with respect to research data management, literature management and research workflows. However, we restrict the discussion to those groups which participated in the study.

3.1 Behavioural sciences, natural sciences and neuroscience (BehNatNeur)

Research group profiles Members of this branch of research either belong to one of the traditional natural sciences (e.g. biology or physics) or are characterised by a high degree of experimental work involving living beings like humans or animals in order to investigate neural or psychological aspects of cognition. In Table C.1 and the following, we give brief descriptions of each of the groups which we classify as belonging to this research branch, along with a brief description of their approximate size, main research objectives and primary research instruments.

Table C.1 Flowchart of the literature analysis process

Name	Research topics	Members	Computers PCs/ servers	Further instruments	Cooperations internal/ external
Active Sensing	Sensory capacities of electrical fish, neural mechanisms of parallel processing, as well as hydrodynamics	4	7/1		1/3
Biological Cybernetics	Sensory control of behaviour, esp. motion sequences	15	15/1 (and 10 set-up PCs)	Three motion capture set-ups (one Vicon MX10 with eight cameras and two custom-made ones with three Basler-A602 cameras each, objectives and ringflash system)	3/3
Neurobiology	Visual information processing in the brains of flying insects	19		Flymax	
Neurocognition & Action Biomechanics	Human perception and sensomotorics, cognitive representation and motion intelligence in humans and robots, cognitive biomechanics and augmented reality, neuromotion and neurosimulation	25	35/0	Virtual and augmented reality, motion tracking, electroencephalography, electromyography	5/15

Neurocognitive Psychology	Attentional control of visual perception and of sensori-motor actions	9		Motion capture set-ups, eye-tracking	
Physiological Psychology	Memory and memory deficits in humans	11	16/0		3/5

Empty table cells indicate that no information has been provided.

Research data Table C.2 shows the different types of data created by a typical series of experiments on a specific research topic in the behavioural sciences, natural sciences and neuroscience. As can be seen, video data especially arise in large quantities and sizes (more than 1000 files and in the terabyte range), followed by "other types of data". Here, groups indicated mainly Vicon[4] data files, electrophysiological measurement files, electroencephalographic data and spectra, as well as other binary data.

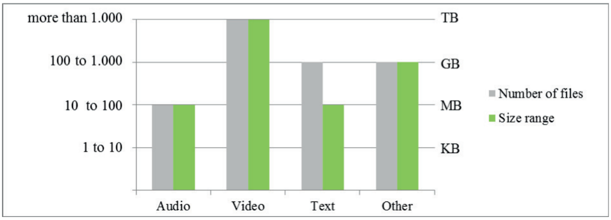


Figure C.2 Data types in terms of number of files and sizes in BehNatNeur

Three out of four groups further indicated that there are established standard formats in their field and that they use them either very frequently or with rare exceptions. With respect to metadata, two groups indicated that they annotate their data with metadata, with the other two stating that they do not.

Research data lifecycle Groups indicated the following stages in the data lifecycle, with bold stages being those shared by at least half of the groups.

1 **Data collection**
All groups stated that they begin with the collection of data in experiments, with two groups indicating that the data collection process may take up to several months.

[4] Vicon is a widely used motion capture system; see http://www.vicon.com.

2 Archiving
Only one group mentioned an archiving step of primary data on a file server accessible to several institutes (Bioserver).

3 Processing
All groups indicated that they process the primary data, either by means of manual analysis in statistics programs like Microsoft Excel or SPSS or computationally using Matlab, for example, or performing video and cluster analyses.

4 Archiving
Three groups indicated a backup step involving DVDs, external hard drives, file servers or individual PCs.

5 Re-use/enrichment
Two groups indicated that the data are re-used at a later stage, for example in the context of new studies, or processed and analysed further.

6 Archiving
One group indicated a final archiving step of the analysed data, again on the Bioserver and individual PCs.

Research data and Open Access As was mentioned at the beginning of this section, research in behavioural sciences, natural sciences and neuroscience involves a considerable amount of data gathered in experiments with humans. As a result, primary data are expected to be problematic with respect to the application of Open Access principles. This is supported by the results of the questionnaire, which shows that only secondary data could conceivably be shared with the public, albeit to a limited extent (Table C.3). For primary data, one group indicated that they would not even share data with close colleagues.

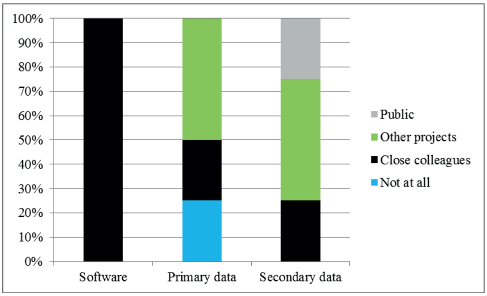


Figure C.3 Willingness to share software and primary and secondary data in BehNatNeur

Research literature Research groups in behavioural sciences, natural sciences and neuroscience at CITEC primarily use Endnote, Mendeley and Reference Manager as well as the university-wide publication repository PUB[5] for managing their scientific literature. With respect to publication preferences, groups stated that both print and electronic publication media are preferred and established in the field.

Literature and Open Access Three out of four groups indicated that Open Access is hardly established in their field, and one group even indicated that Open Access is not established at all. This is supported by the empirical website analysis, which revealed that, of the 231 publications analysed, only 16 (6.93%) were Golden Open Access publications and 79 (34.20%) were Green Open Access publications. The remaining 136 publications were unavailable following the strategy explained above (see 2.5 Website analysis of publication behaviour).

Table C.2 Research groups in behavioural sciences, natural sciences and neuroscience

Name	Research topics	Members	Computers PCs/ servers	Further instruments	Co-operations internal/ external
Applied Computer Linguistics	Dialogue systems, dialogue, conversation, language interaction	3	3/1	Motion tracking lab, audio recording studio	1/10
Clinical Linguistics	Language, cognition, interaction (basic functions and dysfunctions)	10	15/0		3/5
Emergentist Semantics	Interaction of children between 3 months and 6 years of age, mothers and fathers, human–machine interaction	12	20/0	Four high-definition cameras, several camcorders	2/1

[5] http://pub.uni-bielefeld.de.

Gender and Emotion in Cognitive Interaction	Emotions and gender stereotypes and their role in human–machine interaction	8	10/1		3/7
Language and Cognition	Language understanding, influence of visual context on language understanding	10	15/1	Eye-tracking, video cameras, reaction time PCs	2/5
Phonetics and Phonology	Prosody, human–machine communication, speech synthesis	6	10/1	Audio recording studio, electroencephalography, video cameras	3/6
Psycholinguistics	Interaction, gesture, priming	8	10/1	Two video cameras, audio recording equipment	1/2

Empty table cells indicate that no information has been provided.

Linking research literature and data All groups indicated that it is currently possible to publish literature and data together.

3.2 Social sciences and humanities (SocHum)

Research group profiles Members of this branch (Table C.2) either belong to social sciences or humanities in the traditional sense, such as social anthropology and language studies or they focus on the immediate application of such studies in a computational context. As with BehNeurNat, members of this research branch carry out experiments involving humans.

Research data Figure C.4 shows the different types of data arising in the course of investigating a typical research topic in the social sciences and humanities. Similar to the findings in BehNatNeur, video data constitute the largest part of the data (between 100 and 1000 files and in the terabyte range), followed by audio and text data in the gigabyte range. No other data types were specified in the questionnaire, although the above list of research instruments suggest that at least eye-tracking data and electroencephalographic data arise.

Five out of six groups mentioned that they annotate their data with metadata, although only two groups stated that they are aware of existing standard formats in their field.

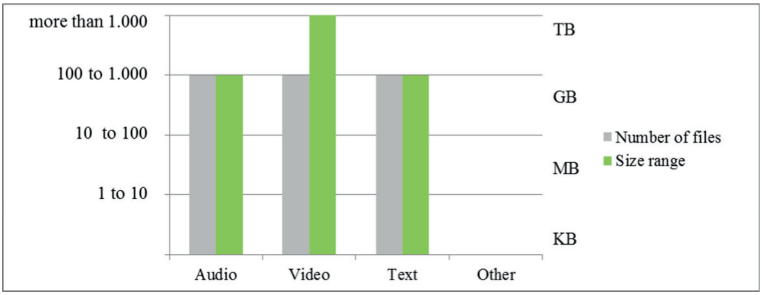


Figure C.4 Data types in terms of number of files and sizes in SocHum

Research data lifecycle Groups identified the following stages in the data lifecycle, with bold stages being those shared by at least half of the groups.

1 **Data collection**
As was mentioned above, groups collect primarily video and audio data in experiments involving humans.

2 **Archiving**
Two groups indicated an intermediate archiving step.

3 **Processing**
The collected data are, in some cases, analysed statistically, whereas some groups mentioned post-processing steps like cutting and compressing.

4 **Archiving**
At least one archiving step is involved in all data lifecycles described.

5 **Enrichment**
Three groups mentioned time-consuming manual annotation and transcription steps, as well as semi-automatic annotation using tools such as ELAN and Praat.

6 **Re-use**
Two groups stated that they re-use the data, for example to generate natural stimuli on the basis of their transcribed and annotated data.

Research data and Open Access Similar to BehNatNeur, research in social sciences and humanities involves experiments with humans. In contrast to BehNatNeur, however, groups in SocHum seem to be more willing to share their data (Table C.5). In general, the majority of groups are willing to share software and primary and secondary data beyond the level of close colleagues. However, only secondary data could, in principle, be made publicly available.

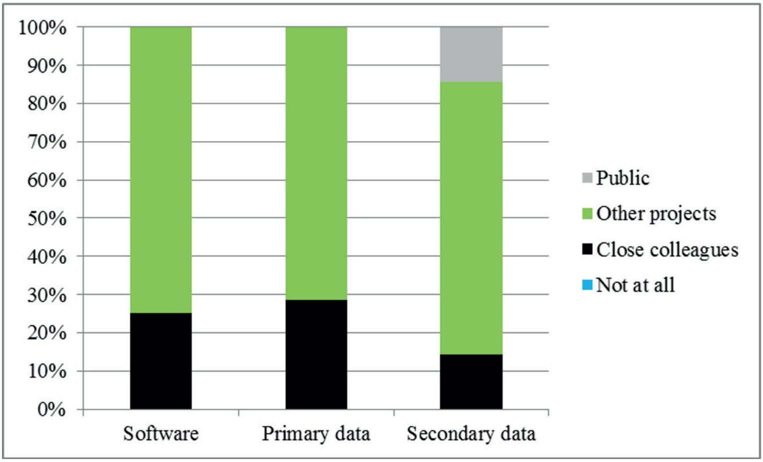


Figure C.5 Willingness to share software and primary and secondary data in SocHum

Research literature Research groups in the social sciences and humanities at CITEC primarily use BibTeX to manage their publication metadata and Citavi, Zotero and Mendeley for managing their scientific literature. They typically create their own publications collaboratively using Google Docs and Subversion. As with BehNatNeur, most groups stated that both print and electronic publication media are generally preferred and established in the field, although one group indicated that only the print medium is established.

Literature and Open Access Two out of seven groups stated that Open Access is not established in their field and four indicated that it is hardly established. Again, this is supported by the empirical website analysis, which showed that of the 38[6] publications analysed, only one (2.63%) was a Golden Open Access publication and 22 (57.90%) were Green Open Access publications. The remaining 15 publications were unavailable.

Linking research literature and data Only two groups mentioned that they are aware of solutions for publishing literature and data together. However, all groups but one agreed that it would be a reasonable and desirable development.

[6] The number of analysed publications of SocHum groups is so low because – as of February 2011 – most of these groups either do not have a group website or do not list their publications.

3.3 Theoretical and applied computer science (CompSci)

Research group profiles Members of this branch (Table C.3) deal primarily with the development of algorithms and computational models and have software as primary output of their research. In contrast to the areas discussed above, experimental studies generally do not involve humans or animals.

Research data Table C.6 shows the different types of data arising in a typical study in theoretical and applied computer science. In contrast to the findings for the previous research areas, text data make up the largest part of the data (more than 1000 files and in the terabyte range), although it can generally be said that all types of data arise in large amounts in this research area. Other types of data indicated in the questionnaires and interviews comprise dialogue data and spectra.

One out of three groups indicated that they annotate their data with metadata, with three groups stating that they are unaware of established standard formats.

Table C.3 Data types in terms of number of files and sizes in BehNatNeur

Name	Research topics	Members	Computers PCs/ servers	Further instruments	Cooperations internal/ external
Computer Graphics and Geometry Processing	Acquisition, modelling, optimisation and animation of virtual 3D objects or characters	10	20/1	3D scanners, motion tracking systems	3/6
Genome Informatics	Theoretical and algorithmic bioinformatics with applications in genome research	15	1/0	Use of the computer infrastructure at the partner institution CeBiTec	0/15
Semantic Computing	Knowledge representation and management, Semantic Web, information retrieval	8	12/2		5/10

Theoretical Computer Science	Neural computation and methods of computational intelligence, esp. prototype-based learning approaches as well as learning theory and self-organisation	5	15/0		1/10

Empty table cells indicate that no information has been provided.

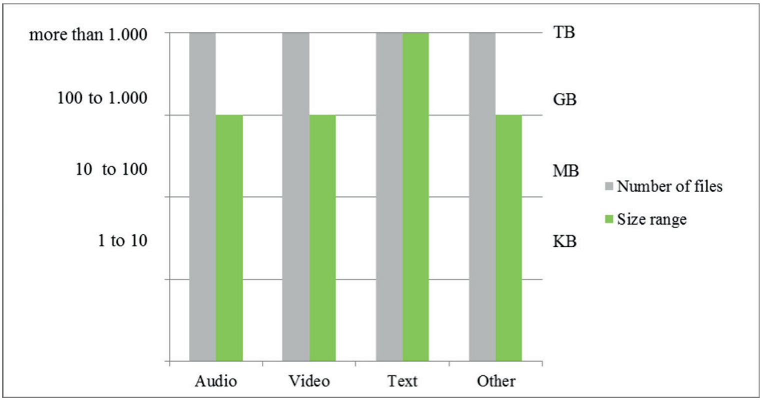


Figure C.6 Data types in terms of number of files and sizes in CompSci

Research data lifecycle Groups identified the following stages in the data lifecycle, with bold stages being those shared by at least half of the groups.

1 **Data collection/re-use**
As was mentioned above, in contrast to the previously discussed research areas, the data collection step does typically not involve experimental work. In fact, groups tend to start by re-using existing data, such as those available on the World Wide Web. However, these data are typically not research data as such (i.e. this step should not be considered as re-using research data), but rather data from social media like Twitter or Flickr.

2 **Enrichment**
One group indicated that they first annotate the initial data set with further information before they start processing the data algorithmically.

3 **Processing**
In CompSci, this step typically marks the central stage in the research workflow, as it is concerned with the application of the algorithms developed by the researchers.

4 **Archiving**
As with the previous research areas, at least one archiving step is involved in the data lifecycles observed.

5 **Re-use**
One group indicated that the archived data are typically re-used at a later stage for testing and comparing the performance of newly developed algorithms.

Research data and Open Access In contrast to BehNatNeur and SocHum, groups in CompSci are generally more open when it comes to sharing data. In particular, all data types could conceivably be shared beyond the level of close colleagues, with public availability being accepted by the majority of groups both with respect to software and secondary data (Table C.7). This shows that Open Access (or Open Source in terms of software) is a well-established practice in CompSci.

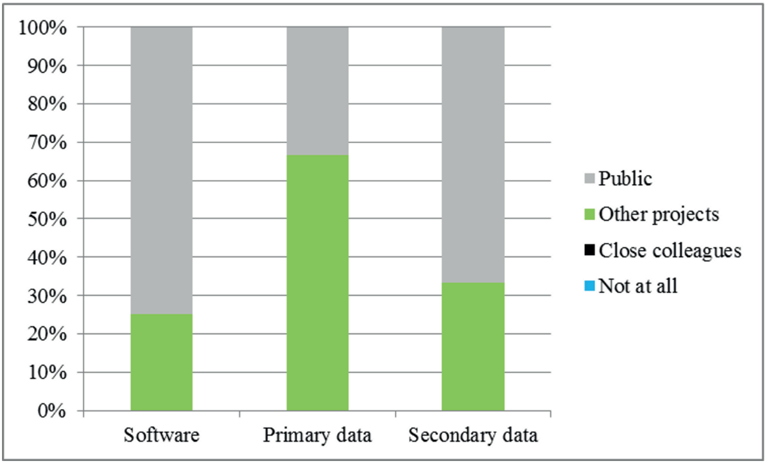


Figure C.7 Willingness to share software and primary and secondary data in CompSci

Research literature Research groups in theoretical and applied computer science at CITEC primarily use BibTeX to manage publication metadata, as well as the Drupal content management system for managing metadata and the publications themselves. Besides this, Mendeley and the university's PUB system were mentioned, as well as Subversion for collaboratively creating literature. In terms of publication media, electronic publications seem to be preferred over publications in printed media.

Literature and Open Access The openness with respect to the willingness to share research data is also reflected in the status of Open Access to literature, as only one out of five groups considers Open Access not to be established in the field. However, it needs to be said that the actual status of Open Access to literature as suggested by the empirical website analysis is still very similar to the BehNatNeur and SocHum. In fact, only one out of 100 publications was a Golden Open Access publication. Green Open Access publications, however, made up the vast majority of the publications, namely 78%. The remaining 21 publications were unavailable.

Linking research literature and data Two out of three groups indicated that they are unaware of possibilities for publishing literature and data together, agreeing, however, that it would be desirable.

3.4 Robotics and engineering (RobEng)

Research group profiles Members of this branch (Table C.4), if not concerned with the actual engineering of machines, typically have software and models as one of their primary research outputs as well. In contrast to the previous branch, however, work in robotics and engineering at CITEC typically involves experimental studies with humans and robots, and the software and models are generally directly applied to robots or other engineered systems.[7]

[7] It should be noted that the research focus of the Applied Informatics group has changed towards robotics in recent years, and a considerable amount of research deals with experimental studies involving, for example, the interaction between humans and robots. Therefore, although being traditionally rooted in the computer science field, it has been classified as belonging to the robotics and engineering branch.

Table C.4 Willingness to share software and primary and secondary data in BehNatNeur

Name	Research topics	Members	Computers PCs/ servers	Further instruments	Co-operations internal/ external
Applied Informatics	Pattern recognition, computer vision, software engineering, evaluation of cognitive systems, human-inspired memory, social robotics	40	50/2	Robotic platforms, 3D cameras, headmounted displays)	10/8
Cognitive Robotics and Learning	Neural learning methods, esp. recurring reservoir networks, transfer of other machine learning approaches to interactive scenarios	8	8/2	Several robot platforms	2/2
Cognitive Systems Engineering	Motion generation using dynamic systems, software and systems engineering for cognitive robotics, architecture of intelligent systems	10	10/1	iCub (humanoid robot), several special hardware platforms, usage of the general CoR-Lab infrastructure	3/1
Cognitronics and Sensor Systems	Cognitronics, microelectronics, CPU design, sensor systems	20	40/10	High-performance measuring instruments with network connection, research platform “tele-workbench” with network cameras and video and data servers, special hardware platforms for rapid prototyping of microelectronic switches based on FPGAs	3/15

Neuroinformatics	Data mining, brain–computer interfaces, evolutionary computation, complex systems integration	32	60/1	Brain–computer interface system, eye-tracking devices, three cybergloves, depth cameras, in-house developed tactile sensors, robot set-up (2 PA10 arms, two shadow hands), robot set-up (two Kuka lightweight arms, Schunkhand), manual intelligence lab (14 Vicon cameras)	10/10
Sociable Agents	Development of systems for intuitive and natural human–machine interaction	9	20	3D camera, two time-of-flight cameras, eye-tracking, two cybergloves, headmounted display, two 60” displays	

Empty table cells indicate that no information has been provided.

Research data As with the previous disciplines, data arise in large quantities also in the robotics and engineering groups. Figure C.8 shows that video and other data types constitute the largest portions of the research data (both in the terabyte range), where eye-tracking, motion capturing, tactile, simulation and design data, as well as binary data, were named among “other data types”.

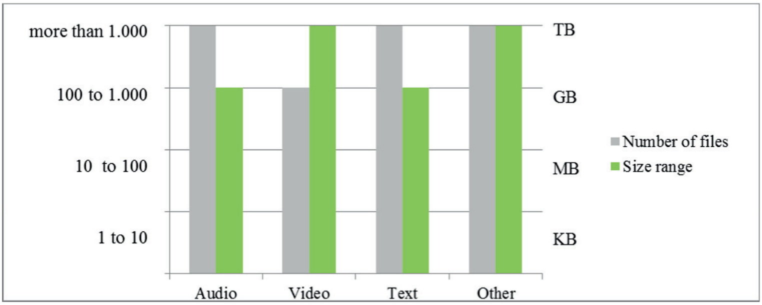


Figure C.8 Data types in terms of number of files and sizes RobEng

As with the CompSci groups, half of the groups stated that they annotate their data with metadata and three of the four groups stated that they are unaware of established standards in the field.

Research data lifecycle Groups indicated the following stages in the data lifecycle, with bold stages being those shared by at least half of the groups.

1. **Data collection**
   All groups begin by collecting data in experiments, typically involving the interaction between humans and machines, humans performing cognitive tasks or autonomous robots performing tasks.
2. **Processing**
   One group indicated a processing step consisting of post-processing the data recorded by the Vicon system, as well as compressing the recorded videos.
3. **Enrichment**
   Two groups indicated that they annotate their data, using a tool like Anvil, for example.
4. **Processing/analysis**
   All groups analyse their data, for example by applying different machine learning algorithms to them.
5. **Archiving**
   As with the previous research areas, one archiving step is part of the data lifecycle.
6. **Re-use**
   Two groups stated that they re-use their data and experimental set-ups at later stages.

Research data and Open Access Similar to the observations for the CompSci groups, research in robotics and engineering is rather open with respect to the willingness to share data. As can be seen in Figure C.9, software and primary data are mostly considered to be made publicly available. One group indicated, however, that in some projects the amount of generated data is so large that it is considered to be too much for being shared.

Research literature Research groups in robotics and engineering at CITEC primarily use Drupal, Endnote, BibTeX and Subversion for handling their publications. There seems to be no preference with respect to publication media, with printed and electronic publications being well established.

Literature and Open Access Two of the three groups which answered the respective question in the questionnaire stated that Open Access is hardly

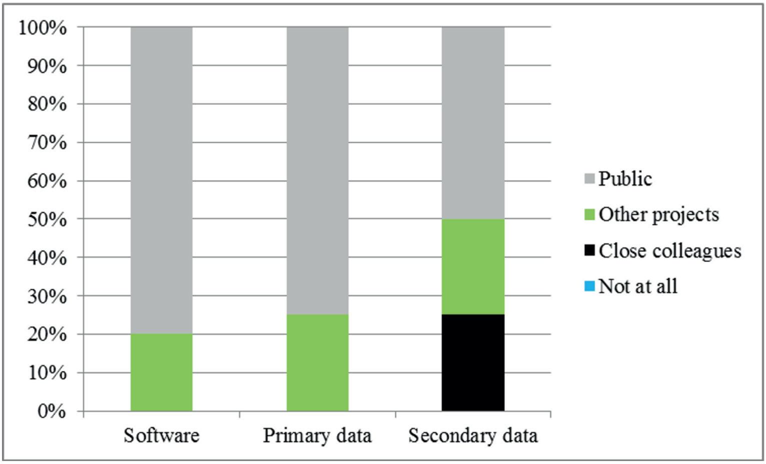


Figure C.9 Willingness to share software and primary and secondary data RobEng

established in their field. As with the other research areas, the empirical website analysis confirmed that only a tiny fraction of the analysed publications are Golden Open Access publications. In particular, five (1.17%) of the 428 publications were Golden Open Access publications and 292 (68.22%) were Green Open Access publications. The remaining 131 publications were unavailable following the strategy explained above (see 2.5 Website analysis of publication behaviour).

Linking research literature and data Two groups indicated that it is currently possible to publish literature and data together, and the two groups which were not aware of such possibilities stated that it would be a desirable development.

4 Representativeness of this case study

The methods presented above (see 2 Methodology) were applied to a number of research groups in each of the research branches just introduced. Whenever possible, it was attempted to apply each method to at least one representative of each branch, which succeeded for almost all methods.[8] However, since all branches are well covered in the questionnaire that forms the basis of the

[8] One of the reasons for not having observed experiments in CompSci is due to the fact that such are rarely carried out.

semi-structured interview, we do believe that this study gives a representative description of the entire case CITEC nonetheless. The overall participation according to methods and research branches is given in Figure C.10, and can be seen as an attempt to quantify the representativeness of this study. It should be noted, however, that representativeness refers to "representative of the institution" in this context, not to "representative of the field of ICT". We are aware of the fact that while this chapter gives a representative account of CITEC, it describes only one of many possible examples within the field of ICT.

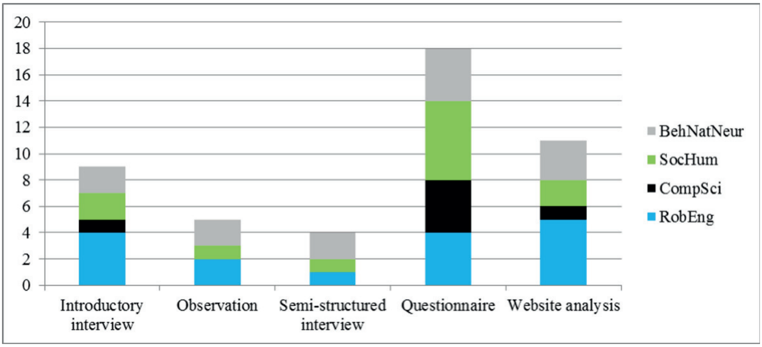


Figure C.10 Overall participation in the case study according to methods and research branches

As was mentioned above, because the semi-structured interview had been guided on the basis of the questions in the questionnaire, a group either participated in the semi-structured interview or completed a questionnaire. As such, 21 out of 32 research groups (65.63%) answered detailed questions (i.e. either participated in a semi-structured interview or completed a questionnaire) and 23 groups (71.88%) were covered by this study in some way or another. Figure C.11 summarises the participation according to research branches and overall.

5 Current status of research infrastructure

This section presents a detailed and consolidated view on the current research infrastructure at CITEC, focussing first on available infrastructural facilities and services. Afterwards, we will discuss the types and amounts of data dealt with at CITEC, as well as the various stages they pass through.

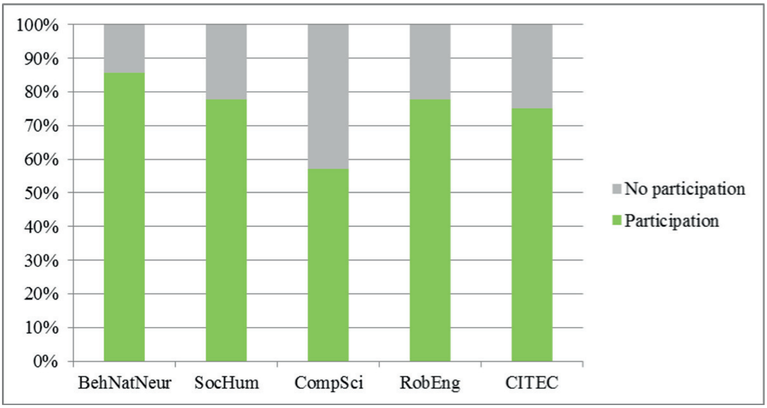


Figure C.11 Overall participation in the case study

5.1 Infrastructural facilities and services

5.1.1 Computing and network infrastructure

CITEC hosts a computational infrastructure that is maintained in large parts by the *Rechnerbetriebsgruppe* (RBG; IT services group) of the Faculty of Technology at Universität Bielefeld, which operates the computational infrastructure at the Faculty of Technology. Based on the figures given above (see 3 Case narratives), a research group has on average 18.67 desktop PCs or notebooks, the majority of which run on Unix-based operating systems like Mac OS X or Linux. This is supported by a survey carried out in a different context prior to this case study, which revealed that the use of non-Unix operating systems (such as Microsoft Windows XP) was well below 10%. Although this figure cannot be taken at face value, it nonetheless gives a hint as to the actual distribution of operating systems at CITEC, at least in technical disciplines. In the BehNatNeut groups, however, Windows seems to be the predominant operating system. In fact, the interviews as well as observation sessions have shown that groups are sometimes forced to resort to a non-Unix operating system in case they use special hardware or commercial software which depends on such.[9] These systems are, however, in most cases maintained by the researchers of the groups, have restricted network access and are thus largely independent of the infrastructure operated by the RBG. RBG further operates a *Subversion* revision control system[10] server that can

[9] The importance and use of special hardware and commercial software in the data lifecycle is discussed in more detail below (see 5.1.3 Research instruments and 5.2 Overview of the data lifecycle).

[10] http://subversion.apache.org.

be used to store group-related or project-related data. Finally, the CITEC Compute Cluster (C3) provides powerful computational support and is operated by *Central Labs Facilities* (CLF) and connected to the network operated by the RBG.

In addition to the facilities offered and maintained by the RBG, there is a wide range of further services available at CITEC. Although these are still hosted on servers provided by the RBG, the services themselves are generally offered by CLF and consist of, for example, collaborative research environments like the *CITEC Social Network*, as well as central repositories like the *CITEC OpenSource Server*. These will be introduced in the following subsections.

5.1.2 Social network and collaborative services

The *CITEC Social Network*[11] is a platform based on the OpenSource social networking engine Elgg,[12] where researchers at CITEC can create and join groups in order to exchange opinions, discuss particular research topics or upload documents. One of these groups is concerned with, for example, the CITEC Software Round Table, a regular strategic meeting aiming to discuss issues regarding the creation of a cognitive interaction toolkit, as well as exchange experiences and best practices regarding software and frameworks used. Members of this group can participate in these discussions and have access to the documents presented at sessions of the regular colloquium. In addition to this, every member of CITEC has access to an *instant messaging service*, further supporting exchange and collaboration among researchers. As of February 2011, the *CITEC Social Network* has more than 400 members and around 30 different groups.

The *CITEC Project Development Platform*[13] is a collaborative development environment which builds on the Redmine[14] project management web application and features an integrated Wiki and discussion forum. For each project, the platform provides a file space for work documents and project deliverables, as well as issue tracking and milestone management. Moreover, the platform is connected to a *project repository farm*, a revision control system that provides a dedicated Subversion repository for each project.[15] Access to a project repository requires a login and is determined on the basis of the role that the respective person has in the project. Moreover, projects in the

[11] https://social.cit-ec.uni-bielefeld.de.

[12] http://www.elgg.org.

[13] https://projects.cit-ec.uni-bielefeld.de.

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

[15] https://projects.cit-ec.uni-bielefeld.de/svn/<project-id>.

farm are arranged by topic, which allows for collaboration beyond the level of research groups or institutes.

The *CITEC OpenSource Server*[16] provides a central repository for depositing and obtaining open-source software. While primarily aiming at storing software developed at CITEC, the *CITEC OpenSource Server* openly invites researchers from other institutions to not only use the software, but to contribute and collaborate on software development as well, with the goal of creating an open library of software related to cognitive interaction technology. Similar to the Project Development Platform, the OpenSource Server is connected to revision control system repositories. However, both the OpenSource Server and Repository Farm are publicly accessible. Figure C.12 summarises the collaborative research infrastructure available at CITEC. As can be seen there, all components access a *directory service*, an LDAP server hosting a directory with information on all CITEC members and associates, such as contact details and affiliations.

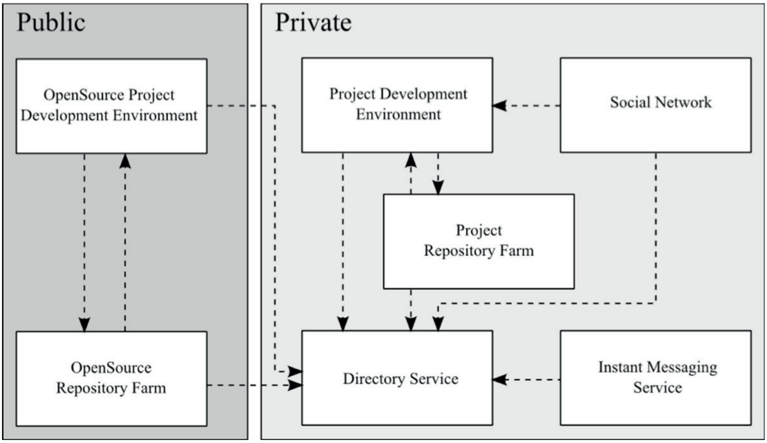

Public
Private
OpenSource Project Development Environment
OpenSource Repository Farm
Project Development Environment
Social Network
Project Repository Farm
Directory Service
Instant Messaging Service

Figure C.12 Summary of the collaborative research infrastructure at CITEC[17]

5.1.3 Research instruments

As was mentioned above (see 1 History, structure and mission), research in cognitive interaction involves a considerable amount of empirical investigation of humans and animals, focusing in particular on the way they handle certain

[16] http://opensource.cit-ec.de.

[17] This image is based on a figure created by Thilo-Paul Stueve presented at the CITEC Software Round Table colloquium.

cognitive tasks. In the following, we will illustrate the variety of instruments being used in such studies by means of a concrete example. In particular, we discuss a collaborative experiment between the groups Neurocognitive Psychology and Neuroinformatics that aims at analysing learning, interaction and automatisation in speedstacking. Speedstacking consists in stacking and destacking ten plastic mugs in several predefined formations in the shortest possible amount of time.[18] At the beginning of one speedstacking exercise, a previously untrained participant is asked to rest his or her hands on a hand timer, a device measuring the time needed to complete the exercise. The exercise starts with the participant taking his or her hands off the timer. The participant then performs the task and puts his or her hands back on the hand timer as soon as the task has been completed. In particular, the participant tries to complete as many individual speedstacking task iterations as possible within 3 minutes. This is repeated for five times and makes up one complete experiment.

In the time between each experiment, the participant is asked to practice the speedstacking task for at least 45 minutes per day. In these experiments, the previously mentioned goals are investigated along different lines. For example, the focus of the eyes of the participant is recorded by means of a so-called eye-tracker, in order to examine whether the position changes over time (e.g. at a certain point in time and 1 week later). Moreover, measuring the progress that the previously untrained participant makes after a certain amount of training is taken as a learning indicator, where progress is measured primarily on the basis of the decrease in the time needed to perform the task over time. Finally, the hand movements of the participant are recorded by means of special markers attached to them, as well as special cameras tracking those markers. This is done to have three-dimensional trajectories of the movements of the hand in digital form, which means that they can be transferred to an artificial system, such as a robot, at a later time. Due to these different aspects, each experiment involves a number of different instruments serving different purposes. In addition to the ones just mentioned, for example, irregular speedstacking completions (e.g. falling mugs or any other kind of disruption) are manually annotated as containing mistakes, in order to ensure that the times measured during such iterations is not taken into account when analysing the overall learning curve and thereby ensuring a certain level of quality of the data observed within an experiment. This annotation is done in a different computer than the one used, for example, to record the eye-tracking data. Figure C.13 shows the complete set-up of instruments used in the particular experiment that was observed.

[18] https://www.cit-ec.de/research/ALIAS.

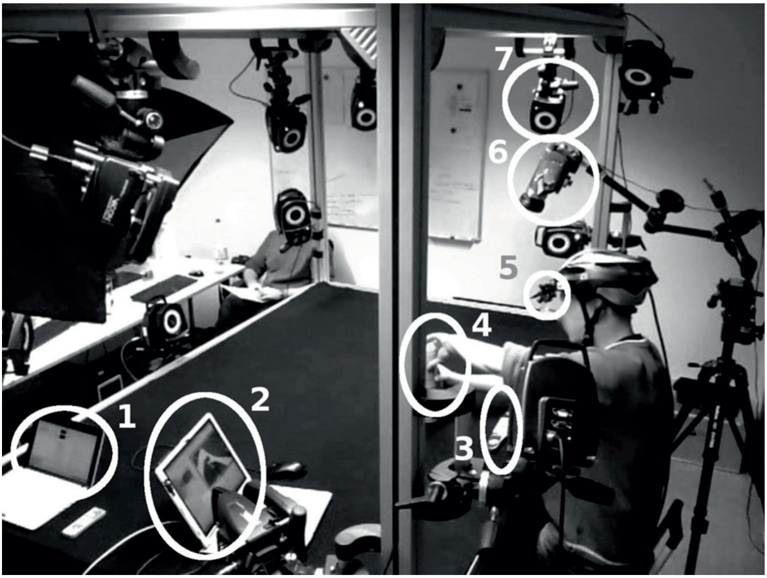


Figure C.13 Set-up of a collaborative experiment of the groups Neurocognitive Psychology and Neuroinformatics

In particular, the following instruments have been used, according with their purpose:

1 **Laptop “MacBook Pro”:** for recording the amount of time needed to complete one speedstacking iteration and to manually annotate whether there has been a mistake or disruption in this iteration.

2 **Laptop “Windows XP”:** for handling the data recorded by an infrared camera (no. 5) that is attached in front of the eyes of the individual, as well as of a further head camera. The recorded videos are displayed on the laptop in real-time.

3 **Hand timer:** recording the amount of time needed for stacking.

4 **Six special markers:** (three per hand) allowing a camera (no. 7) to track the movements of the participant’s hands in 3D.

5 **Head camera and eye infrared camera:** for recording the view the participant has, as well as where the participant is looking.

6 **Scene camera:** to have a further perspective of the stacking scene. Here, all mugs are always in sight, which is not necessarily the case with the images recorded by the other cameras.

7 **Fourteen Vicon cameras**: recording the 3D coordinates of the markers

8. **Computers "Windows XP":** (not shown in Figure C.13) with special commercial software processing the images recorded by the Vicon cameras.

Table C.5 below summarises the other research instruments used at CITEC, as observed in other experiments or indicated by answers in the questionnaire (see 3 Case narratives).

Table C.5 Research groups in social sciences and humanities

Instrument/platform
3D scanners
Audio recording studio
Brain–computer interface system
Data gloves
Depth cameras
Electroencephalography
Electromyography
Electrophysiology (at least five set-ups including binoculars, intra- and extracellular amplifiers, analogue-digital converters, PCs, oscilloscopes, micromanipulators, stimulators, frequency generators and electrode pullers)
Eye-tracking
fMRT
High-performance measuring instruments with network connection
Hydrodynamics
iCub
Microsoft Kinect
Motion capture (at least four set-ups, e.g. Vicon MX10 with 8–13 cameras, two self-built set-ups with three Basler-A602 cameras each, objectives and ringflash system)
Research platform "Tele-workbench" with network cameras and video and data servers
Robot platforms
Special hardware platforms for rapid prototyping of microelectronic switches based on FPGAs
Tactile sensors
Video cameras
Virtual and augmented reality

5.1.4 Data management

As mentioned above (see 5.1.1 Computing and network infrastructure), CLF provides central revision control repositories for archiving project-related data. However, the overall analysis of the various interviews and questionnaires clearly suggests that there is no general data management strategy that is followed by all groups. In addition to this, however, the interviews especially revealed that there is no general archiving strategy within a group, but that it rather depends on the particular project as well as the partners involved in a project. For example, data created in EU-wide projects are frequently stored in external repositories which are hosted by one of the participating project partners. Some CITEC-internal projects make use of the collaborative research environment operated by CLF, whereas others make use of the storage infrastructure provided by project X1 of the Collaborative Research Centre 673 "Alignment in Communication". Finally, projects involving either a rather small amount of people, such as PhD projects, for example, or group-internal projects seem to be primarily stored on group-internal – or even personal – storage devices. Figure C.14 summarises the overall data management strategy at CITEC, represented by the answers given to the question which devices the group uses for storing their data.

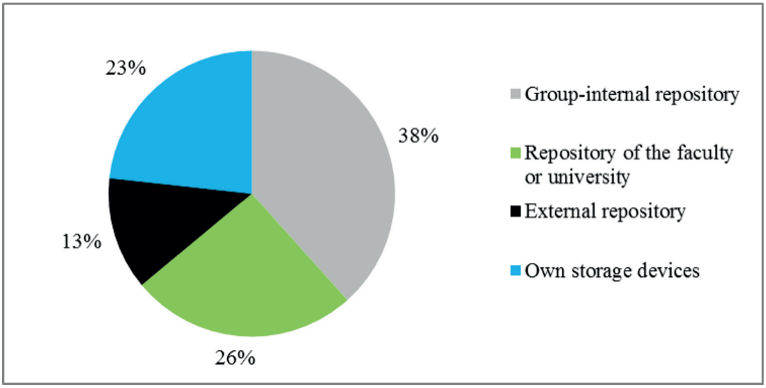


Figure C.14 Overall data management at CITEC

As can be seen in the figure, despite the availability of a central data management infrastructure, there is no homogeneous data management strategy building on this infrastructure, since only 26% of all research groups make use of it. On the one hand, this is certainly in part due to the reasons mentioned in the introduction to this chapter, namely that CITEC is a very young institution involving previously existing research groups, some of which adhere to the management procedures they had previously established. On the other hand, however, this may in part be because groups do not have designated personnel for dealing with questions of data management and may therefore be not perfectly well informed about the available infrastructure, unable to make use of the infrastructure, possibly due to a lack of sophisticated technical background knowledge, or have their own independent infrastructure.

This is supported by the answers given by groups as to whether they have personnel in charge of data management questions. The results are given in Figure C.15.

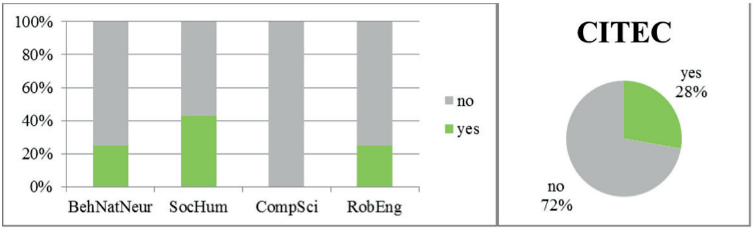


Figure C.15 Groups having a person in charge of data management

The results show that only 28% of the research groups at CITEC have a person in charge of data management. A follow-up question revealed that of the 72% which do not have such personnel, 69% would like to have such personnel. The distribution according to research branch is given in Figure C.16.

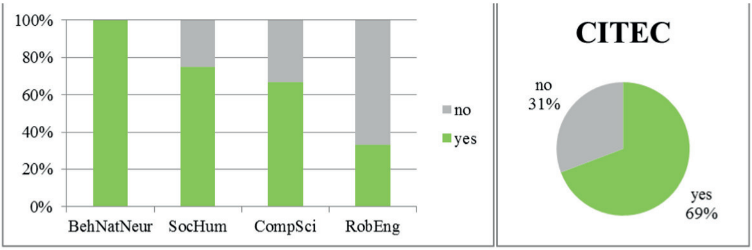


Figure C.16 Groups wishing to have a person in charge of data management

In case groups motivated their choices, the main reasons indicated in favour of having a person in charge of data management are the large amount of data being dealt with and the possibility to obtain a better general overview of the data being managed in a group. This would in turn increase the sustainability of the data, allowing for better reuse and thus comparability. Main reasons against having a person in charge of data management were that it was not a primary task area and as such not is financeable or that it works as it is – either due to an easily manageable amount of data or because researchers themselves are well grounded in data management issues. Especially the latter seems to be the case in CompSci and RobEng groups. While this is certainly true, however, it clearly explains the heterogeneous distribution shown in Figure C.14 above, since data management tasks are transferred to the researchers themselves. Therefore, there is the strong requirement to have budget assigned to the task of managing research data. This does not mean, however, that each group needs to be given a new member that takes over this task – although in some cases this may certainly be reasonable – but rather that groups are provided with budget to be able to appoint and finance a specific member of the group to take over data management tasks, such as participating in strategic meetings of CLF in order to be informed and to agree on global data management practices at CITEC.

In addition to the resource-based aspects just discussed, the data management strategy taken strongly depends on the diversity of data types that are to be managed. For primary data obtained in experiments involving humans, for example, it may in many cases not even be permitted to store them in external repositories, because other people may have access to them. This means that a more fine-grained analysis of data management is necessary, which is given as part of the discussion of the data life cycle (see 5.2 Overview of the data lifecycle).

5.1.5 Publication management

With regard to publication management, groups were asked which tools they use for managing internal and external literature (with multiple answers allowed). Here, 26% of all groups use BibTeX to store publication metadata and one-fifth use the Drupal CMS to manage internal publications, including metadata – which can in turn be exported in BibTeX format – and, in some cases, the publications themselves. The most frequently used tools are shown in Figure C.17. It should be noted that the vast majority of tools used are freely available, with Endnote being the only exception. Moreover, the figure suggests that publication management is done on a group-internal basis. This is to say that no group answered that they use a group-external repository for depositing publications. This is obviously because no such repository was available at the time of writing. However, as will be discussed later in this chapter (see 9.1 Literature management), current developments are clearly headed into this direction.

5.2 Overview of the data lifecycle

In the following, we will discuss the data lifecycle extrapolated from the descriptions given by researchers in the questionnaire. Figure C.18 shows the general stages that can be identified (see 3 Case narratives for a description of the data lifecycle in the various disciplines).

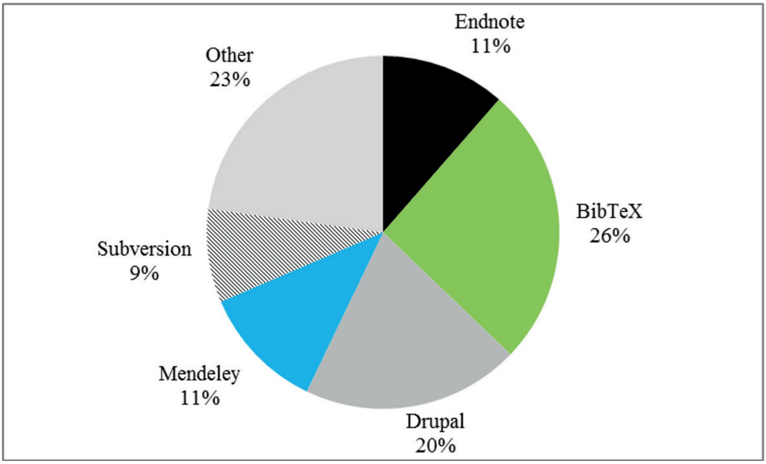


Figure C.17 Overview of publication management tools used at CITEC

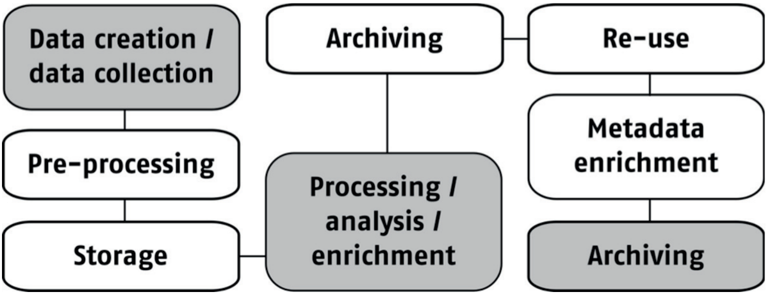

Data creation / data collection
Pre-processing
Storage
Archiving
Processing / analysis / enrichment
Re-use
Metadata enrichment
Archiving

Figure C.18 Different stages in the data lifecycle

The grey boxes in the figure represent those stages which could be identified by (almost[19]) all groups participating in the survey. In the following subsections, we will discuss each stage in more detail, starting with data creation/data collection in the top left-hand corner.

5.2.1 Data creation and collection

The variety of research instruments being used in a single experimental study, as well as the other instruments that are typically used at CITEC, has been illustrated earlier in this chapter (see 5.1.3 Research instruments). These in-

[19] In case one or two groups did not list a particular stage, we still considered it as representative and marked the respective box in grey.

struments do, of course, produce very different types of data. In the following, we will discuss the different types of data that arise, as well as the scale (in terms of storage requirements) at which they are created. In addition to the primary data created this way, the following also includes cases of collecting primary data. This is, for example, the case in which the primary data of a group consist in material found on the web, such as images or texts on general websites.

Data types and scales In order to get an overall picture of the types of data being created at CITEC, we asked research groups to specify the types of data typically arising in the investigation of a particular research object, as well as a rough estimate of the quantity in terms of number of files and memory requirements. The distribution according to types of data is shown in Figure C.19 and Figure C.20.

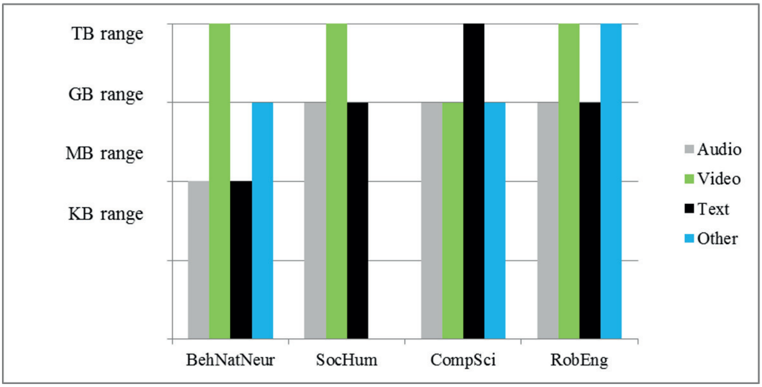


Figure C.19 Data types occuring in different disciplines (memory requirements)

The figures show that video data arise in all research areas in considerably large quantities, posing very high storage demands in almost all of them. For example, the Neurobiology group reported behavioural experiments which involve high-speed cameras recording 500 images per second, each of them with a resolution of 1 megapixel. With up to three high-speed cameras in an experimental set-up, each second recorded thus requires around 1.5 GB of disk space, and a currently running PhD project has thus created around 9 TB so far. In addition to audio, video and textual data, several groups specified other types of data occurring in considerable quantities and size, such as eye-tracking data, electroencephalograms or nuclear magnetic resonance spectra. Overall, it can thus be said that all research areas at CITEC pose high

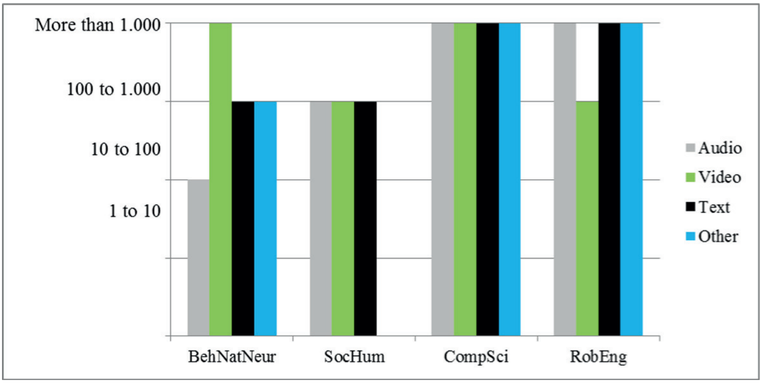


Figure C.20 Data types occuring in different disciplines (number of files)

demands on storage infrastructure. How these are managed will be discussed below (see 5.2.2 Pre-processing).

Software In addition to the different types of data arising in the different disciplines, groups were asked to specify whether they rely on commercial software in order to obtain primary data. Here, Figure C.21 clearly shows that more than half of the groups rely on commercial software at least to a considerable extent, with 11% relying (almost) entirely on commercial software. This further suggests that groups depend on the hardware required by the software in order to run successfully.

5.2.2 Pre-processing

Two groups indicated a pre-processing step taking place in the data lifecycle. For example, this was the case for the experiment described above (see 5.1.3 Research instruments). Here, the data recorded by the Vicon cameras (i.e. the trajectories of the markers which are attached to the participant's hands) are directly reviewed in the Nexus[20] tool after the experiment, before the data are transferred to other storage media. This is done in order to make sure that the cameras were able to trace the hand movements correctly. In cases where this is not the case, the data can be corrected manually in the tool. Other groups mentioned processing steps like digitisation, cutting of audio and video data or data compression. In all cases that include such a processing step, the answers indicate that raw primary data are not used in later stages of the workflow.

[20] http://www.vicon.com/products/nexus.html.

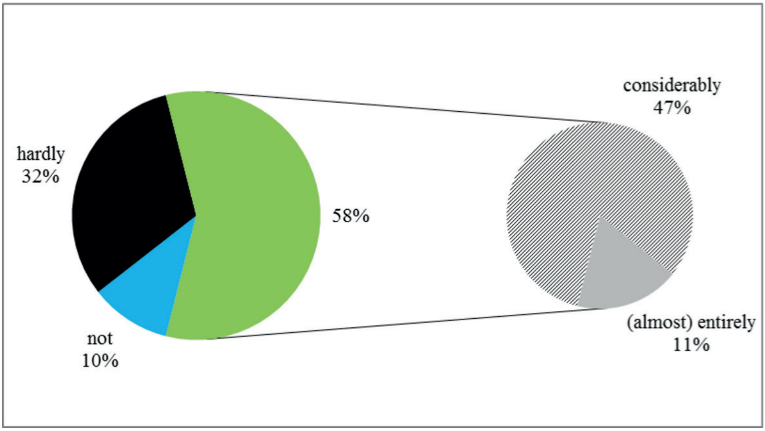


Figure C.21 Overall dependence on commercial software for generating primary data

5.2.3 Storage and transmission

After the data creation (or collection) and a possible pre-processing step, many groups indicated a first archiving step. However, since most groups did not mention this step, we do not consider it to be a crucial step in the data lifecycle. For those groups which did indicate it, this step consists in archiving either the storage media on which the data had been recorded, such as digital audio tapes, or other storage media to which these data had been transmitted, such as DVDs or external hard disks. In other cases, this step simply consists transferring data to the hard disk of the respective researcher.

5.2.4 Analysis and enrichment

All groups which create or collect primary data mentioned an analysis step, which mainly consists in analysing the primary data by means of statistical tools like SPSS or programming languages like R or MatLab. In many cases, this includes an enrichment step, in which secondary data are obtained by annotating the primary data with tools like ELAN or Praat. Since in some cases enrichment precedes analysis and in others vice-versa, we grouped these two steps together in one stage.

5.2.5 Archiving

In most groups, this intermediate archiving step does not exist. It is, however, an integral part of those groups involved in the creation of the so-called

Manual Interaction Database,[21] an effort to create a strong empirical basis for investigating research questions in the area of *Motion Intelligence* (see 1 History, structure and mission). Here, groups store the post-processed data from their experiments in SQL databases in order to allow for future re-use.

5.2.6 Re-use

Re-using, for example, algorithms, methodologies, models or annotated data either for follow-up experiments or – in the case of software components – even in other contexts is a very common step in the data lifecycle. In addition to the re-use of manual interaction data just mentioned, models for classifying data, as typically created by machine learning approaches, are frequently applied to data sets other than those on the basis of which they had been obtained, primarily in order to measure their performance on these previously unseen data sets. Similarly, physics-based models representing the physical properties of a human hand are re-used in robotic systems in order to achieve comparable behaviour in a robotic hand. Finally, annotated data are frequently used in other settings to investigate other research questions, such as linguistic phenomena in the case of text corpora. For a number of research groups, however, data re-use – aside from publishing the findings obtained on their basis – is not a common procedure, with one group indicating that data are in fact re-used too rarely.

5.2.7 Metadata enrichment

Most groups annotate their data with metadata, which is shown in Figure C.22 below.

A follow-up question asked whether existing standards are used for this annotation or whether groups use custom formats. All of the groups which annotate their data with metadata use existing standards do so either quite frequently (57%) or almost always (43%). Reasons for deviating from standard formats were the lack of metadata fields for annotating proband information. Of all groups, 59% indicated that there are no established metadata standards in their field.

5.2.8 Archiving

As was discussed above (see 5.1.4 Data management), different data management strategies are followed at CITEC. The strategy chosen depends on different factors, one of which is the type of data that is being managed. In line with general practice, we distinguish between primary and secondary

[21] http://www.cit-ec.de/research/MINDA.

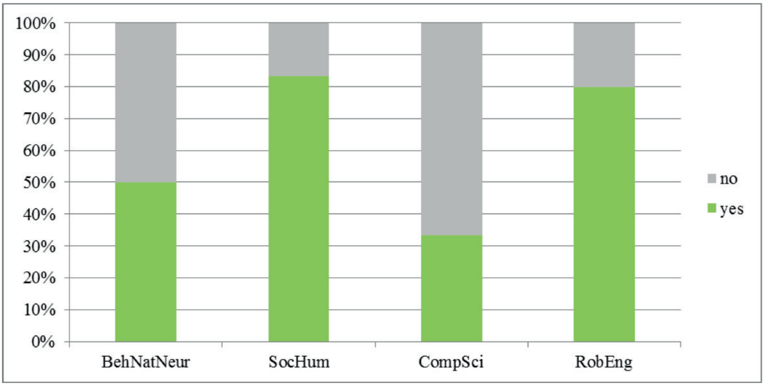


Figure C.22 Metadata enrichment according to research branches

data. Although software could be considered a special kind of secondary data, we treated it separately in this study – mainly due to its importance in a computationally oriented research field and the expected difference in how it is managed in contrast to primary or secondary data. Figure C.23 summarises the different archiving strategies according to types of data.

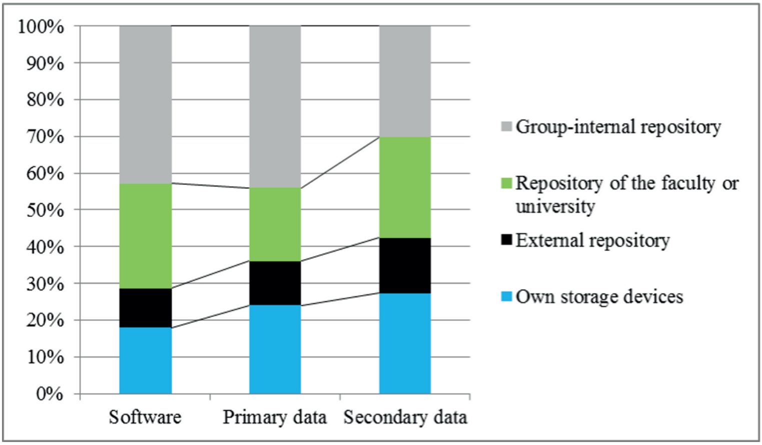


Figure C.23 Archiving strategies according to data types

As can be seen in Figure C.23, it is far more common to use repositories provided by the faculty or university for storing software and secondary data than it is for storing primary data. The figure suggests that the latter are typically stored on own storage devices or those provided internally by the group. In general, it can be clearly seen that using external repositories for storing data is rather uncommon for all kinds of data, though slightly more common for secondary data.

As far as backup strategies are concerned, most groups mention that they generally perform regular backups with standard backup systems, on file servers and/or individual computers. In addition to this, indirect backups are in many cases achieved by using a revision control system like Subversion, since people working with the data stored there usually have a local version of the respective repository. However, it is generally the case that no complete snapshots are backed up this way, which means that this strategy cannot be considered a standard backup procedure. In contrast to this, groups which store their data on repositories of the IT services department of the Faculty of Technology at Universität Bielefeld can make use of the backup policies followed there, which includes backing up complete snapshots of the data stored in the repositories. Finally, some groups cooperating with the Collaborative Research Centre 673 at Universität Bielefeld make use of the server provided by infrastructure project X1 in order to archive their data.[22]

6 Current status of Open Access to literature

The results presented in the following sections are based on literature-related questions in the questionnaire, as well as the empirical website analysis as described earlier in this chapter (see 2.5 Website analysis of publication behaviour), which was carried out in February 2011. A discussion of the results is given below (see 6.3 Discussion of results).

6.1 Results of questionnaire

In the questionnaire, groups were asked to state whether Open Access is established in their group and field of study. The results – grouped according to research branch – are given in Figure C.24.

6.2 Results of empirical website analysis

The following figures illustrate the distribution of Golden and Green Open Access publications, in relation to all publications of a particular group (see 2.5 Website analysis of publication behaviour for classification criteria). Figure C.25 shows the results grouped according to research branch, and Fig-

[22] http://www.sfb673.org/projects/X1.

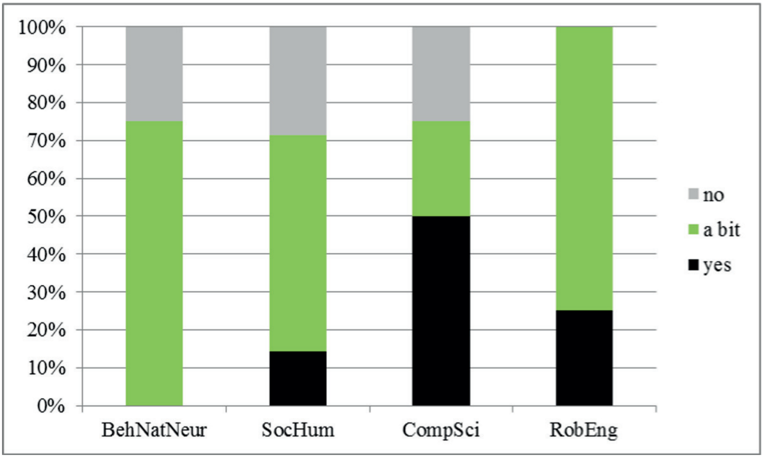


Figure C.24 Groups' replies to whether Open Access is established in their group or field of study

ure C.26 summarises the overall publication behaviour at CITEC. Absolute figures are given in Table C.6.[23]

Table C.6 Data types in terms of number of files and sizes in SocHum

Discipline	Publications analysed	Golden Open Access publications	Green Open Access publications	Unavailable publications
BehNatNeur	231	16	79	136
SocHum	38	1	22	15
CompSci	100	1	78	21
RobEng	428	5	292	131
CITEC	797	23	471	303

6.3 Discussion of results

As the results given above show, Open Access seems to be established in all research groups at least to some extent. In particular, of all groups participating in the questionnaire, 78% answered "yes" (22%) or "a bit" (56%). When looking at the results of the empirical website analysis, it becomes clear that

[23] As was mentioned above (see 3.2, the number of analysed publications of SocHum groups is so low because – as of February 2011 – most of these groups either do not have a group website or do not list their publications.

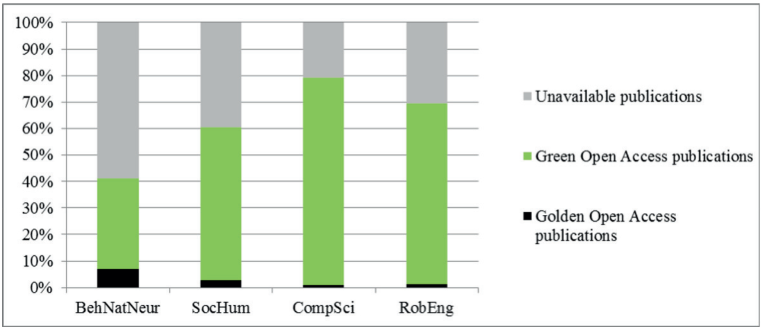


Figure C.25 Golden and Green Open Access publications according to research branch

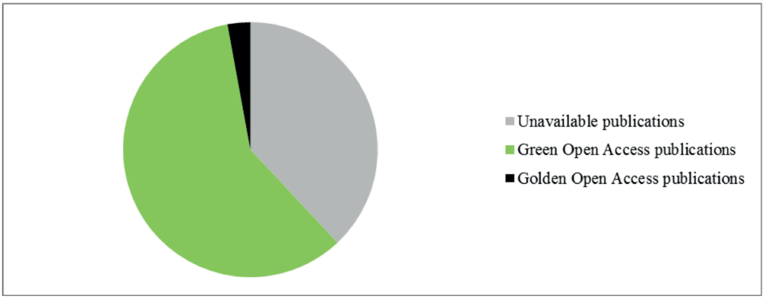


Figure C.26 Overall publication behaviour at CITEC

this can only refer to Green Open Access publications, since only 3% of all publications analysed were Golden Open Access publications, according to the criteria given above (see 2.5). In addition to this, it should be noted that a freely available publication is not necessarily freely available with the authors' knowing about this. For example, the American Physiological Society mentions that articles may be made temporarily free as part of a press release or for other promotional purposes, which means that the actual number of Open Access publications may well be below the one given here. Nonetheless, it can be clearly seen that the majority of publications are available online in some way, with the empirical analysis suggesting an estimate of around 62%.

7 Current status of Open Access to research data

We have given an insight into the existing research infrastructure, as well as the places in which research data are managed (see 5 Current status of research infrastructure. On the one hand, the existing research infrastructure distinguishes between data to which access is restricted to members of the respective project in which the data had been created or collected, and data which can be accessed publicly. On the other hand, the results of the questionnaire have shown that not many groups actually make use of these facilities. In particular, it was shown that archiving of both software and primary and secondary data is mostly done in group-internal repositories (see 5.2.8 Archiving). In the following, we will discuss which policies are followed by the with respect to exchanging data and/or making them publicly available, as well as which kinds of data are generally eligible for exchange.

7.1 Policies and limits

The interviews with the groups revealed that exchange of research data is generally done on a per-request basis. This is to say that researchers from other research institutions who have become aware of a certain data set being created or used in a study, typically by reading a paper describing the respective data, contact the authors of the respective paper to ask for access to the data. Although even this is up to now only very infrequently the case, it does happen occasionally. In such cases, groups generally appoint the person responsible for a particular data set with the task of determining whether the data can be made available to other institutions or not. Here, the general rule for granting access is that the requester acknowledges the cooperation in future publications based on this data set. In addition to this, all groups which gave answers to these questions stated that the general rule is that the set of primary data is believed to have been fully analysed. The reason for this is mainly that the amount of financial and human resources that has gone into the creation of primary data is typically too high to just give the data away “for free”. On the other hand, some groups are realistic about the fact that some data sets cannot possibly be fully analysed by a single research group in a reasonable amount of time. Likewise, when asked for conceivable benefits of making primary data available, however, groups tend to see added value in getting additional and even alternative analyses of the same data sets, primarily for reasons of comparability. Only one group (from the SocHum branch) indicated that they make their data available on a public platform, in order to achieve wider (re-)distribution.

Given that the data are believed to have been fully analysed, there is a clear tendency towards providing Open Access to these data. In addition to this, more concrete plans of developing the necessary infrastructure for providing Open Access to research data will be discussed below (see 9 Future developments). However, since this infrastructure had not been put into place at the time this case study was being carried out, the survey focused on the question as to whether it is generally conceivable for groups to make data available, and if so, which kinds of data and to what degree. In particular, this means that groups were asked to specify whether they would make software and primary or secondary data to close colleagues (only), to other research groups or even to the general public. The following section presents the results of this enquiry.

7.2 Willingness to share data

Figure C.27 to Figure C.29 illustrate the willingness to share primary data, secondary data and software respectively. In addition to this, Figure C.30 shows whether groups which share the software they develop do also make the corresponding source code available.

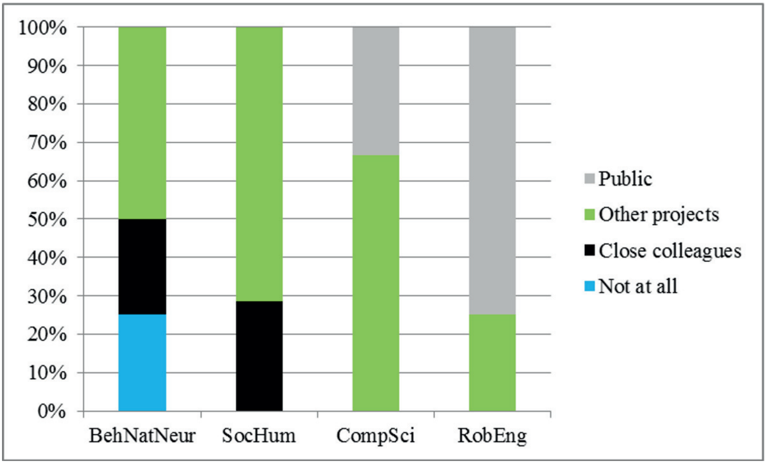


Figure C.27 Willingness to share primary data

7.3 Discussion of results

The results shown above indicate that the willingness to share data beyond a circle of close colleagues differs significantly both between types of data and between types of groups. Starting with primary data, it is clear that groups in behavioural and natural sciences – as well as those in social sciences and

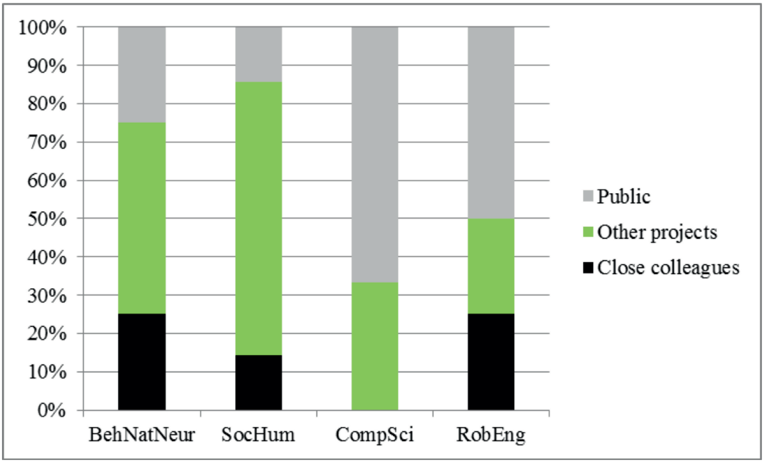


Figure C.28 Willingness to share secondary data

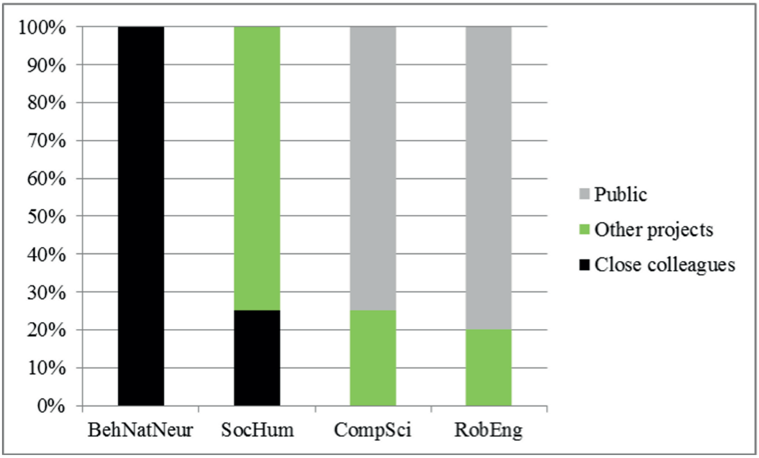


Figure C.29 Willingness to share software

humanities to some extent – are more restrictive than groups in computer science and robotics. In fact, one group in behavioural and natural sciences even indicated that they would not share primary data at all. A straightforward explanation for this is that primary data arising in the former two research branches very often deal with experimental data involving humans. As a result, the free availability of primary data is in many cases not possible from a legal perspective, or at least not desired, which suggests that Open

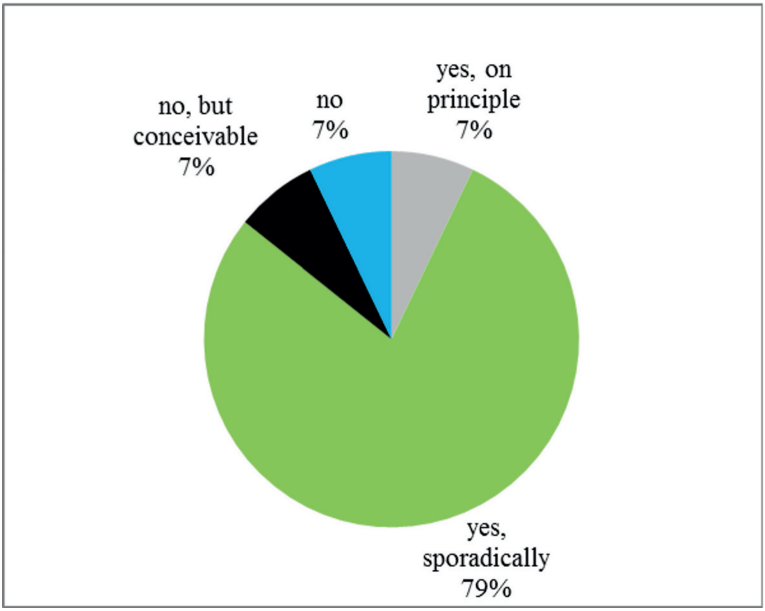


Figure C.30 Willingness to share source code with software

Access to primary data is, at least in these disciplines, a complicated issue. In the latter two research branches, however, primary data are often not created by the group itself – such as in the Semantic Computing group, which frequently uses data available on the World Wide Web – or involves non-human individuals, such as animated characters or robotic devices. As such, legal restrictions like personal rights are less of an obstacle in these disciplines, and the general willingness to share these data is considerably higher. However, potential re-use of the data is limited due to the lack of standard formats in the field. Moreover, one group of the robotics and engineering branch indicated that there are projects whose primary data they would not share at all. The group indicated, however, that this is because the amount of primary data produced exceeds a limit beyond which exchange does not seem reasonable. The situation is slightly different with respect to secondary data. Here, some groups in BehNatNeur and SocHum consider their data to be suitable for being made available to the public, and CompSci groups are less restrictive as to making their data available.

With regard to software, the situation is again very different between the different disciplines. In general, however, the figures seem to suggest that software – at least in those disciplines in which it represents one (if not *the*) primary research output – could very conceivably be shared with the general public. On the one hand, this could be because well-established platforms for sharing Open Source software exist which enable straightforward sharing of software, such as Sourceforge[24] or Google Code.[25] In addition to this, however, it seems to be the case that software as is generally far less in terms of size than, for example, primary or secondary data. In fact, it seems highly unlikely (in fact almost impossible) that software developed within a single project would ever reach the range of terabytes. Given that software is generally written by humans,[26] this would mean an incredible amount of code being created by hand. While software packages including source code, compiled binaries and the libraries on which the software depends may reach the range of several gigabytes, even this is typically not the case. In line with the above findings for primary data, this suggests that size of data may have an influence on the ease and willingness to share data.

On the basis of these figures, we investigated whether the availability of standardised metadata formats for the description of primary and secondary data is correlated with the willingness to make data available. As was mentioned before, 59% of all groups indicated that there are no standardised metadata formats in their field (see 5.2.7 Metadata enrichment). Therefore, we checked for the remaining 41% whether they could conceive sharing primary and secondary data beyond the level of close colleagues. The results are shown in Table C.7.

Table C.7 Willingness to share software and primary and secondary data in SocHum

	Field has standard format	Conceivable exchange of primary and secondary data
BehNatNeur	3	0
SocHum	2	2
CompSci	0	0
RobEng	2	2
Overall	7	4

At first sight, the figures seem to indicate a rather low correlation, which is in fact 0.23. However, when analysing the figures more closely, it becomes evident that this is due to the difference in the behavioural, natural and neural sciences, where none of the three groups that indicated the availability

[24] http://sourceforge.net.

[25] http://code.google.com.

[26] We are ignoring code generators like Apache Velocity since they are not believed to constitute the main part of software development.

of standard formats are willing to share neither primary nor secondary data beyond the level of close colleagues. In line with what has been discussed above, however, this is mainly because BehNatNeur groups generally tend not make primary data available, due to the reasons mentioned before. In fact, two of the three groups indicated their willingness to share secondary data, which would suggest a correlation of 1.0 using this laxer interpretation of "willingness to share".[27]

Summing up the findings in this section, the figures suggest in general that technical disciplines like CompSci and RobEng are less restrictive when it comes to making data available, be it to other projects or to the general public. As was discussed above, this is in part due to the different extensions of the individual types of data in each discipline, with primary data being a primary concern in BehNatNeur and SocHum. Abstracting from individual research branches, Figure C.31 summarises the overall willingness to share data, according to the types of data. As can be seen there, software and secondary data could far more conceivably be made available to the public, whereas primary data – though being conceivable to be shared with other projects – are either unsuitable for general Open Access or would require very flexible licensing schemes.

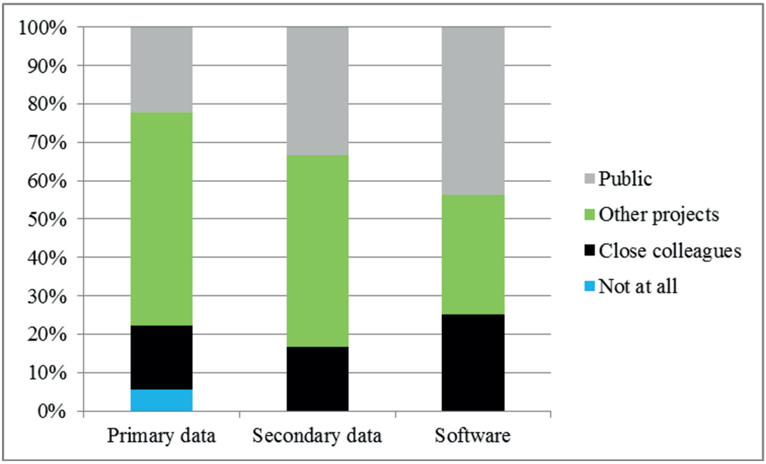


Figure C.31 Overall willingness to share data, according to types of data

[27] Note that we have used a strict interpretation of "willingness to share" – in the sense that willingness to share both primary *and* secondary data was counted as positive evidence only – since the laxer interpretation (i.e. "willingness to share primary *or* secondary data") is true for almost every group.

8 Challenges

In this section, we will discuss the general challenges for an Open Access infrastructure as suggested by the findings above. In general, it has been shown that, in most cases, the data management strategy followed is up to the individual researcher, which results in rather heterogeneous strategies being followed not only between groups, but also within groups. The risk of data sets becoming unavailable due to a researcher leaving the institute is therefore rather high. It should be noted here that this challenge is not solved by creating central repositories alone, since the potential re-usability of a data set is determined by a number of factors. On the one hand, the nature of an experiment or study has a deep impact on re-usability. For example, it seems reasonable to assume that in BehNatNeur, experiments are typically carried out in order to verify a specific hypothesis under very strict conditions. This means that the data collected in such experiments are less likely to be useful in other contexts, which would need to be tested under different conditions. This is certainly different in other disciplines such as CompSci and the data collected there are more likely to be re-used. On the other hand, especially in those disciplines where data are in principle suitable for exchange, it is the degree of documentation of a data set that decides whether it is re-usable at a later point in time or not. In the following, we address different infrastructural challenges with respect to data and publication management.

8.1 Data management

8.1.1 Models for data types, provenance and access rights

Given the variety of data types generated at CITEC, an immediate challenge is to develop models which are capable of representing all aspects of a particular data set. In addition to very general aspects such as type (e.g. audio vs. video), these include the following:

- Given the dependence on proprietary hardware and software identified above (see 5.2.1), it is vital to **document any hardware or software requirements** that need to be fulfilled in order to be able to view or process the data set, as well as other technical aspects like encoding – similar to software package dependencies known, for example, from the popular Linux distribution Debian GNU/Linux.
- Given the guidelines for data sustainability issued by the German Research Foundation, it is necessary to **develop policies and storage infrastructures for short-term, mid-term and long-term archiving** of research data.

- Given the lack of standardised metadata formats in some research areas, it is necessary to **find a reasonable balance with respect to what can be expressed** about a data set, in order to support the re-use of metadata categories whenever possible and thus enhance the interpretability of metadata annotations.
- **Guidelines for ensuring the quality of published data** need to be developed.
- It seems reasonable to **assign data management responsibilities to particular persons** in each group, in order to make sure that all research groups are aware of the available infrastructure.
- It should be possible to **link data sets with publications and vice-versa**, in order to enhance the ways in which both can be explored. Here, it is recommendable to **use technologies and practices developed in the Semantic Web**,[28] in order to ensure that this challenge is addressed in a principled way and achieves appropriate impact.
- Data generated at CITEC poses challenges for storage and backup strategies. For example, given experiments in which 1.5 GB of video data are generated per second, it is vital to have **reasonable backup strategies**. It is understood, however, that this is even more of a requirement in other research areas besides ICT.

The willingness of people to share data with external people, be they researchers involved in other projects or members of the general public, has been discussed above (see 7.2 Willingness to share data). The primary finding was that research in cognitive interaction technology – primarily due to its high degree of experimental work with humans and animals – raises a number of concerns regarding personal rights, and unrestricted Open Access does not seem feasible here. For other cases, excluding those in which access to data is completely impossible due to legal restrictions, it is necessary to have a sound model of access rights to individual data sets – which may even require entirely new licence models, especially with a view on re-usability and modification. Here, it is necessary to encode the provenance of a data set, in order to document its source and development history. As with other challenges mentioned above, this should be approached by making use of available vocabularies as much as possible, in order to achieve interoperability between resources. In addition to this, it is necessary to have a functioning system that implements this model of access rights. As trivial as this aspect may seem, it should be noted that a security leak in the system – or even accidental publishing of confidential data – may have far-reaching legal consequences.

[28] http://semanticweb.org.

8.1.2 Rules, incentives and limits to research data exchange and Open Access

As was just mentioned, Open Access raises legal concerns especially with respect to primary data obtained at CITEC. Moreover, as was mentioned above, the sheer amount of primary data produced may be a limit to exchange in itself (i.e. if the data exceeds an amount at which sharing the data does not seem feasible; see 7.2 Willingness to share data). In addition to this, especially in cooperations with industrial partners, confidentiality agreements have to be signed which restrict the future use of the data in other projects, let alone its publication to the general public. Besides this, however, we have shown that actual data exchange is still performed to a rather limited extent, with exchange upon request, and only after the data are believed to have been fully analysed, being the main policy for data exchange (see 7.1 Policies and limits). It should be noted, however, that researchers admit that it is, in most cases, not possible to say when a specific data set has been fully analysed, while in other cases it is not even possible for a research group to fully analyse a specific data set in a reasonable amount of time. Finally, groups indicated that they expect the amount of maintenance work (e.g. documentation) required to transform a data set into a state in which it can be released to the general public to be very high and the resources that would be needed cannot be allocated – on the one hand due to lack of funding for such tasks and on the other due to lack of scientific reward or appreciation by the community. This is further supported by the analysis presented above (see 5.2.8 Archiving), which showed that only a small number of groups deposit their data on external repositories. Here, a concern was that – given that a data set is, for example, made available to the scientific community but not to the general public – how would it be possible to trace where the data actually end up, after having been downloaded by a large number of people? Finally, it may be possible that experimental approaches will experience a dramatic decrease in the number of probands, because they have to sign very complex data privacy statements. Here, the general trend towards freedom and openness that can be observed on the World Wide Web today faces the desire for more privacy and protection of personal rights at the same time.

On the other hand, many groups expressed the benefits of Open Access to research data. Some of the incentives for Open Access stated by researchers are given below.

- **Increase data transparency**, which would enable researchers, federal agencies or members of the general public, to obtain a better overview of the data generated at a research institution or in a research field.
- **Benchmarking and contrastive analyses** being carried out by different institutions on the same data sets.

- **Support from other institutions** in analysing a particular data set, and thus faster progress in a research field.

8.2 Publication management

In addition to the challenges for data management just discussed, there are a number of requirements on publication management as well. In particular, there is no CITEC-wide publication repository, and publication management is therefore handled very differently not only between research groups, but also within groups (see 5.1.5 Publication management and 5.2.8 Archiving). Therefore, what is needed is, on the one hand, a shared technical infrastructure for depositing publications and, on the other hand, general guidelines and policies regulating deposit and access. In the context of Open Access, we identify the following requirements:

- The interface to the publication deposit process – be it the user interface, application programming interface or web service interface – needs to allow the depositing client to **upload both metadata and the full text of a publication**.
- It should further be possible to **specify the rights (e.g. copyright) that the depositing client is in possession of**, in order to determine whether the client has the permission to set further access rights for the full text of the publication.
- If the client is in possession of the appropriate permissions, the system should allow him or her to **specify the restrictions that possibly apply to the full text**, such as whether it is publicly accessible or only accessible to people belonging to a certain group of users.
- It should further be possible to **determine whether it is permitted to search or crawl the full text and/or metadata of a publication**.
- In order to be able to interoperate with other literature management tools, **metadata should be exportable in several (de-facto) standard formats**, such as BibTeX or Endnote.

Depending on the input by the client on the previous points, the system should then be able to select the appropriate measures for storage and access and allow for flexible search and retrieval. For example, the literature analysis carried out as part of this case study would have been greatly facilitated by being able to search for all downloadable publications of a specific group (or of all CITEC, with results grouped by research group) or for publications which have been written in cooperations between groups. Finally, as was mentioned in the previous section, it should be possible to link publications to research data sets, in order to enhance the information services provided by the system.

9 Future developments

As part of its second funding period, CITEC has very concrete plans for the future development with respect to managing literature and research data, in particular in the direction of linking the two in order to obtain an ecosystem of semantically enriched descriptions of all kinds of research artefacts. First steps into this direction have already been taken and implemented during the course of this case study and the following subsections discuss these current and upcoming developments in more detail.

9.1 Literature management

9.1.1 Interaction with central facilities provided by the university

As was mentioned above (see 5.1.5 Publication management), research groups at CITEC generally take care of literature management themselves, which means that they host and make the descriptions as well as – to some extent – the full texts of the publications authored or edited by members of the group available. Recently, however, the library of Bielefeld University has released the PUB system, a university-wide repository intended to host metadata and full texts of all publications created at Bielefeld University. In order to make use of this repository while still being able to annotate publications with metadata fields not provided by the PUB system, CITEC has developed a module based on the widely used Drupal CMS allowing for a smooth interaction between group-administered publication repositories and PUB. In particular, the module enables the management of a local publication repository which is synchronised with the PUB repository. Here, the module ensures that the local repository always contains at least the group-relevant publications available in PUB, with the possibility of containing additional publications not available in PUB. This concerns, for example, those items which have not been published yet and whose descriptions are therefore not complete yet. Even though PUB provides way for handling such cases as well, authors may prefer not to expose their manuscripts on the university-wide repository until they have been published. In addition to this, the module allows for attaching the aforementioned additional metadata descriptions to a group's publications, which will be described in more detail below.

9.1.2 Semantic enrichment

CITEC is taking concrete steps towards annotating the locally stored publications with additional metadata. On the one hand, this concerns the use of standard schemas for the description of bibliographic entities, such as Dublin

Core.[29] On the other hand, however, CITEC aims at making the descriptions not only useful for human users navigating to a group's website, but also for machines harvesting the website for information. Here, formalisms developed in the context of the currently evolving Semantic Web, such as the *Resource Description Framework* (RDF) or the *Web Ontology Language* (OWL), as well as formal models for representing bibliographic entities by means of these Semantic Web formalisms are of particular interest. In addition to those established, this concerns the analysis and exploration of bibliographic ontologies currently under development, such as the *Semantic Publishing and Referencing* (SPAR) ontologies,[30] which includes Semantic Web versions of established bibliographic models like the *Functional Requirements for Bibliographic Records* (FRBR).[31] Such ontologies are particularly interesting since they provide a rich vocabulary that not only allows for a formal representation of bibliographic entities, but also of the relations between them. Beyond the usual citation relation, this concerns relations such as *usesDataFrom* or *disagreesWith*. It is clear to see that having such relations between entities would greatly enhance the ways in which publications could be queried not only by humans, but also by machines. Here, current development focuses on the integration of such descriptions into the aforementioned module in order to provide such enhanced services.

9.2 Data management

In addition to the management of literature, CITEC has recently launched a research data management task force involving the leaders of several research groups as well as members of the university library and the Collaborative Research Centre 673. The goal of this task force is to design and implement a strategy for achieving sustainability and reusability of all kinds of data created at CITEC. In the first instance, this development is concerned with providing an appropriate framework for storing the data in a way that enables a smooth integration into the existing research infrastructure as explained before (see 5 Current status of research infrastructure) and implements the necessary procedures for enabling Open Access to the data. A second development phase deals with providing suitable vocabularies that allow for a fine-grained description of all aspects of the data, as well as interlinking with other descriptions, such as those of literature already mentioned. The final phase of this development then deals with aspects of making the data available. Here, planning has begun on extending the already existing OpenSource

[29] http://www.dublincore.org/documents/2010/10/11/dces.

[30] http://purl.org/spar.

[31] http://www.ifla.org/files/cataloguing/frbr/frbr_2008.pdf.

server to an OpenData server[32] that on the one hand provides direct access to the data and on the other hand enables access to metadata descriptions by metadata harvesters like CLARIN via the Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH). As a result, the vision of research data management at CITEC is ultimately an open one, where all kinds of research artefacts created at the institution – including literature and data which do not affect personal rights – are made available to the research community as well as the general public.

10 Implications for Open Access infrastructure

10.1 Technical implications

- **The diversity of data types** arising even in individual experiments on a single research topic requires mechanisms that allow for linking heterogeneous data types in a way that allows flexible and intuitive exploration.
- **The amount of data** being generated requires the storage infrastructure to be able to deal with data in very large quantities and sizes.
- **The dependence on non-standard formats and proprietary software,** including non-free operating systems needed for the operation of specific research instruments entails a number of issues like backward (in)compatibility, maintenance and licensing that require exact specifications, for example, which software version is needed in order to be able to process the data file in the intended way. These need to be stored and linked with the data in order to make the data re-usable.
- **Privacy issues** of experimental primary data involving humans, as well as data arising in cooperations with industrial partners, pose special requirements on the security of the data, as misuse or accidental release can have far-reaching legal consequences.

10.2 Scholarly implications

- **Fine-grained licensing schemes** regulating access, re-use, linking, manipulation and redistribution of research data need to be developed, as current schemes cannot handle critical cases, for example where anonymised primary data lose their anonymity by other data sets linking to them.

[32] At the time of writing, the CITEC OpenData server has officially gone live at http://opendata.cit-ec.de and published the first freely available data set of manual interaction data.

- **Rewarding and acknowledgement schemes for data creation, curation and publication** need to be developed and established, as these tasks typically take up much more time and effort than, for example, the creation of a scientific article, while they are at the same time not recognised as indicators or measures of quality of research as the latter.
- **Rewarding of golden Open Access publications** in order to establish it as a recognised means of publication.
- **Institutional, disciplinary and/or funder-driven** guidelines and policies for data exchange need to be established in order to provide a framework and incentives for data exchange.
- **Advertising the availability and benefits of the infrastructure** in a way that allows researchers from less technical fields to know what is available and where to find it.
- **Educational support in using the infrastructure** so that researchers not only find available services, but also know how to use them and benefit from them.
- **Funding for designated resources** dealing with data management issues, since data curation is currently done at a subjective level, instead of being a designated part of the general research agenda.

11 Acknowledgements

We would like to thank all research groups who have taken the time to participate in this study, as well as Jochen Steil and Sven Wachsmuth for their helpful comments and suggestions.

12 List of figures

13 List of tables

D e-Infrastructures Area

Leonardo Candela, Akrivi Katifori and Paolo Manghi

1 Introduction

Quoting the e-Infrastructure home page[1] of the FP7 ICT Research Unit of the European Commission:

"The e-Infrastructures activity, as a part of the Research Infrastructures programme, focuses on ICT-based infrastructures and services that cut across a broad range of user disciplines. It aims at empowering researchers with an easy and controlled online access to facilities, resources and collaboration tools, bringing to them the power of ICT for computing, connectivity, storage and instrumentation. This allows for instant access to data and remote instruments, 'in silico' experimentation, as well as the setup of virtual research communities (i.e. research collaborations formed across geographical, disciplinary and organizational boundaries)."

In other words, e-Infrastructures support research infrastructures from the "virtual" perspective, by enabling community "actors" (researchers or their applications) to exchange their "resources" (research data and literature) by means of a controlled, regulated, digital environment. Specifically, researchers in the field of e-Infrastructure investigate solutions and methodologies enabling and facilitating the realization of e-Infrastructure platforms capable of supporting the activities of domain-specific research communities (e.g. Agriculture, ICT, Social Sciences). In general, e-Infrastructures can be considered as a combination of (i) established policies, standards and best practices and (ii) a set of technologies and tools, which together support an environment where researchers of a given domain can accomplish their daily activities in a collaborative and synergic fashion (Atkins et al., 2003; Ioannidis et al., 2005).

The main purpose of this chapter is to report how researchers investigating in the area of e-Infrastructures organize their activities of "data and publica-

[1] http://cordis.europa.eu/fp7/ict/e-infrastructure.

tion management" and themselves rely on research infrastructures to do so. Due to the early age of this field and its rather multidisciplinary computer science character, no well-established research infrastructure is available and researchers tend to follow "infrastructure-flavoured" solutions local to their organizations. As a consequence, the authors of this chapter (from the D-Lib research group at CNR, Italy and the MADGIK research group at the University of Athens, Greece) opted to approach this study by collecting a number of experiences from relevant stakeholders in the field in order to identify "local infrastructure" commonalities and "research infrastructure" desiderata.

We shall first elaborate on the strategy adopted to run this investigation, based on questionnaire-driven interviews to a number of representative organizations in the field. Subsequently, we shall present the specific case narratives, before finally drawing a summary of the current status and elaborate on possible future challenges.

2 Methodology and representativeness of the study

In order to investigate on the current status and future challenges of research infrastructures in the area of e-Infrastructure, we adopted a methodology based on questionnaire-driven interviews to experienced researchers in the field. The questionnaire[2] contains a structured set of the questions, which we perceived as crucial to gather the information necessary to gain in depth understanding of the research workflow lifecycle at the researcher's group or organization. Crucial is the distinction between literature and data, where issues such as management, exchange and Open Access are somewhat more cross-domain and mature for scientific publications and heavily domain specific and not as thoroughly investigated for research data. In the process, we collected a list of "community desiderata", intended as current issues and/or envisaged solutions which interviewees believed could contribute to improve the overall research activities of the community. The general outline of the questionnaire is the following, concentrating on four main question groups:

- **Research group profile:** general information about the research group, interests and available service and computing infrastructures.
- **Research data:**
  - **data and metadata typologies:** information on which kinds of research data the organizations deals with and which kind of metadata formats are used to describe research data.

[2] https://spreadsheets.google.com/viewform?hl=en&formkey=dGN4bnp1QWJONkdXZ3FRbEtmb2tlZ2c6MQ#gid=0.

    - data in this field are mostly *software* (source code), *software instances* (software in execution, also known as “process”), *benchmarks* (domain-specific research data collections or corpora used to validate software instances), *logs* (recorded history of actions or events, typically used to evaluate and monitor the activity of a software instance) and *statistics* (often derived from logs to evaluate software instance activities).
    - metadata can be “structured”, i.e. machine interpretable and consumable records/profiles or “unstructured”, i.e. documentation such as user manuals, specifications, installation guides, in any format (wikis, websites, document files).
  - **data lifecycle:** information on how data and metadata are produced, processed and stored.
  - **data management aspects:** information on aspects such as data and metadata versioning, provenance and preservation.
  - **data exchange:** information on how data and metadata are exchanged by researchers internally and externally to the organization.
  - **data and Open Access:** information on the awareness and status of application of Open Access principles to research data within the organization.
- **Literature**
  - **literature management:** information on the publication lifecycle established at the organization, from survey, drafting and publishing of literature.
  - **literature and Open Access:** information on the awareness and status of application of Open Access principles to publications within the organization.
- **Combination of literature and research data:** information on the awareness and status of application of literature and data interlinking within the organization.

Based on the questionnaire, we arranged interviews with a selection of key stakeholders in the European domain of e-Infrastructures. Our strategy has been that of selecting a set of organizations and individuals which are representative of wider classes of research institutions and companies, with respect to the size of the organization and research scopes. As e-Infrastructure is a rather new and multidisciplinary topic, the selection criteria cannot aim at providing a full coverage of the methodologies and research aspects carried out in the field. However, we believe the adopted perspective allows one to gain an adequate view of the European status for this novel research field.

More specifically, we approached research institutes (D-Lib Research Group, National Documentation Center and Greek Research & Technology Network (GRNET)), universities (MADGIK Research Group) and private companies (Agro-Know and Engineering R&D Unit on Clouds and Distributed Computing Infrastructures). In the following, section 3 Case narratives presents the information collected in the interviews, section 4 Current status synthesizes the interviews and reports on the current status on research infrastructures for e-Infrastructures, while section 5 Desiderata and future directions concludes the chapter elaborating on researchers desiderata and identifying future challenges to address them.

3 Case narratives

In the following sections we present the summary of the interview for each organization. For each case narrative, we provide:

- general information about the organization, which includes allocation of people over research activities and a description of its local service and computing infrastructures;
- a description of the organization research objectives and projects;
- a description of the organization's typical workflow in the production of literature and data.

3.1 D-Lib research group

3.1.1 General information

The D-Lib research group, led by Dr. Donatella Castelli, consists of around five researchers, 15 technicians and three administrative staff. It is part of the Networked Multimedia Information Systems Laboratory (NeMIS), which consists of 48 researchers and technicians conducting research and development activities on algorithms, techniques and methods for information modeling, access and handling, as well as new architectures and system services – P2P and Grid-based (Foster and Kesselman, 1999) – supporting large networked multimedia information systems. The NeMIS laboratory is in turn part of the Institute of Information Science and Technologies (ISTI) of the Italian National Research Council (CNR), which is organized in 16 laboratories and is committed to producing scientific excellence and playing an active role in technology transfer.

Organization of activities D-Lib group research activities are organized in two parallel tracks: research subjects and projects. Each research subject is managed by one researcher and is assigned a group of co-researchers and technicians to address prototypes and products releases; both researchers and technicians can be assigned to multiple branches. Each project is assigned to one researcher, who becomes responsible and ISTI representative for the project, and generally involves one or more research subjects. In order to serve the project needs, the project responsible is also in charge of coordinating the researchers in charge of the individual subjects to accomplish the project objectives.

Computing infrastructure In order to accomplish research and development tasks, researchers are equipped with personal workstations and can count on a shared computer infrastructure, offering a central processing unit (CPU) cluster equipped with a separate storage area network as described in Table D.1

Table D.1 D-Lib computing infrastructure

CPU	Cores: – 10 × dual AMD Opteron Processor 252 (no hvm) – 2 × dual Quad-Core AMD Opteron Processor 2356 – 2 × dual Six-Core AMD Opteron Processor 2427 – 2 × dual Quad-Core HT Intel(R) Xeon(R) CPU E5630 – 2 × single Dual-Core AMD Opteron Processor 1222 – 2 × single Quad-Core IntelQ6600 – + other miscellaneous hardware: total Total: 88 cores (104 cores considering hyper-threading) Total: 516 GB ram on the cluster
Storage	Protocol: SCSI, SAS, SATA Disks: 42 drives, raid1 pairs, effective 5.7 Tb
Storage area network	Protocol: AoE Disks: 16 sata drives, raid1 pairs, effective 7.2

3.1.2 Research objectives and projects

The team focuses on the following research and development activities regarding the realization of sustainable e-Infrastructures for research:

- foundations and data models of digital libraries;
- digital library management systems: design and realization of systems for the construction of digital library systems (Candela et al., 2008);

- data management and curation services: e.g. authority file management, bulk-data feature extraction and transformation, time-series management, compound objects management (DRIVER-II project[3]);
- design and development of frameworks (middleware): enabling large-scale data infrastructures (D-NET software Toolkit[4] and gCube Toolkit[5]);
- Cloud services: (Dikaiakos, Katsaros, Mehra, Pallis and Vakali, 2009), service on-demand frameworks providing abstractions over different Cloud platforms (VENUS-C project[6]);
- design and development of virtual laboratories or virtual research environments: in the context of large-scale data infrastructures (D4Science-II project[7]);
- foundation elements of "global" infrastructures and "ecosystems" of infrastructures: (see GRDI2020 project[8]).

The team has been involved in many EU-funded projects relevant to the topics of e-Infrastructures, namely:

- **FP6 projects:** DILIGENT (no. 004260, Scientific Coordinator) – see project description in 3.5 MADGIK research group – BELIEF (no. 026500) and DRIVER (no. 034047).
- **FP7 projects:** EFG (no. 517006), DRIVER II (no. 212147), D4Science (no. 212488), BELIEF II (no. 223759), D4Science-II, HOPE, VENUS-C, GRDI2020 and OpenAIRE.

Among these, the most relevant and still ongoing are:

- **DRIVER Targeted Project (IST FP6) and DRIVER II CP/CSA (INFRA FP7):**[9] DRIVER is a multiphase effort whose vision and primary objective is to create a cohesive, robust and flexible pan-European infrastructure for digital repositories. DRIVER has established a network of relevant experts and Open Access repositories. DRIVER-II aims to consolidate these efforts and transform the initial test-bed into a fully functional, state-of-the art service, extending the network to a larger confederation of repositories
- **OpenAIRE**:[10] OpenAIRE aims to establish and operate a data infrastructure for connecting EC FP7 projects with the scientific publications funded under such projects. The infrastructure allows the Commission

[3] http://www.driver-community.eu.
[4] http://www.d-net.research-infrastructures.eu.
[5] www.gcube-system.org.
[6] http://www.venus-c.eu.
[7] http://www.d4science.eu.
[8] http://www.grdi2020.eu.
[9] http://www.driver-repository.eu.
[10] http://www.openaire.eu.

and organizations participating to EC project to measure the impact of the Open Access mandates (Clause 39) across FP7 projects in several research areas. The group is responsible for the realization of the enabling layer of the infrastructure (core infrastructure services: e.g. information service, orchestration services) and for the data management and curation part.

- **D4Science CP/CSA (INFRA FP7) and D4Science II CP/CSA (IP FP7):**[11] D4Science and its continuation, D4Science-II, is a European e-Infrastructure project, co-funded by the European Commission's Seventh Framework Programme for Research and Technological Development. D4Science-II will develop technology and methodologies that will enable sustainable interoperation of multiple, diverse and heterogeneous data e-Infrastructures that have been established and are currently running autonomously, thereby creating e-Infrastructure ecosystems that can serve an expanded set of communities dealing with complex, multidisciplinary challenges whose solution is beyond reach with existing resources. Furthermore, D4Science-II will use the existing D4Science e-Infrastructure as a hub to bring and hold together several established scientific e-Infrastructures and, thus, set up a prototypical instance of such an e-Infrastructure ecosystem. The group is responsible for the realization of the enabling layer of the infrastructure (core infrastructure services: e.g. information service, orchestration services) and for the data management and statistics part.

Research data With respect to research data, the team produces open source software, software instances, technical websites, logs and test results, with related benchmarks. In particular, software is produced by adopting rigid programming policies, from development and testing to integration and production.

Researchers and technicians store their data relying on a local service infrastructure integrating tools such as TRAC (road maps and tickets), SVN (software versioning), BSCW (document and calendar sharing) and wikis, made available across several projects to a pool of "single sign-on" authorized users.

Software data, when possible, are searched and fetched from well-known software web sources (e.g. SourceForge, Google projects, Apache projects) and re-used as part of the resulting products. Similarly, the team may contribute to the open source community.

Software instances are also regarded as available and exchangeable research data. In this context, a software instance is a *service*, i.e. running instance of

[11] http://www.d4science.eu.

software accessible through the web, made available for access by authorized consumers through a service-oriented infrastructure.

Structured metadata formats for research data are mainly proprietary (e.g. software and software instances) and may change depending on the infrastructure implementation. For example, services are described by metadata properties (obliged to include the URL of the service) which enable its discovery based on given criteria and subsequence usage. Such metadata are typically proprietary and target the requirements of service consumption raised by the application domain. In other cases, for example documentation (see unstructured metadata below), metadata formats are imposed by the specific tool's default (e.g. BSCW for technical reports).

Unstructured metadata are continuously produced to support the software lifecycle (e.g. specifications, software documentation, user and installation manuals, websites) and to describe software results or applications (e.g. white papers, technical reports).

Desiderata: most of the software products (research data) in the literature are prototypes and therefore available only through organizations, groups or researchers' websites. As such, they cannot be easily discovered, located and re-used. A community e-Infrastructure serving the purpose of software and documentation sharing would ultimately benefit the community, by guaranteeing standard metadata descriptions, collaborative development and degrees of quality certification.

Literature The team is very active on publication production, as it considers it an important mean of dissemination. The survey phase of publication is typically carried out relying on known publication sources, such as Google, Google scholar, Wikipedia and publisher websites (e.g. Elsevier, ACM, IEEE) and less known but specific sources, such as the DRIVER infrastructure. The phase of publication drafting is typically carried out by physical meeting and multi-hand editing, using shared editors such as Google docs and file-sharing tools such as Dropbox, BSCW and email.

It is mandatory for researchers at ISTI to upload publications metadata and full text, with proper access policies, into the PUMA-ISTI repository.[12] Through PUMA, publications are made available to Google Scholar or other aggregators, such as the DRIVER infrastructure and BASE.

Desiderata: there is no web source focusing on scientific publications on e-Infrastructure research. Relevant results in the field are to be discovered with parallel searches across several websites and cumbersome refinement and skimming cycles, often by reading the article abstracts or full text. A community e-Infrastructure serving the purpose of sharing e-Infrastructure

[12] PUblicationMAnagement, http://puma.isti.cnr.it.

literature would ultimately benefit the community, by guaranteeing standard metadata descriptions and tailored focus.

Desiderata: there are no conferences or journals focusing on e-Infrastructure research. Only a few conferences, such as TPDL (formerly ECDL) or IFLA have "special tracks" dedicated to the topic. Most submissions of scientific publications are therefore sent to conferences and journals whose main topic "touches" that of this research, i.e. service-oriented architectures, digital libraries, knowledge management, Grid (Foster and Kesselman, 1999), etc. In some cases, conferences and journals specific to the application domain of a given e-Infrastructure may also accept submissions of "methodological" papers. The domain of e-Infrastructure has reached sufficient maturity to deserve special venues and classification in the computer science world.

Combining literature and data The group always refers from publications the websites of products cited in the narration. However, this practice follows common sense rather than given policies. Although it would be desirable in many cases, the team is not aware of any best practices or tools for managing or providing combinations of literature and data.

Open Access The team is well aware of Open Access mandates, as it works on projects such as DRIVER and OpenAIRE which are trying to advocate and promote its adoption across Europe and beyond. In particular:

- **research data:** data are stored within ISTI infrastructure and not made openly available to third party consumers, which on request can be granted access to the data, i.e. Open Access policies are figured out case by case. Exceptions are made for software data, which are open source (hence Open Access) and directly available from the product websites;
- **literature:** researchers, when having to choose between equivalent publication venues, tend to prefer those supporting Open Access policies. Unfortunately, most relevant forums in the fields often rely on publishers that do not support Open Access rights.

3.1.3 Research workflows

The typical research production workflow of the team consists of the following phases:

1. problem identification, based on experience and intuition;
2. survey of the literature and data (software, documentation, reports) to find similar or useful (i.e. reusable) resources and "certify" the validity of the intuition;

3. design of a solution, possibly reusing existing data (e.g. software);
4. production and maintenance of unstructured metadata (e.g. software documentation, installation guides, roadmaps, technical reports);
5. development of prototype;
6. definition of benchmarks and testing;
7. release of a product;
8. publication writing and publishing.

Such steps are accomplished by exploiting the local service and computing infrastructure available at ISTI in combination with the above mentioned web tools for discovery, collaborative production and sharing of literature and data.

3.2 Agro-Know

3.2.1 General information

Agro-Know Technologies[13] is a new research-oriented enterprise that focuses on knowledge-intensive technology innovation for agriculture and rural development. The company focuses on realization of systems and services for organization and delivery of agricultural knowledge, promoting the usage of semantic web technologies and Web 2.0 tools. It also explores their deployment and testing in application domains such as education and training, commerce and public administration.

Agro-Know spun off from a group of researchers working in R&D projects in GRNET SA[14] (Greek Research & Technology Network) and today counts 15–20 employees, assigned to research and innovation, design and development activities.

Organization of activities Agro-Know is internally organized in three research teams of about five people, where one or two members are dedicated to software development. In parallel with the research teams, the company has a technical development team, led by one technical coordinator, whose purpose is to support cooperation and sharing of resources among the research teams.

Computing infrastructure The company supports an intranet connecting workstations of researchers and developers, plus common servers for file sharing. In many cases, research teams rely on computing infrastructures provided by the organizations they cooperate with or they work for.

[13] http://www.agroknow.gr.

[14] http://www.grnet.gr.

3.2.2 Research objectives and projects

The main e-Infrastructure research objectives of the company are:

- e-Infrastructures for agricultural research data;
- e-Infrastructures for museums of Natural History with extensive content on biodiversity, botany, etc.;
- e-Infrastructures for education;
- repository platforms adaptable to diverse application scenarios.

Agro-Know gives special emphasis in understanding the needs of the user communities they work with. They feel that a lot of interesting e-Infrastructures research issues can be identified through efficient observation of the user practices, the in-depth understanding of the problems they face and the support that the researchers need in their everyday work.

Among the projects that the Agro-Know team has been or still is involved are:

- **Organic.Edunet:**[15] a multilingual federation of learning repositories with quality content, which support the awareness and education of European Youth about topics related to Organic Agriculture and Agroecology;
- **Natural Europe:** an integrated effort to make knowledge residing in a vast array of Natural History Museums (NHMs) commonly accessible. Accessibility means that the impressive abundance of high-quality digital content is pedagogically structured and presented to the consumer in personalized and contextualized ways;
- **ARIADNE:**[16] an infrastructure of a distributed network of learning material repositories.

Research data Agro-Know creates and processes data such as software, system logs and analytics described by structured and proprietary metadata and by unstructured metadata (e.g. documentation, XML/RDF data models, websites).

The data are produced on the workstations (or private laptops) and then stored for sharing and exchange on the local computing infrastructure through version systems (e.g. Git).

Research data are often exported through project websites (e.g. software, technical reports).

Desiderata: privacy policies at different organizations have hindered reuse and publication of log-file data. An e-Infrastructure for research data in this area could also impose common protection policies and access protocols and ensure these are respected by participating organizations.

[15] http://portal.organic-edunet.eu.

[16] http://www.ariadne-eu.org.

Literature Researchers survey and share the literature through Google Scholar and Mendeley, but in general no collaborative tool (e.g. Google Doc-like) is used. Publications are mostly drafted on workstations (and private laptops), exchanged by email and eventually stored within the company's file server folder structure. However, when drafted in collaboration with external research teams, web tools such as BSCW, Dropbox, Google Docs and wikis may be adopted.

Desiderata: researchers find difficult to share their bibliography, i.e. to exchange their references in a meaningful and organized way. An e-Infrastructure for literature in this area could offer to researchers in the field services for ensuring controlled sharing of publications and bibliographies.

Combining literature and data Researchers at Agro-Know are not aware of publishers that allow the combination of literature and data nor of policies and best practices that would enable such combination.

Desiderata: although the benefits of this approach are evident, based on the experience at the company their application may encounter the issues of:

– metadata: standard representation formats for most of the data do not exist;
– privacy issues: in the case of log files the publication of the datasets may not be possible due to privacy laws/policies;
– unavailability: some data are not available for external referencing, i.e. not available through the internet, e.g. logs on a server.

Open Access issues Agro-Know supports and promotes Open Access. In particular:

– literature: researchers favour publishers supporting Open Access. When possible, publications are public in the project websites and also on the company website as a draft with a link to the editor site;
– research data: the software produced by the company are made available as open source and the educational material with a Creative Commons licence.

Desiderata: researchers believe it is crucial that funding agencies impose Open Access for the results of the projects they fund. As a side effect, this would push publishers at finding new business models.

3.2.3 Research workflows

The general workflow employed in each of the Agro-Know projects is the typical specification, design, development and documentation and evaluation cycle:

1. understanding and defining: describing the objectives of the project along with the partners that may be involved;
2. requirement analysis: producing requirement analysis documents by close interaction with the recipients of the technology to be delivered;
3. design: producing functional and architectural specifications of the technology to be developed, in strict collaboration with the recipients;
4. development: implementation of the technology based on the given specifications. Developers tend to re-use, adapt and customize core technology developed at Agro-Know and to re-use third-party open source software;
5. documentation: in parallel to development, researchers focus on technical reports or publications writing in collaboration with the technology recipients (e.g. user communities) and with project partners;
6. testing and deployment: after strict testing and evaluation, the technology is released and put into production. The underlying software is made available openly to the public, unless project copyright obligations are involved.

3.3 National Documentation Center (EKT)

3.3.1 General information

The National Documentation Centre[17] (EKT) is the Greek national infrastructure for scientific documentation, online information and support services on research, science and technology. The Centre was founded in 1980. It is integrated with the National Hellenic Research Foundation (NHRF) and is supervised by the General Secretariat for Research and Technology of the Ministry for Development.

EKT is both a major e-Infrastructures developer in Greece and one of the main providers for science and technology services and content, as it operates, among others, the Science and Technology digital library, including the digital library of Greek PhD theses.

Organization of activities EKT operates as partner of several projects and to each of them it assigns one coordinator supported by a research team. Research teams may share members and operate over more than one project. EKT elects one of the project coordinator as research supervisor of all projects, in order to maximize re-use of resources and collaboration.

EKT also undertakes close collaborations with external research teams. The most relevant experiences are with the institutes of the National Hellenic

[17] http://www.ekt.gr.

Research Foundation (the Pandektis project[18]) and with GRNET, in the context of GÉANT project.[19]

Computing infrastructure The EKT computing infrastructures is described in Table D.2

Table D.2 EKT computing infrastructure

CPU	– Virtualization platforms comprising 8 servers, 64 processing cores, 192 GB of memory in high availability configuration – 77 physical and virtual CentOS Linux, Redhat, Windows 2003 and Sun Solaris servers – 36 high-end 64-bit Intel Xeon, AMD Opteron and Solaris SPARC physical servers
Storage	– Storage Area Networks, coupled with 5 FC switches providing 83 TB of raw disk space – LTO3 and LTO4 tape libraries with 156 TB raw capacity
Storage area network	– Fully redundant IP network featuring no Single Points of Failure, Gigabit Ethernet end to end, redundant 1 Gbps firewall, border/core router configuration, VPN – Active Directory/LDAP infrastructure, high capabilities work stations, Gigabit Ethernet until the end user

3.3.2 Research objectives and projects

EKT research teams have expertise in the following research topics and activities:

- aggregation of heterogeneous resources;
- Open Access infrastructures;
- websites;
- digital library technologies;
- repository platforms;
- digitization;
- organizing national and international working groups for thematic studies to produce best practice or policy documents.

In particular, EKT participates and in some cases coordinates several research projects, both European and national. Those related to e-Infrastructures include:

[18] http://pandektis.ekt.gr.

[19] http://www.geant.net.

- **EuroRIs-Net and its continuation EuroRIs-Net+:** EuroRIs-Net is a coordination action supports the network of national contact points for Research Infrastructures;
- **OpenAIRE[20] project (EC FP7):** see project description in 3.1 D-Lib research group;
- **Pandektis:** Pandektis aimed to provide free access to 11 integrated and scientifically elaborated collections produced by the three humanistic Institutes of the National Hellenic Foundation for Research: Institute of Greek and Roman Antiquity, Institute of Byzantine Research and Institute of Neohellenic Research;
- **Argo:**[21] Argo aimed at realizing an environment which facilitates Open Access and search across bibliographical information resources available in Greece as well as abroad.

Research data Software is the main forms of data that EKT produces, together with unstructured metadata in the form of technical reports. Data and unstructured metadata are stored for internal sharing between the research teams in common file servers at EKT, with different access rights for different groups of users and over different projects. For software, a version control management system is used, as well as issue tracking (Mantis), while technical reports are drafted collaboratively as wikis.

Data exchange with groups of other organizations is accomplished mainly through the project websites.

Research Literature Researchers survey the literature through Google Scholar[22] and Scopus[23] and manage references using CiteULike.[24] For collaborative drafting they use SVN. Finally, preferred venues for publications are conferences such as TPDL (formerly ECDL) and IFLA and journals related with the topic of interest. Interestingly, some PhD theses have been followed in cooperation with universities and research centres.

Publications are made available for web search and access through the Helios[25] repository, realized at EKT.

Combining literature and data Combining data and literature is considered a good practice at EKT. On the other hand, the absence of best practices and tools available to support it does not make it an option.

[20] http://www.openaire.eu.

[21] http://argo.ekt.gr.

[22] http://scholar.google.com.

[23] http://www.scopus.com.

[24] http://www.citeulike.org.

[25] http://helios-eie.ekt.gr/EIE.

Open Access issues EKT is one of the first organizations in Greece to actively adopt and promote Open Access and one of the first to sign the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. It is the creator and owner of the main website for Open Access in Greece,[26] which provides information on best practices, policies and existing repositories that have adopted Open Access, for example.

3.3.3 Research workflows

EKT adopts clearly defined procedures for research and development, specifically:

1. requirement analysis: producing requirement analysis documents by close interaction with the customers;
2. design: Producing functional and architectural specifications of the technology to be developed;
3. development: Implementation of the technology based on the given specifications, possibly reusing EKT software. Progress is monitored by the project coordinator and by the EKT research supervisor;
4. documentation and publications: In parallel to development, researchers focus on technical reports or publications writing in collaboration with the technology recipients (e.g. user communities) and with project partners;
5. testing and deployment: After strict testing and evaluation, the technology is released and put into production.

3.4 Greek Research & Technology Network (GRNET)

3.4.1 General information

GRNET SA[27] operates the Greek Research & Technology Network, according to the operating model described by the EU Research and Education Networks. It operates both at a national and international level and constitutes the setting for the development of innovative services for the members of the Greek research and education communities. GRNET SA connects more than 90 institutions, including all Greek universities and technical and research institutes, as well as the public Greek School Network, supporting more than 500,000 users all over the country. Moreover, it provides local interconnection services to the main Greek Internet providers, through the Greek Internet Exchange/GR-IX[28] infrastructure. GR-IX started operating

[26] http://openaccess.gr.
[27] http://www.grnet.gr.
[28] http://www.gr-ix.gr.

in 2008 and provides interconnection at Nx10 Gbps, enhancing the quality of internet service and infrastructure nationwide.

Organization of activities GRNET's technical personnel are organized in research groups, which in turn can be assigned to one or more projects. Researchers and developers can participate to and collaborate with several groups and projects, for both publication writing and software development activities.

Furthermore, GRNET collaborates via EC projects with major European institutes that work on infrastructures, such as CERN.

Computing infrastructure GRNET's computing infrastructure is presented in Table D.3. Occasionally, the activities may require Cloud (Dikaiakos et al., 2009) resources rental, to acquire CPU and data storage capabilities on demand.

Table D.3 GRNET computing infrastructure

CPU	– 26 servers – 512 cores
Storage	– 200 TB storage

3.4.2 Research objectives and projects

The main research topics at GRNET are:

- e-Infrastructures for research infrastructures;
- Grid solutions (Foster and Kesselman, 1999);
- service Cloud solutions;
- access to digital content.

Among the projects that GRNET has participated in are:

- **StratusLab:**[29] StratusLab is developing a complete, open-source Cloud distribution that allows Grid and non-Grid (Foster and Kesselman, 1999) resource centres to offer and to exploit an "Infrastructure as a Service" Cloud. It is particularly focused on enhancing distributed computing infrastructures such as the European Grid Infrastructure[30] (EGI).

[29] http://stratuslab.eu.

[30] http://www.egi.eu.

- **Organic.Edunet:**[31] Organic.Edunet had as its aim to facilitate access, usage and exploitation of digital educational content related to Organic Agriculture (OA) and Agroecology.

Research data GRNET researchers deal with research data such as software, virtual machine images/appliances, websites and system logs. These are often accompanied by unstructured metadata in the form of manuals, documentation and technical reports. Data and metadata are stored and archived in server storage devices private to the groups. Unstructured metadata are often in Latex and multi-hand drafted with the support of a version control system. Similarly, software is organized and managed through version control systems.

As for metadata, GRNET tends to use proprietary formats for software and currently is designing metadata standards for virtual machines in collaboration with external groups (Dublin Core model and RDF encoding).

Data exchange between members of the group and across several groups is made possible through wikis, which are used as structured and organized directories to the data files. In general, data are open for others to use, except when external collaborators require a non-disclosure agreement.

Desiderata: exchanging research data with external groups in different projects is made difficult by the adoption of different version control systems. An e-Infrastructure for this research community may offer services for storing and sharing research data based on common formats and policies to be adopted as standards by the community.

Research literature GRNET researchers focus more on software development than on publication writing. As such publication management is not accomplished through specific tools. When surveying and drafting Google and Mendeley might be used to search publications and manage references. Researchers mostly publish at conferences and journals.

Combining data and literature Combining data and literature would be considered very useful but is not yet an option as there are no best practices to follow or wide-spread tools available to support it.

Open Access issues GRNET is aware of the advantages of Open Access policies, but is not pursuing them actively. Specifically:

- literature: researchers do not prefer Open Access publisher to others and do not invest in buying Open Access licences;

[31] http://www.organic-edunet.eu.

- research data: software data are usually available as open source (e.g. Apache 2 licence) and technical reports are available with a Creative Commons licence.

3.4.3 Research workflows

The GRNET e-Infrastructures team focuses mostly on software development, less on publication writing, but does not implement strict development procedures. To achieve its objectives, the team exploits collaboration tools, both for software development and technical report writing, and adopts design and development methodologies that may vary from project to project.

3.5 MADGIK research group

3.5.1 General information

The Management of Data, Information, & Knowledge Group[32] (MADGIK), led by Prof. Yannis Ioannidis, is part of the Department of Informatics and Telecommunications[33] of the School of Sciences of the National Kapodistrian University of Athens. Research and development activities within the department cover a wide spectrum of information and communication technologies. The group has a rich and long experience in several topics of computer science including digital libraries (information integration and access, Grid-services, cultural heritage systems) and e-Infrastructures.

Organization of activities The MADGIK group counts around 40+ members, including five faculty staff, several R&D staff and students at all educational stages. Being active in research and development, it includes 15 full time technical people, organized in R&D project-dedicated teams, each led by team leaders and supervised by the scientific coordinator.

The group is in close collaboration with other groups of the same organization for publication writing and software development issues and has a strong and long tradition of cooperation with groups of other organizations.

Computing infrastructure The group has a local storage and computing infrastructure, consisting of personal workstations and shared servers in a local network, organized in virtual machines. In projects such as D4Science-II, part of this infrastructure joins a larger development and execution environment that consists of a cluster of 110 CPUs with 300 GB RAM and 15 TB of storage.

[32] http://madgik.di.uoa.gr.

[33] http://www.di.uoa.gr/en.

3.5.2 Research objectives and projects

The MADGIK group has the following general research objectives:

- databases and information systems: Data repositories, query optimization, personalization, intelligent databases, etc.;
- digital libraries;
- human computer interaction: user interface for databases, complex data visualization;
- scientific repositories: scientific experiment management, data repositories, workflow management.

It participates and has participated in a large number of national and European projects related to e-Infrastructures which include:

- **OpenAIRE project (EC FP7):**[34] (see project description in 3.1 D-Lib research group) the group focuses on designing and developing user interfaces and end-user functionality services, as well as on services for the integration of access statistics collected from European repositories.
- **DRIVER Targeted Project (IST FP6) and DRIVER II CP/CSA (INFRA FP7):**[35] (see project description in 3.1 D-Lib research group) the group focuses on end-user functionality services, such as user profiling, user recommendations and “generic” portals dynamically adaptable to match functional requirements of end-users of different communities;
- **D4Science CP/CSA (INFRA FP7) and D4Science II CP/CSA (IP FP7):**[36] (see project description in 3.1 D-Lib research group) the group focuses on optimized and distributed search services, as well as on highly configurable data transformation services.
- **DILIGENT Integrated Project (IST FP6):**[37] the main objective of DILIGENT (Castelli, Candela, Pagano and Simi, 2005) has been to create an advanced testbed for knowledge e-Infrastructure that will enable members of dynamic virtual e-Science organizations to access shared knowledge and to collaborate in a secure, coordinated, dynamic and cost-effective way.

Research data The group produces mostly research data in the form of software, software instances, benchmarks, experimental data, XML, system logs and websites. Software is stored and versioned through SVN services, in some cases shared with project partners.

[34] http://www.openaire.eu.

[35] http://www.driver-repository.eu.

[36] http://www.d4science.eu.

[37] http://diligent.ercim.eu.

Unstructured metadata, in the form of technical reports and project deliverables, are compiled (possibly in collaboration with other project partners) and exchanged through e-mail when edited. Tools such as Google Docs or common project wikis may be adopted for collaborative editing but are not the rule.

Depending on the domain of the e-Infrastructure to be delivered, domain-specific research data may be collected and used as benchmarks; e.g. images and raw scientific data, audio and video, publication full texts, big data, time-series. Interestingly, the work space resulting from the D4Science project is used for benchmark data storage and exchange by the group itself. This platform has been developed to support scientific research in general with environmental and maritime data as the use cases and allows data management and exchange through web user interfaces. Similarly, metadata formats of domain-specific research data may be regarded as benchmarks; examples vary from standard, e.g. Dublin Core, Darwin Core, SDMX, ISO for geographical data, to proprietary formats.

For software and software instances, custom metadata may be used, in agreement with the specific project requirements, which in turn depend on shared development policies. The use of custom metadata for software instances has been the result of user or system needs, as the standards were not defined or sufficient (e.g. an example is the need to record in the metadata service dependencies). Technical reports are rarely annotated with metadata but it is planned to make this annotation standard within the group in the near future.

Desiderata: after the end of an EC project, consortiums have an obligation to keep the resulting reports only for a few years. The EC project BELIEF provides a digital library where documents of past projects can be stored for future storage in time. However, it would be desirable if funding agencies, such as the EC, would provide a “place” (namely an infrastructure) where past and ongoing projects could store and retrieve their data outcomes, from software to technical reports and deliverables.

Research literature Scientific publications are exchanged through e-mail and rarely edited through collaborative tools, like Google Docs. In some cases, some of the authors may be reluctant to learn and use a new collaborative tool, so e-mail exchange is the more common practice.

In order to search for publications, tools like Citeseer and Google Scholar are more commonly used and, to a lesser extent, the DRIVER infrastructure. The group’s preferred publication forms are conferences, online and print journals and PhD and MsC theses.

Combining data and literature The group believes in the publication of data combined with literature, as a mean to verify the experimental results and conclusions of the publication. However, it does not implement those practices, due to the lack of standards and tools.

Open Access issues The group supports Open Access for publications and also adopts it, although not as a strict policy. The reason is that the top conferences and journals touching the fields typically do not implement Open Access business models.

Most of the research data and metadata are open but exceptions exist:

- software data: the group tends to adopt GPL licences and open source;
- unstructured metadata: technical reports and documentation are available openly on the wiki, except for the project managerial/financial ones;
- logs: service activity logs are used for debugging purposes and for measuring the usage of the infrastructure from several perspectives, including end-users and applications. As a consequence, logs can be released to third parties only after proper permissions, as they may be used to infer private information.

3.5.3 Research workflows

The group works on system design and development based on research findings and on relative scientific publications. When operating in the context of a project whose aim is to deliver an e-Infrastructure, the typical workflows consists of the standard phases of requirement analysis, design and implementation, by reusing, experimenting or devising research achievements and solutions of the research group. Design, development and testing of software are often carried out in cooperation with project partners, by sharing hardware and supporting tools. Research papers are often presenting a system or part of it, together with experimental results which prove its effectiveness or quality.

3.6 Engineering R&D Unit on Clouds and distributed computing infrastructures

3.6.1 General information

Engineering Group is Italy's largest systems integration group and a leader in the provision of complete IT services and consultancy. Engineering Group has about 6500 employees and 35 branch offices, throughout Italy, in Belgium and (outside the EU) in Brazil. The Engineering Group operates through seven business units: Finance, Central Government, Local Government and Healthcare, Oil Transportation and Services, Utility, Industry and Telecom, supported by an SAP transverse skills centre and by its Central Office for Research & Innovation, with researchers active in Italian and EU projects. Engineering was one of the first Italian companies to adopt the Quality standard ISO 9001 in the early 1990s. Since 1996 the company has adopted NATO standard AQAP 2110/160 certification. And recently the production units have been certified CMMI® level 3. The Pont Saint Martin Service Centre (PSM) provides to more than 100 Italian and international customers, 40,000 workplaces, 1000 remote connections, 10,000 electronic mail boxes and about 7000 SAP users. The R&D department is organized to work in strict cooperation with business divisions in order to facilitate knowledge and technology transfer.

The Engineering R&D Unit is involved in the NESSI[38] and NEM ETPs[39] initiatives and in a number of Grids (Foster and Kesselman, 1999) and Cloud (Dikaiakos et al., 2009) related initiative including VENUS-C (see 3.1 D-Lib research group), VisionCloud, Passive, TEFIS,[40] ERINA4Africa,[41] ERINA+,[42] ARISTOTELE[43] and D4Science-II (see 3.1 D-Lib research group). The Engineering team interviewed consists of 16 members.

Organization of activities

Research and development activities are managed by dedicated teams that are formed by taking into account the requirements of the specific activity and evolve during the activity itself, e.g. new members can be added or members having different expertise might replace previously allocated members. Members of the group partake to multiple activity teams. The overall goal is to maximize the use of human resources.

Computing infrastructure

The infrastructure supporting the activities of the interviewed group consists of 16 workstations (the policy is to have one workstation per group member) plus the computing resources listed in Table D.4. In addition to that, the team makes use of resources acquired through one or more Cloud infras-

[38] http://www.nessi-europe.com.

[39] http://www.future-internet.eu/news/view/article/the-cross-etp-vision-document.html.

[40] http://www.tefisproject.eu.

[41] http://www.erina4africa.eu.

[42] http://www.erinaplus.eu.

[43] http://www.aristotele-ip.eu.

tructures, including Windows Azure, Barcelona Supercomputing Center and Engineering Group data centre.

Table D.4 ENG computing infrastructure

CPU	12 Servers (bi processor – quad processor)
Storage	2 TB
Storage area network	1 SUN (1.7 TB)

3.6.2 Research objectives and projects

The Distributed Computing R&D group focuses on a number of research and development activities including:

- software configuration, build and testing;
- authorization, authentication and accounting in distributed infrastructures including service-oriented architectures (Lomow and Newcomer, 2005), Grid (Foster and Kesselman, 1999) and Cloud domains;
- Grid and Cloud computing (focusing on their exploitation in Real Business ENvironments).

The team has been involved in many EU-funded projects relevant to the topics of e-Infrastructures, namely:

- **D4Science-II:**[44] actually the third phase of a project started with the name of DILIGENT (Castelli et al., 2005) where the Engineering has been involved since the beginning. D4Science-II is developing an infrastructure enabling the interoperation of diverse infrastructures that are running autonomously, thereby creating ecosystems that can serve a significantly expanded set of communities. In this project, Engineering mainly works on the design and implementation of security-related solutions, focusing on interoperability aspects and takes care of the overall coordination of the integration, testing and distribution activity;
- **VENUS-C:**[45] an FP7 Research Infrastructures project, coordinated by the Engineering team is building open source facilities to provide an easy-to-use and service-oriented Cloud infrastructure. From a technical standpoint, Engineering leads research and technological development activities dedicated to Monitoring, Accounting and Billing while also contributing to activities related to Application Security. Engineering is also the lead partner to evaluate new business and sustainable models for scientific computing in close synergy with partners from enterprise as part of the activities pertaining to Communication and Sustainability;

[44] http://www.d4science.eu.

[45] http://www.venus-c.eu.

- **ERINA+:**[46] a project that is developing and applying techniques for measuring the socioeconomic impact of the project funded by the European Commission within unit F3 (Research Infrastructures) by enhancing and applying the socioeconomic methodology for the impact evaluation and assessment, already conceived and experimented during the ERINA study. Engineering is the coordinator of the project and is leader of the activities on the dissemination of project results;
- **ETICS 2:**[47] a project (the second phase) that developed an out-of-the-box software build and testing infrastructure, powered with a build and test product repository, and automatic collection of software quality metrics. Engineering was involved in tuning, improving and integrating the Grid Quality Certification Model (Meglio, Bégin, Couvares, Ronchieri and Takacs, 2008), with other established certification procedures and standards as well as developing and maintaining a web client to facilitate the interaction with the ETICS service.

Research data With respect to research data, the team mainly deals with software artefacts, project reports and technical documentation leading to websites, wiki pages and manuals. Unfortunately, although scientific paper production is encouraged, it is not frequent.

These research data are shared mainly among teammates by relying on tools that might depend on the activity the team is involved. Among these tools there is intranet, CSV and ETICS (for software artefacts) – which are exploited by all the teams – as well as tools like BSCW[48] and TRAC[49] – which are mainly used in the context of specific teams because are somehow a working practice imposed by the activity, e.g. they are imposed in a research project like D4Science-II.

The metadata collected depend on the tool/software they are conceived for, e.g. the metadata equipping software artefacts designed for ETICS are based on ETICS specifications. There is no metadata standard that the team is requested to use but those resulting from the tools they rely on to perform their activities.

Desiderata: the team is discussing the benefits and drawbacks in making the research data they produce publicly available, although they are regulated by policies. This holds mainly for software artefacts. On one hand, this practice is conceived to be a good practice leading to enhancement of organization visibility and business; on the other hand, it is conceived to be a

[46] http://www.venus-c.eu.

[47] http://etics.web.cern.ch/etics.

[48] http://public.bscw.de.

[49] http://trac.edgewall.org.

"dangerous" practice because of the risk of reducing the organization's competitiveness. The desiderata are to have facilities for enhancing the visibility of the data that guarantee visibility of policies regulating data access and provenance.

Literature With respect to literature, the production of scientific papers is limited while the consumption is encouraged. Engineering team mainly relies on known publication sources, such as Google, Google Scholar and publisher websites (e.g. Elsevier, ACM, IEEE). As regards paper production, the team relies on "standard" editing tools (namely Microsoft Word) and file-sharing facilities, e.g. the intranet, email attachment, Dropbox.

Desiderata: because of the limited activity, there are no major desiderata but the overall team is interested in having a seamless access to all the literature. In particular, this seamless access should simplify the discovery of the so-far produced literature on a specific topic.

Combining literature and data With respect to linking data and literature, it is common to provide the paper with the URL(s) of the software artefacts the paper is documenting or is related to. In addition to that, it is quite common to have websites/web pages dedicated to document software artefacts.

Desiderata: the mechanisms for linking data and software artefacts should be strengthened. In addition to a simple link, a bunch of metadata should be either explicitly added or dynamically derived with the goal to enrich the paper with characteristics of the software artefact, such as the licences, technical requirements and software dependencies. These metadata should be machine oriented as to promote the implementation of tools benefiting from these data.

Open Access With respect to Open Access, there are no established policies within the group. Open Access strategies aiming at enhancing research and development results are encouraged. However, it should be possible to define fine-grained access policies.

3.6.3 Research workflows

The typical research production workflow of the team is pragmatic and quite standard since it is mainly oriented to produce new software artefacts. It includes the following phases (this is a simplistic view, the phases are organized in loops where decisions taken at certain points can be reconsidered thus leading to multiple iterations):

1. requirement analysis: producing requirement analysis documents by close interaction with the customers;
2. problem characterization and analysis;
3. survey: of existing tools (off-the-shelf solutions) and approaches that can be (re-)used in the context of the problem domain;
4. design: of a technical solution resolving the specific problem by promoting the (re-)use of existing technologies and standards;
5. implementing and testing: of the envisaged solution;
6. release: of the software artefact with the related documentation.

4 Current status

From the analysis of the interviews, it appears that researchers in the field of e-Infrastructure follow similar research workflow patterns, mostly in the direction of producing software data (to be used in the construction and maintenance of production infrastructure systems) and relative publications. The e-Infrastructure community, however, has not reached common agreements on policies, standards and best practices in the production of research data and literature. Depending on their focus (e.g. companies and research institutions), organizations and research groups tend to grow their own research infrastructures, based on proprietary best practices, policies, data formats, etc., in order to enable their researchers to collaboratively discover, produce, store, share and publish online both research data and literature. Typically, as illustrated in Figure D.1, such infrastructures are obtained as combination of:

- **local service and computing infrastructures:** examples are hardware (e.g. machine clusters), services such as SVN and TRAC for software data versioning and development and repository systems for literature storage and publishing;
- **web infrastructure:** as many other computer science research communities, the e-Infrastructure community makes heavy usage of the plethora of online tools for literature drafting (e.g. Google Docs, discovery, e.g. Google Scholar, BASE, OAIster, DRIVER) and sharing (e.g. SourceForge, Google projects, Apache projects, Dropbox). Among such online tools are included also local infrastructures which offer web access to their literature and data, e.g. institutional repositories, open source SVN systems.

Due this “local” approach, the e-Infrastructure research community has not established standards for data formats and classification or metadata for data resources, nor either policies and rules for interlinking data and literature.

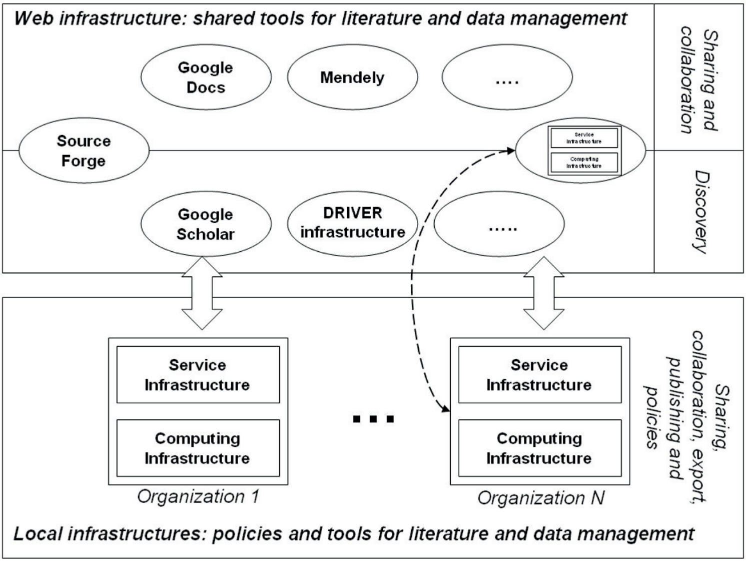


Figure D.1 Current status of research infrastructures for e-Infrastructure research

Overall, the research community has not grown a shared research infrastructure and, as a consequence, an e-Infrastructure, from both the organizational (i.e. policies, standards and best practices) and technical (i.e. services) point of views. Through such e-infrastructure, research data and literature in the field could be (possibly openly) collected, shared, exchanged and linked to each other, based on well-established participation and access policies and standard formats. Although researchers agree on the potential benefits that such infrastructure would bring, no plan in this direction is being undertaken. The reasons for this are many: for example the existence of practical and powerful online tools, reluctance to change methodologies, lack of funds and logistics and the youth of the discipline.

The unavailability of a common e-Infrastructure leads to two main drawbacks:

- **interoperability costs:** whenever organizations need to cooperate in the production of data and literature, for example within collaborative research projects, they have to bear a cost of interoperability of content (e.g. data and metadata exchange) and of learning new cooperation tools (e.g. file sharing, publication drafting, software versioning).
- **hardly reachable data and literature resources:** in order to discover and identify data and literature of interest to the field, researchers

need to access and search the plethora of web sources available for publication and data sharing ("aggregators", e.g. Google Scholar, DRIVER, SourceForge), but also websites of organizations (e.g. to find software products and documentations), often reachable through generic searches on "The Web" (e.g. Google, Yahoo).

The following sections summarize the results gathered through the interviews, trying to cover all aspects of the typical e-Infrastructure research workflows and identifying the possible improvements that would derive by the establishment of a common research infrastructure.

4.1 Research data

4.1.1 Data types and metadata

Research data typologies are:

- **software:** intended as programming language code or the results of code compilation, such as installation packages;
- **software instances:** intended as software running on a machine (e.g. web services), often described by a so-called "profile" (Grid terminology) and therefore discoverable and reusable for interaction or "orchestration" by authorized applications;
- **benchmarks:** intended as collections of data available through any kind of storage support (e.g. file system, DBMS) and used for testing purposes. Typically their format, size and storage support vary depending on the application domain and can included videos, images, table data, database tables, files and folders;
- **logs:** intended as recorded histories of actions or events, typically used to evaluate and monitor the activity of a software instance. Their storage modes and formats vary, ranging from databases to text files;
- **statistics:** intended as qualitative or qualitative measures often derived from logs analysis to evaluate software instance activities (e.g. number of requests to a software instance in a given period).

Regarding metadata typologies, in general, data come with metadata information in order to make it available for discovery and re-use within and outside the local infrastructures.

Generally, structured metadata (produced in the form of records/profiles which are interpretable by a machine), can obey to proprietary or standard formats depending on the typology of data. In some cases, as for software and software instances data (e.g. D-Lib research group), proprietary metadata structures are introduced to be able to describe domain-specific properties of the data (e.g. dependencies of software packages). Metadata standards are also adopted (e.g. Dublin Core for technical reports), often imposed by the tools integrated in the local service and computing infrastructures (e.g. repository platforms, BSCW).

Researchers also heavily rely on unstructured metadata in such forms as roadmap specifications, functional and architectural specifications, policy specifications, guidelines, usage and installation manuals, software documentation and technical reports. Documentation is made available in various standard formats, such as file formats PDF, docx, Latex, DocBook or through web formats, such as Web 2.0 wikis and more "traditional" websites.

4.1.2 Data management aspects

Organizations provide a wide range of data storage and export solutions, whose adoption depends on the typology of data and on the Open Access policies adopted. When asked, interviewees confirmed that they do not implement literature or preservation policies and no desiderata have been suggested in this direction.

Storage Organizations' local infrastructures are equipped with version control management systems (e.g. SVN, Git) and issue trackers (e.g. TRAC, RedHat Issue Tracker), through which they manage software data. Similarly, technical reports and benchmarks are stored using standard document management tools, such as repository platforms (e.g. DSpace, ePrints, Fedora, PUMA) and sometimes version control systems. Typically, such tools are under the control of the organization and to authorized users and applications.

Production Some organizations have adopted a systematic approach and have grown a local service and computing infrastructures where data can be managed across several projects and research activities, under controlled access policies. In other cases, such tools are deployed as independent instances, dedicated to the research activities of the case. In some cases, Cloud technology is exploited, in order to outsource the cost of temporary or high peaks of storage and computing power demand. For example, this may be useful when testing highly distributed algorithms to be run on Grid-oriented (Foster and Kesselman, 1999) research infrastructures. Typically, Cloud CPU rental enables the arbitrary growth of CPU or storage demand (especially, peaks of demand) at a cost that is lower compared to the one of purchasing and maintaining the machines required to run the same tests.

Collaboration Researchers collaborate in the production of software, unstructured metadata and benchmarks by exploiting the functionality offered by the tools available to them through the local service and computing infrastructures and through online tools such as Google Docs for technical reports.

Export and policies Research data, both data and metadata, are shared and published by means of local services, such as SVN or organization/project websites. Research data are subject to confidentiality and protection policies that depend on the organization and, within the organization, on the typology of data and on the project or research undertaken. The trend is for companies in the field to be reluctant on openly sharing the data they produce for business (e.g. Engineering). Such data are generally accessible within the boundaries of the organization and sometimes not available outside to the owning research group. On the other hand, research institutions tend to publish and disseminate their results through all possible means, to promote and give visibility to the results of their activities. In general, when disclosed to the world, the usage of software and unstructured metadata may be restricted according to standard licensing schemes and non-disclosure agreements.

4.2 Literature

e-Infrastructure researchers follow a typical literature lifecycle, made of phases of: (i) survey and analysis of the literature and (ii) drafting and publishing of an article, of course prior to submission, reviewing and acceptance to a venue, such as a conference, or a journal. Both phases are largely affected by the interdisciplinary nature of e-Infrastructure research, which is placed somewhere in between service-oriented architectures/infrastructures, Grid infrastructures, digital libraries, multimedia storage, information retrieval, big-data (NOSQL solutions) and the specific functionalities of the research field for which e-Infrastructures are necessary.

Survey and analysis There is no dedicated online literature source for e-Infrastructure research. Researchers rely in general-purpose online aggregators, such as Google Scholar, Citeceer, the DRIVER infrastructure (see 3.1 D-Lib research group), BASE,[50] OCLC-OAIster,[51] Scopus[52], publishers websites, such as Springer, Elsevier and ACM or the Web, with Google, Yahoo and other search engines typically used by the majority of computer science researchers. Similarly, some of them also exploit online tools such as Mendeley and CiteUlike[53] to share their favourite reading lists.

[50] http://www.base-search.net.

[51] http://www.oclc.org/oaister.

[52] http://www.scopus.com.

[53] http://www.citeulike.org.

Drafting As many researchers in computer science, articles are written exploiting online free tools for collaborative editing and file sharing, such as emails, Google Docs, Dropbox and SVN servers (e.g. for Latex articles).

Publishing Due to the interdisciplinary nature of the research field, only a few venues specific to e-Infrastructures are available, e.g. some tracks on Theory and Practice for Digital Libraries conference (formerly ECDL) and IFLA (International Federation of Library Associations and Institutions) conference series. As a consequence, articles in the field end up being submitted in journals and conferences related with digital libraries, service-oriented architectures, Grid and discipline-specific venues, those for which e-Infrastructures are constructed (e.g. biology, cultural heritage, grey literature). Some organizations from academia, research and industry also support and fund PhD and MsC theses.

4.3 Linking literature and research data

Researchers do reference their software and unstructured metadata from their publications by means of URLs indicating project website or downloadable files and as bibliography references. Moreover, data such as table data and graphs are placed/embedded within the publications text or, when too large for the publication body, as an appendix. This attitude reveals the awareness of the benefits of pointing readers to actual evidence of the results, but also shows the necessity of a more structured approach. In this process of linking publications and data, both writers and readers follow their intuition and not agreed-on rules, e.g. how to point to data, how to describe data properties and provenance. A more structured approach would enable better evaluation of the quality of the publication, avoid falsified data and enable discovery and re-use of the data, for example in order to improve previous scientific results. Interviewed researchers generally agreed on the benefits of such a combined approach for publication and expressed the need for both policies and tools to support its diffusion.

4.4 Open Access

It appears that most organizations are aware of the existence of the Open Access initiatives and agree with their mission and goals. In fact, many of them also actively promote it among their own researchers and in other communities (e.g. D-Lib group, EKT, MADGIK group). This is typical for research and academic institutions, whose interests are the dissemination of their achievements through Open Access literature and open source software data and unstructured metadata (e.g. technical reports). On the other hand, many organizations have interests which conflict with the consequences of Open Access especially on the side of software data (in this case Open Access translates in open source). In this context, Open Access may have the undesirable side effect of disclosing technology to third-party organizations, thus potentially reducing the possibility to sell it to customers (e.g. Engineering).

4.4.1 Literature Open Access issues

In general, although many organizations in the field are supporting and promoting Open Access, it seems that none of them has imposed Open Access policies as obligatory to its researchers. This choice has mainly to do with the lack of Open Access publishers linked with relevant conferences or journals in the field, i.e. those giving more value and thus visibility to research results, and with the high costs of purchasing gold Open Access licences from them (e.g. "Open Choice" publishing model from Springer).

Organizations store their publications in local repository platforms or websites in order for third-party organizations and researchers to follow their activities and get hold of the actual documents (for Open Access material) or to reach the toll-gate sources from which these can be requested. Since such sources are reached by online aggregators such as Google Scholar, DRIVER, etc. e-Infrastructure literature can be considered today discoverable through accurate and selective search activities.

Overall, no e-Infrastructure-specific literature sources are available on the web and researchers are required to tentatively search for publications in the field across online collections pertaining to several research domains.

4.4.2 Data Open Access issues

Open Access for data depends on the typology of data and on the specific policies of the organization involved.

For software data, Open Access, namely open source, is always considered a possibility and generally ruled by means of specific software licences, from GPL, Apache and non-disclosure agreements. Organizations make software available through product websites, local software repositories and sometimes through shared open source software repositories, such as SourceForge.

For software instance data, Open Access translates in open interaction with the APIs of running software. However, this is rarely the case. API access policies are often controlled through authentication and authorization protocols or, more simply, through white lists and black lists of IP addresses.

For unstructured metadata (e.g. technical reports, specifications), Open Access is a common practice, although often decided on a case by case basis. Organizations make available their unstructured metadata through product websites and local repository platforms, which are often aggregated, i.e. web crawled or OAI-PMH harvested (Lagoze and de Sompel, 2001), by online search engines, such as Google Scholar.

For benchmarks and log kind of data, Open Access policies are not frequently applied for a number of reasons. In some cases, these are simply not perceived as resources possibly reusable by the community. In others, they are produced in proprietary formats and may therefore result not interesting or not be easily re-used by third-party consumers. Finally, as for web log files or benchmarks obtained by protected information, there may be privacy issues that prevent such data to be openly disseminated.

Overall, e-Infrastructure data are available from the individual organization stores, websites and repositories, given these are made accessible from the Web and not only within organization intranets. This well-established attitude makes research data in the field hard to expose and discover, hence to re-use or reference by researchers.

5 Desiderata and future directions

The interviewees also suggested a number of desiderata on which aspects of e-Infrastructures could/would improve the current research workflows. In the following, such ideas are collected and presented according to the structure of the questionnaire: research data, literature, linking data and literature and Open Access. Finally, these are combined to figure out how an e-Infrastructure for e-Infrastructure research that meets such desiderata may impact on and benefit the overall community.

5.1 Research data

Controlled data sharing In general, e-Infrastructure researchers are willing to share their data so that they can reach and consume data produced by others. Sharing policies may range from open source licences and toll-gated copyrights to non-disclosure agreements, but the (marketing) principle is that data resources should be reachable and potentially accessible by researchers interested in them. For example software, unstructured metadata, benchmarks and logs should be always discoverable and reachable through community-oriented web tools, together with a metadata description of their degree of Open Access.

Data unreachability on the web In many cases, researchers find it hard to reach data they might need outside the boundaries of their organizations. For example, this is the case for software and unstructured metadata when these are not published on shared repositories such as SourceForge or exposed through repository platforms and product websites to be then crawled by web search engines. Researchers need to agree on best practices and policies for data publication and require community-specific tools for leveraging discovery of their data according to such policies.

Lack of metadata description standards In those cases where research data are available through web tools (e.g. software through Apache projects), the relative metadata properties are not peculiar to e-Infrastructure resources. This makes it hard for researchers to distinguish and identify the resources they require. Researchers need standards for data descriptive metadata and for data unique identifiers (e.g. DOIs, web handles).

Service and computing infrastructure sharing Typically software is developed, tested and integrated on local service and computing infrastructures featuring adequate CPU and storage quotas. Maintenance of services and hardware leads to high sustainability costs, hardly affordable by many communities. These costs could be reduced by adopting e-Infrastructures for sharing computational resources across multiple organizations according to a combination of service Cloud (Dikaiakos et al., 2009) and Grid resource sharing (Berman, Fox and Hey, 2004). This economy of scale approach would maximize the usage of resources and therefore minimize the overall cost of maintaining very large infrastructures and realizing complex e-Infrastructure software.

5.2 Literature

Lack of common classification schemes for literature The community calls for a clean classification scheme of the research field, in order to organize its scientific production and facilitate its discovery.

Lack of services for sharing literature e-Infrastructure literature is not easily discoverable through well-known web publications sources, mainly due to its interdisciplinary nature. The community calls for common services enabling the collection and discovery of publications in the field.

5.3 Linking literature and research data

Researchers realize the advantages of interlinking publications with research data in a meaningful way, from reusability of data to more effective validation of the results. To this aim, they need to agree on common policies for specifying references to data from within a publication text or from within the publication metadata. This work should be realized in conjunction with the definition of standards for metadata and unique identifiers for data resources.

5.4 Open Access

As in other research fields, e-Infrastructure researchers realize the importance of Open Access for both data and literature. On the other hand they are also aware of (i) the "certification of excellence" implied by peer review mechanisms, which often lead to retention of copyrights, and (ii) the return-of-investment principles behind the production of data for business. Hence, as for other research fields, to enforce Open Access, researchers need innovative business models.

5.5 A research infrastructure for e-Infrastructure researchers

The researchers' desiderata presented in the previous section seem to converge to the realization of an e-Infrastructure providing policies and services for sharing and collaboratively constructing research data and literature resources in the field of e-Infrastructures. As illustrated in Figure D.2, such an e-infrastructure would be complementary to the current local infrastructures. The combination of the two layers would give life to an effective research infrastructure for e-Infrastructure researchers. This would be spontaneously maintained by organizations willing to benefit from its services, based on well-known economy-of-scale principles. Its benefits would derive from a combination of organizational and technological efforts:

- **Organizational**
  - promote standards and policies for data and literature exchange (formats) and description (metadata);
  - promote standards and policies for interlinking research data and literature;
  - investigate on new business models capable of reaching the right compromise between publishers business and open access policies, without compromising the evaluation and publication process of research results.
- **Technological**
  - services for safely sharing and curating research data and literature in the field;
  - services for discovering and interlinking research data and literature in the field;

- services for collaboratively constructing research data and literature by reusing existing resources.

Investigations and studies on how communities could gradually move towards the realization of these objectives in a collaborative and synergic fashion are being undertaken in the EC project OpenAIRE. Experimental solutions in interlinking of research data and research literature have been realized in the EC project DRIVER-II, e.g. enhanced publications (Woutersen-Windhouwer, Brandsma and Hogenaar, 2009) and will be implemented in the EC project OpenAIREplus (to be started in December 2011).

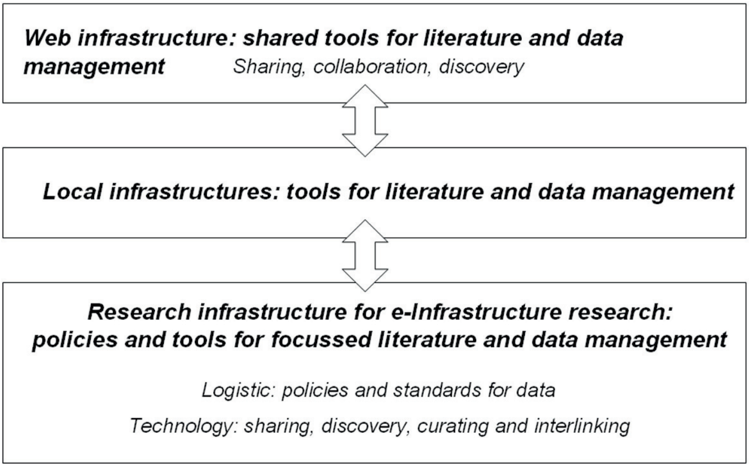

Web infrastructure: shared tools for literature and data management
Sharing, collaboration, discovery
Local infrastructures: tools for literature and data management
Research infrastructure for e-Infrastructure research: policies and tools for focussed literature and data management
Logistic: policies and standards for data
Technology: sharing, discovery, curating and interlinking

Figure D.2 Challenges: future research infrastructure for e-Infrastructure researchers

It is hard to envisage or quantify the cost for organizations willing to work in synergy to realize and maintain such infrastructure, as well as the cost of those organizations willing to join in a second stage, in order to benefit of its services. Certainly, as it happened in the past with other research infrastructures, the initial spark should come for a strongly motivated community, whose history and vision justifies common objectives, goals and risks. Although the e-Infrastructure community is probably the one which can at best realize this goal, its history is still in an early stage and such motivation is likely largely missing today.

6 List of figures

7 List of tables

8 Bibliography

Atkins, DE, Droegemeier, KK, Feldman, SI, Garcia-Molina, H, Klein, ML, Messerschmitt, DG, Messina, P, Ostriker, JP, & Wright, MH. *Revolutionizing Science and Engineering Through Cyberinfrastructure*. 2003.

Berman, F, Fox, G, & Hey, A. *Grid Computing: Making the Global Infrastructure a Reality*. John Wiley & Sons, 2003.

Candela, L, Castelli, D, Ferro, N, Ioannidis, Y, Koutrika, G, Meghini, Pagano, CP, Ross, S, Soergel, D, Agosti, M, Dobreva, M, Katifori, V, & Schuldt, H. The DELOS Digital Library Reference Model - Foundations for Digital Libraries. Version 0.98. February 2008.

Castelli, D, Candela, L, Pagano, P, & Simi, M. *DILIGENT: A DL Infrastructure for Supporting Joint Research*. 2nd IEEE-CS International Symposium Global Data Interoperability – Challenges and Technologies, 2005, 56-69, Society, I. C. (Ed.)

Dikaiakos, MD, Katsaros, D, Mehra, P, Pallis, G, & Vakali, A. Cloud computing: distributed internet computing for IT and scientific research. *Internet Computing, IEEE* 2009, 13, 10–13.

Foster, I & Kesselman, C. *The Grid: Blueprint for a New Computing Infrastructure.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.

Ioannidis, Y, Maier, D, Abiteboul, S, Buneman, P, Davidson, S, Fox, E, Halevy, A, Knoblock, C, Rabitti, F, Schek, H, & Weikum, G. Digital library information – technology infrastructures. *International Journal on Digital Libraries*, 2005, 5, 266–274.

Lagoze, C & de Sompel, HV. *The Open Archives Initiative: Building a Low-barrier Interoperability Framework.* Proceedings of the first ACM/IEEE-CS Joint Conference on Digital Libraries. ACM Press, 2001, 54–62.

Lomow, G & Newcomer, E. *Understanding SOA with Web Services*
Addison Wesley Professional, 2005.

Meglio, AD, Bégin, ME, Couvares, P, Ronchieri, E, & Takacs, E. ETICS: the International Software Engineering Service for the Grid. *Journal of Physics: Conference Series* 2008, 119, 042010.

Woutersen-Windhouwer, S, Brandsma, R, & Hogenaar, A. *Enhanced Publications: Linking Publications and Research Data in Digital Repositories.* Amsterdam University Press, 2009, 212.

E Research in the Humanities and Social Sciences

Arjan Hogenaar, Heiko Tjalsma and Mike Priddy

1 Introduction

The social sciences and the humanities taken together contain a heterogeneous range of research disciplines. Almost all existing methods of research can be found within these two domains. Data handling (collecting, processing, selecting, preserving) and publication methods differ greatly. Attitudes in the field towards Open Access of publications as well to research data vary as well.

It is not possible to cover the total fullness, and complexity, of all the disciplines within these two domains. Our observations will therefore be based upon a number of case studies. Taken together these case studies give a fairly representative picture of the domains, at least of the most common research environments. The main dividing line is between those disciplines creating empirical data, such as survey data in the social sciences and those, especially in the humanities, using existing source material, such as history or text studies. This source material can either be of an analogous or a digital nature. As will be shown in the case studies in many disciplines a mix of created and existing is often combined.

The Data Archiving and Networked Services (DANS[1]) has been chosen as an exemplar within the area of social science and the humanities. DANS promotes sustained access to digital research data. For this purpose, DANS has created the online archiving system EASY[2] which enables researchers to archive and re-use data in a sustained manner, primarily in the social sciences and the humanities. It is expected that this will be extended to other disciplines in the future. In addition, the institute provides training and advice and undertakes research into sustained access to digital information.

[1] http://www.dans.knaw.nl.

[2] http://www.easy.dans.knaw.nl.

Through its activities, DANS is in close contact with a number of researchers in the two domains of this study. The findings in this section are based on interviews with a selection of these. Care has been taken that this selection was as representative as possible for these heterogeneous disciplines. Interviews of approximately 1 hour each were conducted with a range of researchers from both within DANS and with researchers from other institutions that have close ties with DANS: either collaborators on projects or who are using or depositing data in the online archiving system EASY. The interviews were semi-structured with a list of questions and subquestions, but if it was clear that certain groups of questions were not relevant to the interviewee these were not asked. The majority of interviews were conducted with scholars who would identify themselves as working in the Humanities or with humanities data. This emphasis was because DANS had recently conducted similar interviews on usage of digital data and research infrastructures with senior researchers, where the bias was towards the social sciences. A total of 15 interviews were conducted, nine with humanities researchers and six with social scientists. Even with such a small sample of interviews, we attempted to get a broad cross-section of disciplines; however, within archaeology we conducted three interviews to get a deeper insight into one discipline.

The interviews conducted with senior academics and research managers, mostly professors and/or directors of research institutes, occurred in summer/autumn 2010 and in spring 2011. This set of interviews formed part of a strategic plan on widening the scope of social science and humanities disciplines utilizing the services of DANS.

In most interviews, the need for data preservation and (open) data access, as experienced in the specific discipline, were discussed in a very broad sense. The interviews carried out within the humanities and social sciences are of particular importance for the OpenAIRE Project, as they give insight in how researchers within these fields deal with open access to data and publications. They focused on a limited number of fields: economics/econometrics, finances, sociology (survey research) and law.

Furthermore an online survey was used which was carried out into data usage, data archiving and research infrastructures amongst researchers from all disciplines in the Netherlands.

2 Workflows in social sciences and humanities research

2.1 Phases in social science research

Discovery and planning Starting from a theoretical and empirical perspective, the researcher first wants to extend his or her knowledge. The researcher shall need to explore what data will be required to give the best answers to the scientific questions of his or her research project: are there existing (archived) data available or should new data be collected?

Initial data collection This is the phase in which data are actually collected. This could be in the form of a survey held or an experiment carried out, or the acquisition of previously collected data, possibly restructured or linked to other datasets, may form the foundation of the data collection. Essential data management strategies are formulated and executed, including decisions about documentation content and formats.

Final data preparation and analysis In this third phase, the researcher undertakes analysis after having performed final verification and modification of the data. The process of data preparation should be is complete and results are written up.

Publication and sharing In the fourth phase, the researcher will communicate the research findings in publications.

Long-term data management In this final phase, there are two critical goals, seen from the perspective of the wider social science community: providing access to the data and ensuring long-term preservation. Once the data are available for secondary use, they have reached the final stage of the research cycle and could become the start of new projects that begin with their discovery and re-use thus beginning the cycle anew.[3]

2.2 The lifecycle in the humanities

Because of the heterogeneity within the humanities, this example describes historical or textual research in general, but applies less to other domains within humanities, such as archeology.

[3] Green, AG and Gutmann MP. *Building partnerships among social science researchers, institution-based repositories and domain specific data archives.* OCLC Systems and Services: International Digital Library Perspectives 2007, 23, 35–53.

Creation In this first phase, the design of the information structure has to be made through data modeling or text modeling, based on the goals and design of the research project. It also includes the physical production of the digital data either by data entry and text entry tools, or by digitization (optical character recognition) of existing analogous resources.

Enrichment The raw data, in whatever format (images, texts, databases, GIS-files) will have to be enriched with metadata, describing the historical information in more detail and, in particular, the context and provenance of it. Preferably this should be done in a standardized way (Dublin Core for example), but in practice this mostly not the case.

Editing Editing includes the actual encoding of textual information by inserting mark-up tags or entering data in the fields of database records. Enhancement could be considered as a separate phase of the editing process by which data are being transformed. This stage could also include annotating original data with background information, bibliographical references and links to related passages.

Retrieval In this phase, the information should be ready to be selected (by queries), looked up and used (i.e. retrieval). Results of this process should be displayed, possibly in a more advanced visualized representation.

Analysis Analysing information can refer to various activities in historical research, due to the varying methodologies used, ranging from quantitative analysis, using advanced statistical methods, to qualitative descriptions.

Presentation Various forms of presentation are used in the historical sciences, and the humanities generally. Presentation of results could also take place in earlier stages as well.

Presentation of digital historical information may also take quite different forms, varying from electronic text editions, online databases and virtual exhibitions to small-scale visualizations of research results.[4]

Long-term data management Of course, within the humanities, as in the social sciences, the data that are the results of the research should be stored for access and re-use as well, but also long-term preservation should be ensured.

[4] Boonstra, O, Breure, L, and Doorn, P. *Past, present and future of historical information science.* NIWI-KNAW, 2006. pp. 21–23. Available at http://www.dans.knaw.nl/sites/default/files/file/publicaties/Past-present.pdf.

3 Case studies

3.1 Archaeology

The Archaeology interviews were with: (i) an established university-based researcher; (ii) an early career researcher who re-used data; and (iii) a senior-researcher who conducted excavations on behalf of the municipality (local governmental archaeological agency). The three archaeologists had backgrounds in Neolithic prehistory of the Netherlands, Bronze and Iron Age of East Netherlands, and Roman and Medieval history of The Hague. All three researchers continue to work in these areas, although approaches to data and literature do vary. In this archaeology case study, quotes from these interviewees are printed in italic.

For archaeologists it is normal that an excavation produces digital and analogue data, as well as finds that may require digitization. The data collected may be compared to other excavations recorded in the digital literature archives. Additionally digital data may come from other organizations, for example elevation data, which is often used in archaeology. One of the interviewed archaeologists will include GIS and elevation in a database for the specific project. However, the gathering of other digital sources, as well collecting own resources, is often for personal research needs. "*A lot of archaeologists don't use this [digital] data as heavily as I do. Some will and there are some who still don't even create digital data.*" Archaeology is a diverse field and so is the use of digital data. None of the archaeologists interviewed use standardized digital tools, or workflows "*because it is too diverse, the creation of data*" There are a number of key sources used to start the process of gathering data. For one archaeologist, "*part of job is to look at new methodologies, new visualization, new tools.*"

One archaeologist's research is about the habitation development along a river valley, with a focus on 12 Roman sites, but using existing data from digs that took place between 1960 and 1990 and were not studied and published before.[5] Because of the period the original excavations were conducted, only a few files and images are digital. "*Bringing the old data into the digital domain is crucial and very important to this scientific community.*"

Another archaeological senior researcher is also digitizing unstudied work of past excavations.[6] The first excavations were in the 1930s, the researcher was involved in later excavations in the 1980s and 1990s and "*is now digitizing the memory I have in my head*". DANS has funded the scanning of the field

[5] Subsidized by the Odyssee programme that supports the publication of previously unpublished archaeological excavations.

[6] "Den Haag Ockenburgh: een fortificatie als onderdeel van de Romeinse kustverdediging"; also funded by the Odyssee programme.

drawings of the 1930s, new interpretations are made and misinterpretations from the 1930s are corrected. The digital databases from the later excavations are in a number of different formats "*that were not quite good enough.*" This data is now being enhanced and assessed, but updating the database systems is difficult due to organizational planning.

In archaeology most data is generated through field excavations, whereas data re-use is less common and this is borne out by the three researchers interviewed, where only one interviewee re-uses data. The other researchers tend to use the grey literature that are produced from excavations if these sites are similar to their proposed excavation. "*Only occasionally would one want to re-use data, e. g. the incorporation of house or farmstead plans from different sites and comparison between the different plans. Some required analogue drawings that required digitization, but others were already in digital form. One could combine them together at the same scale, make comparison between house plans.*"

However, it is likely, as the quantity of digital archaeology data grows, that researchers will see data sets from excavations as a source of new research in its own right. "*More people will search for existing data as a resource to answer new research questions.*" In the Netherlands, where it is required that all excavation data be deposited in one archive, "*finding data is not too much of an issue as there is only one EDNA, which makes things easier*".

A common issue for archaeology, but also for other humanities and social sciences, is the ability to search for data across current political boundaries (also identified by the political science researcher). Clearly in ancient history current political boundaries did not exist. "*If there were German and Belgian counterparts to EDNA*[7] *and it was possible to search (multilingual) across these archives it would be useful, for example like the ARENA II project.*" Also: "*There is a lot of interest in being able to cross-search countries, for example amber, which came from Denmark in the Bronze Age, you may want to investigate the distribution of amber artefacts available now and how far they have travelled from the source. [. . .] The Bronze Age just kept on going and burial mounds in Germany are the same as in Belgium, France and the Netherlands.*"

For those archaeologists interviewed, all the data required are collected before metadata enhancements are added. For one researcher, any additional metadata or annotations were for personal research use only as part of a new dataset and so did not see it as an enhancement of the existing data. It was felt that most researchers do not see enhancements as part of an ever-increasing corpus from which everyone will be able to draw benefit. "*What*

[7] EDNA is the e-depot Dutch Archaeology. For more information on this service, see 6.2 e-Depot Dutch Archaeology (EDNA).

do I need, how do I get it, what do I need to do, where can I do research, and at the end I will have a publication and a dataset. The publication is based upon work that is acknowledged, but I might have completely reshuffled everything."

In archaeology in the Netherlands, during an excavation project, everything is stored locally, as each member of the team is working on their own dataset. An excavation project is highly dynamic, where new data is created everyday from up to ten specialists. There would normally be a shared network disk for the gathering of the data and post processing after the excavation, which will be backed up daily. Data is very valuable to archaeologists because an excavation cannot be repeated, and therefore it is important to ensure that data is not lost. There will be a back-up in the field, using external hard disk drives, as well as at the university. Once an excavation project is completed, and the database does not have further edits, the data is, obligatorily, archived in EDNA.[8]

The archaeologist that is reusing existing unpublished data combines his gathered information into a database on the 12 sites. The aim is not to create a detailed study of each of the 12 sites but rather a comparative overview. However, the research will result in new data as several sources of data are combined, metadata is added and new maps, for instance of pottery distribution, are created by integrating existing GIS data. Currently the digital data on the sites is stored at DANS under restricted access and project-specific data is stored on a local computer. This will eventually be deposited at DANS when the project finishes in 2012.

Desk-based archaeological research is now possible using archives such as EDNA as the publications (excavation reports) and data are accessible. Time is saved because there is no longer the need to travel to libraries that hold the analogue publications. "*The effect of digital resources within [archaeological] research will only grow during the next few years.*"

Even the early-career archaeologist prefers to publish in highly rated, peer-reviewed journals before putting his work on a website. The choice of "*a valuable scientific journal*" is more important to him than the aspect of Open Access, even when he would receive money to pay for publication in an Open Access journal.

However, another researcher believed that the "*reputation of the journal is not as important in my area compared to other research areas, and citation index is not as important for me to get further funding*". For this researcher, it is more important to get published than getting papers into the right jour-

[8] EDNA (e-Depot Nederlandse Archeologie) is hosted at DANS using the EASY archive. Some university archaeological departments (e. g. University of Groningen) have their own data archive as well.

nals. Other researchers used conference proceedings and thus were less concerned about journals reputation or they published in "not strictly scientific" journals due to time constraints.

Interestingly, not every researcher interviewed has an online list of publications on a personal homepage or institutional website, nor is particularly worried about having an online presence. However, for some scholars, it does appear to be an essential part of their standing in their community. "*One of my experiences is that I gain a lot of visibility by having things put online. [I] always feed the institutional websites with PDFs, [however] institutional websites are a problem because they are not stable, and repositories are a better bet.*"

Two researchers identified the lack of an institutional literature/pre-print repository within DANS. The current ingest procedure into the repository is directed towards data and many of the metadata questions do not apply to a single paper e. g. the [dataset related] question "how many files does it have?" It was felt that a publication repository would give the researchers more visibility and that "*it would be harvested by other repositories that only harvest metadata*" It was suggested that this repository should be for prepress and not for publications, and should contain, for example, longer discussion papers rather than shortened papers for journal-based publication. "*If you wanted to be more descriptive than is possible in an article for a journal with all the details put somewhere and then extract a certain aspect for the journal paper. Ideally it should be kind of peer reviewed. It doesn't always have to be an international consortium, but rather internal or some sort of editorial control to ensure that quality is maintained. A level of quality that is comfortable for all the group of people who are working with it.*"

Both researchers thought that only using individuals' websites was problematic as "*A personal website doesn't have the persistence of an archive as people move their websites around.*" One researcher felt that: "*The fact you have a third party willing to publish your work properly has a psychological effect. The tangible object of a book is still nice. When it arrives boxed, it is too late to change it, and having something in your hand as a product of your effort is nice.*" Furthermore the researcher commented that the linear process, with iterations of improvement, an editorial process and a formal deadline improved the quality of the final work. "*Having a book, I can show it to my mother, rather than saying I've just had a paper published in an electronic journal.*"

Archaeological publications do not follow the normal pattern of peer-reviewed papers in journals or chapters in books, but results from excavations are typically published as reports as a result of the size of the publication. Publication is normally through institutional series. Journal articles are more likely to be in the form of a generalized article or something from an important excavation or team.

The excavation reports are usually under institutional copyright, but are made available digitally through EDNA and in local university repositories. This product of the institute is not peer-reviewed but there is an editorial team that will maintain quality. There are a limited number of national and international journals in this field, but there are conference proceedings, which are peer reviewed.

Mainstream archaeological publications are by “internal” official reports. Dutch state services have a number of series that are published and there are a high number of edited volumes. Paper versions are still the main medium for publishing your information from excavations and other archaeological studies. These are usually available in PDF, but the paper (book) is still the main version. “*I am convinced that the more openly you make your publication and data available, the more your research will be cited and re-used. If it were an Open Access PDF document then you would download it immediately and read the chapter you are interested in and will cite it. So I try to be as open with my data and publications as possible.*” Also: “*Standard [archaeology text] books would be a very valuable addition to EDNA*” available as Open Access, was a comment from one researcher who also taught on a masters course.[9] However, another archaeology researcher commented, “*customers want to have analogue printed reports*” (the PDF of the publication is sent with it as well). For example, when an inventory fieldwork is being conducted to explore if there is archaeology in the ground, the publication of it is sent to the “builder”, and these publications are freely open.

As identified earlier, there are other forms of publication common in archaeology apart from journal and conference papers, such as local leaflets for the public, local history websites and popular books about local history which are on sale. One researcher commented, “*There are lots of limitations to printed journals and books, particularly when it comes to illustrations and interactive [multi] media. For me the printed journal is a little bit out-dated. I also like hyperlinking to other resources on the Web.*” This researcher self-publishes as “rich internet publications” and considers it as a form of scientific publishing, under a Creative Commons share-alike licence, but does publish in Open Access journals as well where there is a policy such that “*you are free to use [your article] for your own academic purposes*” Also: “*What everyone should have is the right to use their own work. It is stupid to sign away your copyright and then have to ask for permission to use work you have created. [It is] completely crazy how the publishing industry is going, in that you work*

[9] One of the top downloads from EASY is the archaeology book “De steentijd van Nederland”.

for free for them then you have to pay if you want to use it (your work). [I] will look for peer-reviewed Open Access journals, but I don't think so much about the (Thompson) ranking of the journals. Colleagues have strong beliefs in other directions, but peer reviewing is very useful, if done in an honest way then it can be very useful."

An archaeologist commented: "*Persistent identifiers (PID) are very important as that will give you the opportunity to cite the data source in a normal way...Data citation [is the] same as literature: author, title, date and PID, plus which file, and this can be verified.*" Gradually, in archaeology people are becoming aware that there is the possibility, a need and a way of citing published data. This is becoming part of the normal workflow of the archaeological community, but mainly since EDNA has been in place. "*Citation of data is growing, but it will take another 10 years before it will become common.*" This awareness is growing due to champions who are promoting Open Access, who believe that it is a good thing that their data is being re-used. Crucial for the re-use of data is the usage of PIDs. DANS is currently evolving the PID system used in EDNA (and EASY) so that a single file can be referenced rather than a collection. There is a move to increasing the granularity of what can be cited.

In archaeology in the Netherlands, the storage of the publication, and possible requirements documentation, with the data (even though the publication is stored digitally elsewhere) means that the publication acts as documentation that aids the understanding of the dataset

For archaeologists in the Netherlands, access to data, grey literature and internal reports are not restricted and this is born out by the comments of the interviewees, who have not found access to data required being restricted.

For the commercial and municipality archaeologists, the list of priorities of excavation and report means that generating data is usually last and largely unfunded. The publications are done but to deposit the larger sets takes time, staff and organization and is currently not part of the workflow.

In the commercial area of archaeology, there are a number of situations in which one does not want to make the information available. There are reasons why embargoes are implemented on archaeological data: for instance, if an inventory of an area is made public, then it could be inferred that the area may be built on or used for other purposes so land/property speculators might use that data, or if there are protests against the location of a road, then it maybe possible for action groups to misuse the archaeological information/data to stall the process. There is an obligation to make everything available because it is cultural heritage; therefore a short-term embargo is used. The Dutch state service has 2-, 4- and 6-year embargoes.

Treasure hunting may also be an issue however; this is not really a problem, as the publications are not released until 2 years (normally) after the excavation has been completed. Treasure hunting is often used as an argument, but it is not a good argument to keep archaeological data private, because if a site is important then it will be classified as a monument and therefore gain legal protection. Before every excavation the archaeologists must inform the state service, who will check that the location is not a listed monument, and provide a report after the excavation has finished with a structured summary. With larger projects, there is a programme (methodologies) requirements documentation that describe how the excavation will be undertaken.

One archaeologist deposited his doctoral data in EASY at the e-depot (EDNA), at first with an embargo to restrict access and then later making it available openly. The reason for doing this was so that the data would be assigned a PID, which could then be cited in the PhD. Only after the PhD was defended the data was made public. Although there is not strictly a legal or privacy requirement for an embargo, it is an example where embargos on data are useful. Shortly after changing it to Open Access it appeared in the top ten downloads in EASY.

Archaeology produces a lot of data but communication of results occurs at the report publication level. Within archaeology, it is very difficult to draw definitive conclusions about data. "*Sometimes it is more interpretation of the type of artefact, the date or the cultural relevance/significance: it is all knowledge-based inference and interpretation. If you gave the same artefact to different archaeologists they may come up with different interpretations of what it is, or what it is used for. Even measuring the length of an artefact is open to interpretation as most are broken. [. . .] Data is open to interpretation, and conclusions are soft, so this makes people cautious of reusing data.*"

Another comment was: "*It is not a formal part of research to share the data as part of an university department, but it is for commercial archaeologists in the Netherlands.*" Normally there is no checking whether data is deposited. "*There was the idea that they [funding organizations] would not pay the last 10% of grant if you did not deposit your data and DANS signed off that you have done so. However, I don't know if they have followed through on the threat.*"

Archaeology in the Netherlands is a small community of researchers, so it is normal to share data on an informal basis between researchers upon request. "*If someone wants to extend your work they may contact you, for example, for your GIS file, and perhaps ask to re-use it.*" One archaeologist is willing to grant individual requests for access to data when it is being studied and generated, but will not yet put the data on open. They are always welcome to ask for the data before publication.

Main issues in the field Archaeology

- There is a need for retro-digitization: digitization of data gathered before the start of the digital era.
- Excavation results are normally speaking published in reports. Excavation results should be distinguished from research data in archaeology. These are selected and collected from the excavation results and subsequently analysed by university researchers as part of their normal academic research. These research results will be published in peer-reviewed academic journals.
- The interest for, and possibilities of, reusing archaeological data is clearly growing. There is, however, still a way to go regarding data sharing and standardization of metadata. National repositories like EDNA can help here.
- Sharing of data may be made easier by the implementation of standard data formats and by clearly defined embargo regulations.

3.2 Political science

One researcher from a political science and computer science background now works with political data from many sources. He enriches these data with all kind of connections between the different data objects. In this way he uses the Dutch parliamentary proceedings as a secondary data source. This procedure leads him to comment upon the quality issues of open and publicly available data. In this case, it could be safely assumed that the data source was "authoritative"; however, when errors were discovered, the researcher was initially "blamed".

This researcher works with a broad collection of data of many formats, including controlled vocabularies (in all 24 EU languages); however, the main data source is the proceedings from the Dutch parliament (and others). Other sources include: political blogs, RSS feeds, Twitter streams and newspaper articles with user-generated comments. The very large Dutch proceedings are all scanned and are complete from year 1814 to current. All the data used exist as digital sources, but some scanning is undertaken. These sources are often structured text, but the structure is not explicit, so there is considerable effort in making the implicit structure explicit and machine-readable

When it comes to political data, the most effective way of managing and enriching the large quantity of data that is used is to download and store all data sources locally to the research team. This is because the most costly process is transformation of the data. "*The raw data is stored on disk, and backed up to a dual-redundant RAID. The transformed data is stored on disk in an eXist XML database which is also backed up. The database has 20 GB of text, which is a significant amount of data.*"

The data coming from the Dutch parliamentary proceedings and other sources have been greatly enhanced. The text from the proceedings is enriched by cross referencing, based on named entities, for example to the names of speakers (their bio page), political parties, dates and controlled vocabularies of political issues (which is the hardest). Hyperlinks to laws, which are identified in the speech (with a specific number that points to the legislation), are also included. Votes are extracted from the data. New queries can be asked, for example "everything said by a person" which is a completely new view compared to the documentary of day by day, and analyse the language used by a speaker over time.

The data enrichment is automated using machine-learning algorithms; however, the rules are hand coded. This is because the structure is quite consistent throughout the corpus. It is also consistent between countries as six countries use the same the proceedings methodology, so it is possible to apply the same automated enrichment. "*Then it is possible, with machine translation, to work with a large database across political boundaries.*" The addition of Dublin Core, TEI (for text mark-up), and ISO country and language code metadata is also automated. "*Persistent identifiers are added to every paragraph to make very specific linking possible.*" This means millions of identifiers are generated for a data collection, which is unique amongst the researchers interviewed.

This research is the only interviewee who published data directly after enrichment and before articles are published. This research team, which is enriching the parliamentary proceedings, use the community of users to check the data, "*which is the main advantage of openness*". They have discovered that it is much better to be completely open and honest, that there are a lot of mistakes and be willing to hear from users and correct the errors. "*If it is open immediately then quality becomes important to you because people will be checking it, using it and building on it. Openness will just improve quality [of data] through these simple social mechanisms. Open data can be a really big thing.*"

The researcher did identify a number of disadvantages to this approach. For example, if someone takes the data, builds upon it, and researchers using that dataset finds that there is a gap or it is not as reliable as they think, then the use of the data is already quite far away from the origin so it is difficult to identify the cause of the problem. Enriching raw data created by others makes this team brokers between the original data producers and the users. Regarding the original sources of data used by him: "*you would think [these] are more authoritative than we are, but actually they are not. This is dangerous and could kill your reputation*" as the researcher will be blamed for the errors in the enriched material which come from the original source material. This method of data publishing is not typical for the field, and the researcher believed it to be "state of the art", since "*most others in field do not publish data, or use it is a bench mark. The data itself is not valued*". Also: "*Everything we do is completely public so it is quite dangerous. Once it is in the eXist database it is public. There is a URL to every item in the database and you can query it. This is highly public and therefore this makes it extremely easy to share. It is important to be fast [with corrections]. Only when people use it you find mistakes.*"

Main issues in the field Political science

- Datasets in many formats and are used with controlled vocabulary.
- In this particular area new sources like blogs, RSS-feeds and Twitter are becoming important, but some of these are also extremely difficult to get hold on for researchers.
- Enrichment of existing data (texts) is an important activity in this discipline.
- Openness of data leads to enrichment by the research community.

3.3 History

3.3.1 Oral history

Quantitative oral history recordings are the sources material for one specific scholar. An important aspect of these recordings is that the data collected from individuals may be sensitive.

The researcher gathering oral history data does so "*with the aim of making this information accessible to others*", and this is "*implicit in the discipline*". The data used can be very sensitive, however, the data is often being "*handed down on the basis of trust*". The researcher considers that for each form of qualitative data it is necessary to have a specific protocol, for the people that are questioned in an interview and for the people that want to use that information, and this cannot be generic, "*because then it would be useless*". In the USA there are strict review boards to assess the research protocols used. The interviewee thought that this might be useful in the Netherlands, but argued that such review boards might make research very inefficient because of the difficulty of getting permission to use the data. The difficulty with the oral history data is the selection being made. Statistically the selection is often not representative of the situation or the whole group of interviews: "*it is the group of people that want to be heard that become part of the selection and not the ones that do not want to be heard*" This is important to realize when using this kind of data for research. It was suggested by the researcher that there should be a pilot project to test the sharing of qualitative data in Europe, but suspects there will be problems and solutions will be found. Eventually the sharing will work but the researcher stressed the importance of respecting peoples privacy and the contract someone signs need to be very clear on this aspect.

For the oral history interviews, the researcher established an embargo commission to restrict access to the recordings and data in sensitive cases. Four people with very different backgrounds who will provide each a different insight on the embargo cases form this commission. People who are being interviewed are mostly not asked to give permission for the use of everything they say by researchers, and this situation should be taken more serious. The commission should prevent sensitive information becomes public when the person in question does not want this to happen.

There is also the issue of how long information should be kept private. Sometimes it is desirable to release the embargo before the set date, therefore it is important to put something to cover that in the contract. For example when the person in question brings his information to a public source, the original holder of that information is then permitted to release the embargo and use the information publicly as well.

For researchers in this field, a major concern is that the research data created as part of their work is not re-used for commercial purposes, particularly when this is interview data gathered.

3.3.2 Cultural, social and economic history

A scholar in the field of historical demography, working with the HSN (historische steekproef Nederland/historical sample of the Netherlands[10]) noted that "*most historians are not able to work with digitized HSN material*", so his assistance is required in preparing the data. In such cases the scholar will be a coauthor of the publication. This researcher also commented that he had "*easy access to everything I need*" with respect to data and literature.

A researcher investigating intergeneration mobility (in the field of professions in education,) also creates dataset based on HSN. However, this researcher commented that "*HSN is not user-friendly and it is difficult to create subsets*". The subsets will not be published or stored, as "*there is no facility for this. HSN should create space for storage of subsets and results of research*".

This researcher re-uses sociological surveys (statistical data) archived on EASY as a historical source and mainly uses digital sources of data. However, there are obstructions to using these sources such as the need for registration

[10] http://www.iisg.nl/hsn/abouthsn/index.html.

and specialist skills to handle the data. Furthermore the metadata may not be of a high-enough quality to make full use of the data source.

Other researchers tend to store digital data locally mainly because it is work in progress or because the researcher is working on their own, but other comments include: "*I lost data while working in a team due to overwriting of files stored in cloud*", "*I have relatively small amounts of data*", "*part of the material is digital, and a part is still analogue*". Back-up on institutional infrastructure seems to be the normal way of ensuring that the data is safe, but some additionally store data at home on a private PC as well.

Working with the HSN (historische steekproef Nederland/historical sample of the Netherlands), the users are obliged to add the release of the project-specific subset of HSN with the publication in order to make it possible to re-use the data. This is a form of data citation, but again the researcher commented that the rules for referencing are not formalized. Another researcher believes that "*data creation should be given credit*" and applies the citation rules set down by the data producer; however, even in his own project this facility has not been implemented.

For one researcher, the right of display of copyright material on academic research websites is problematic. This is because cultural heritage institutions are "*exploiting their rights by asking large sums of money for a single image. Mostly they are generous to academia but it should be free, as we have paid for the paintings and objects in the museums and it should be free on a website for everybody to re-use, but obviously not for commercial gain*". It can take considerable time and money to ask permission and obtain the right to re-use each item, especially in a visually rich digital publication.

There is no obligation for researchers to publish data, and at DANS this is currently free for self-depositors. In the same way as paying for publishing an article in a journal, it is likely that depositors will have to pay for data archiving. This researcher considers this is better than letting the user pay because that is against the idea of Open Access; however, this paid deposit is something that is a point of discussion for the future. " *There is an opportunity for humanities to take the forefront, because we do not have many commercial barriers, and researchers in humanities want a big public. In the humanities a considerable result can be achieved with relatively little energy. Next to intellectual talents, you need extra resources, an infrastructure to store and re-use data and research outcomes.*"

Main issues in the field History

- Openness of the oral history data often conflicts with its confidential character.

- Skills to handle complex datasets are sometimes lacking within the discipline.
- Low quality of the metadata hinders optimal re-use of the data.
- Often cultural heritage institutions ask fees for accessing their data, in combination with copyright barriers, so that researchers have to negotiate with them to (re-)use the data.

3.4 Law

For law research, there are two main sources that can be considered as data. These data, more commonly referred to as resources, have not been created *within* the law research environment, but in the administrative world, including the administration of justice.

The first type of source is the jurisdiction consisting of all the judgments issued by the law courts and other bodies administering justice. In the Netherlands, only a limited number of the latter category is publicly available on the internet, mostly only those cases that are important from a jurisdictional point or having large public attention. In the law world there is an ongoing debate on the desirability of making far more (or in fact systematically all) judgments available online. There is, however, strong resistance against this, mainly based on privacy and financial considerations.

Theoretically a solution to this problem could be to deposit (all) these judgments in a research repository to which only law researchers (i.e. not the general public) have access. The Personal Data Protection Act of The Netherlands (Wet Bescherming Persoonsgegevens, WBP) enables this possibility. Access to data files containing personal data is allowed for “academic, statistical or historical” research. This principle is based on the European Directive on personal data protection. There is, however, a pessimistic view on the materialization of this in the near future. Too often, privacy reasons are used as a way of avoiding publication, but there are certainly ways to get around the issue of personal data protection. There is already software solving this problem available. From the political world, attention is increasing towards the quality of administering justice. This could be seen as a hopeful development, as it could lead to more attention to the problems mentioned. Other sectors of the society can be considered as stakeholders here, such as journalists, but also lawyers. Besides from the privacy issue, there is a financial hurdle: who is going to pay for this? Anyway there is a clear lack of willingness in the juridical world itself to do this and to give these “data” away to another (research) organization.

There are the laws and official regulations at all kinds of level themselves. These are available, nowadays mostly in digital form on the internet, albeit not in the most ideal format. Regarding the publication of laws, another important source for law research, the situation is considerably better, but far from perfect, as it can be very difficult to reconstruct the development of certain laws or regulations over time.

According to some, mostly younger, researchers, research in the law faculty is "*old-fashioned*" legal, mostly textual, work on jurisdiction and law making. Law researchers on the whole are not very much interested in more quantitative analyses, in the direction of social science. This could explain why there are not so many resources available for that kind of research. These younger researchers very much would like to have available a substantially larger corpus of law resources, in particular court decisions on all levels (district, appeal, etc.) than is presented today.

As for academic publications on law research, most journals do not have Open Access policies at all. There are, however, some hopeful developments here as well. There is for example the *Leiden Law Review* of Leiden University. In this repository there are often Open Access articles. Most articles are published in Dutch, unfortunately.

Another field, strongly related to law, is that of criminology. In the way research is carried out here, there is a strong difference with that in law properly, as described above. Criminology is a form of social science and research methods are those as used in the social sciences. There are research projects with strong qualitative elements, but the whole research process still can be considered as social science. Access to data in this field is also mostly very restricted. This is again mainly because of privacy reasons. Most of the research data are of a highly sensitive nature, including personal data on youth offenders or criminals generally. The Ministry of Justice only allows access to the data by researchers under very strict conditions of use.

Contrary to Open Access to law data, both for law and criminological researchers, publications are not in particular problematic regarding Open Access, as these would normally only contain anonymized data.

Main issues in the field Law

- For research reasons, all judgments should be available online, at least for academic researchers.
- Most law research journals do not have Open Access policies yet.

3.5 Economics, social science

Looking from the perspective in what way Open Access might be beneficial to research infrastructures in these fields it should be observed that, as in law, economic researchers are using, for a large part, data which has not been created *within* their research environment. Financial research uses data from commercial firms as well. That means that these data are not always easily available for researchers and, if available, can have (sometimes very strong) restrictions on dissemination after processing. This could, to a lesser degree, also hold true for publications based on these data.

In economics research, data created in the administrative world are far more important than survey data, which are often created by researchers. One institute with a central position here is the CBS – Statistics Netherlands. They have a large quantity of data that could be linked with each other as well as other data easily, technically speaking. In practice, however, things are not that easy. In the world of academic research the CBS is seen as a difficult organization from which to obtain data, in particular microdata to be linked to other microdata. The CBS has a strong policy of protecting personal data. In particular data on the financial world are difficult to obtain. The willingness of the central banks of Europe to make these available varies from country to country. Even more difficult are data from private banks as well as commercial companies generally, and in particular multinationals.

It seems that most researchers in economics are very much in favour of a very limited period of embargo. One researcher sees as a maximum, a period of 10 months, which is considered as short, especially by PhD students. According to this researcher, on average, a period of 2–4 months is the needed for translating all the data labels and relevant information into English before the data are published. In his field (economics, household surveys) a longer period is unacceptable; information would really be outdated. One thing of importance is international usefulness of the data: that is why all documentation is obligatory in English. He would like to have as much Open Access to the data as possible, only restrained by privacy laws or contracts with commercial parties. However, users of the data should always be traceable. Registration is absolutely necessary.

Also for publications, he is very much in favour of Open Access for all publicly funded research. According to him, there is not yet enough awareness amongst researchers for this, when they are publishing. They are giving their copyright away too easily.

Interestingly enough however, there are conflicting views on this in the social sciences. Another researcher, a professor in the social sciences was rather hostile towards the idea of Open Access and she wonders where this comes from. Her main reservation is that the data collected and processed by a team of dedicated researchers over long periods cannot easily be understood or appreciated by other researchers because of lack of methodology and consequently potential wrong use of data. Furthermore coauthorship of the original investigator should always be appreciated. According to her, research teams which do not provide Open Access to data are publishing far more than those who give full access by way of the internet (for example, in the Netherlands the large-scale projects TRAILS versus NKPS). In her major research project, a large data corpus exists which is extended over a number of years. All these additions continue to enrich the dataset. Other researchers are invited to use this material but under strict conditions: they have to consent in members of the research team being coauthors of the publications.

Another point made by her is that, if universities, funders (like NWO in the Netherlands) and others really should enforce Open Access to data on researchers, social sciences undoubtedly would reduce its lead over the medical sciences regarding the "easy" availability of research data. In the medical field data acquisition, in particular from patients, is already a much more bureaucratic process.

Regarding publications she admits that "publication inflation" has been going on in academia for years, being very focused on articles. She would very much prefer a system in which there is more attention for quality than quantity. However, she does not see developments in this direction. In social sciences, hardly any high-quality Open Access journals exist yet.

Main issues in the field Economics, social science

- Financial research data are not always easily available.
- Re-use of financial research data is often restricted.
- Economics resembles fields like biology and chemistry in that only no embargo or a short embargo period is acceptable for the progress of research.
- The number of high-quality Open Access journals is (still) low.
- In social sciences, "Open Access" is still valued very differently.

3.6 Linguistics

Literature is a data source for some humanities research and as identified by one researcher it is sometimes a barrier to conducting research if there is no access to an academic library that pays for journal subscriptions. This researcher would pragmatically tend to use literature that is Open Access "*because it is easier to get hold of*".

Only one researcher was interviewed who came partly from the linguistics world and also partly now works for the linguistics digital infrastructure CLARIN. In linguistics copyright is a very large problem, in particular for using large text corpora and text mining. The large and growing text databases of newspapers for example, is of great interest for scholars of contemporary language use, are not available for research, except in a very limited way. It is his experience, however, that it is possible to circumvent the copyright restrictions of commercial publishers, as long as there are no commercial interests involved. Good communication with commercial partners is essential for this to succeed.

This researcher confirms that he is very much in favour of Open Access, fully in line with CLARIN policies. The only two unavoidable barriers are privacy protection and copyrights. The copyright barrier, in particular, could partly be surmounted with extra effort and ingenuity by the researcher. Privacy protection could also be lifted for strictly academic research. Under certain conditions researchers are allowed to access confidential personal data, as long as they only publish about their findings in *anonymized ways*. Concerning embargos, he would advocate temporary restrictions only on the ground of protecting a PhD student and then until the actual graduation moment.

Main issues in the field Linguistics

- Research is built on existing text corpora.
- Copyright hinders re-use (for instance via text mining) of the resources. There are ways to overcome this problem.

3.7 E-Science

Many of the researchers interviewed are difficult to categorize, as they often conduct research that could be said to be at the interface between disciplines. This can be particularly seen with one interviewee, a senior researcher who has a background in physics, mathematical modelling, philosophy and history of science, and who is now conducting research in modelling the economics of innovation (social sciences).

This researcher uses large bibliographic databases to investigate the growth of a research field and innovation by extracting the network information of collaboration and citations. However, the researcher identified the difficulty with access to "*cleaned data, which is essential*", but these sources are private. Other large data sources, such as the Web of Science database (a favourite) has a very specific view, and in this instance is only a selection and is biased towards English language and specifically American journals. These difficulties in finding suitable openly available sources lead the research team to use a Wikipedia "dump" from 2008. Although, with Wikipedia, they have "*as much reliability with Wikipedia as you have with Wikipedia*" They observed a "*lot of self-correction and self-cleaning going on*"; however, this does not guarantee 100% correctness. "*You do a selection of data sources that are selections in themselves. [...] The bias doesn't come from the data sources available, but I think the bias probably comes from the research attitude and the epistemic knowledge inside of the field.*"

For this researcher whose team is using the Wikipedia dump, a selection of the raw data was made and statistical methods along with clustering algorithms were tested. The volume of data challenged the algorithms, with some 2.8 TB of raw data to process the BigGrid grid infrastructure was utilized to both store the data and run parallelized versions of the analysis algorithms. The data was parsed and extracted, looking at changes over time. Rapid changes that occur in a short period of time which represent vandalism, or "noise", were discarded. The decision of how and when the data was agglomerated to feed into the statistical analysis tools. The algorithms were written in Java, and for clustering existing algorithm were adapted. Analysis of the results files was also run in parallel on the grid. "*As a humanities group they [BigGrid] took us on board as a pet project. We were one of their successful Humanities projects.*" Data generated was also stored on the grid. When the formal collaboration with BigGrid finished the secondary data was downloaded to a laptop hard drive, and a backup was made to an external hard disk drive. The Wikipedia dump was also stored on a desktop computer as a 100 GB zipped file. "*Currently the data is not accessible, but planning to do so.*" The team of the researcher did not make qualitative annotations, "*but more quantitative notes, done very primitively using excel sheets, or looked it up in the raw data and just counted the occurrences e. g. we compared number of times a term came up Wikipedia and UDC library data.*"

The range of publications and forms of publication is again broad in the humanities and social sciences. Clearly, peer reviewing is crucial, and the reputation of the journal is the first priority before accessibility, but some researchers do wish to reach the widest possible audience including other forms of media. "*Definitely peer reviewed but making the pre-press prints available as soon as possible as well.*" Also: "*Would I choose a journal according [to] its public availability? No. I would choose a journal according to its standing in the community rather than gain visibility.*"

The time take to get a paper published in a journal is an issue for some researchers and this does influence the choice of journal. "*We have submitted a paper to a journal more than a year ago, but they are not so fast at putting together the special issue even though they are only short papers.*" And: "*I'm most interested in short turnaround on publishing and don't like to wait for half a year or more for publishing. [In my field of study] things are out of date in 2 months time. I prefer to publish [literature] online as Open Access/open source, e. g. in the online journal of document information, [where] authors keep copyright.*" Also: "*Submitted a journal article some time ago, and recently, before publication we have found that some hyperlinks didn't work.*" In the first case the researcher was concerned whether publishing a pre-press version locally or depositing it with a subject-specific pre-press repository would risk the inclusion in the journal. The researcher also mentioned that the ability to deposit an article in a pre-press repository "*depends upon who you are talking to in the [publisher's] administration and how you bring it [to them]*".

Main issues in the field E-Science

- Open available resources are sometimes difficult to find.
- Sharing information is in certain situation more important than publishing articles in high-quality journals.
- Time between submitting a paper to a publisher and the actual publishing is too long.

3.8 Important general issues

Apart from the discipline-related issues, there are issues relevant to the whole field of humanities and social sciences.

- Time between sending in a manuscript and the publishing of the final article takes too long.
- Researchers should be given credits for creating datasets.
- There is a strong need for searching datasets across political boundaries.
- Persistent identifiers given to both traditional publications and datasets are important for the retrievability and citing of resources.

4 Current status of Open Access

4.1 Data

For the Netherlands, EASY is the main access point to datasets in the humanities and social sciences. DANS gives access to the description of datasets, but the depositors determine the access rights to the deposited datasets. There is no totally free Open Access within EASY, but access free after registration. This registration is used to identify meant to retrieve who the users of the datasets and generate usage statistics.

In EASY there are 19,900 descriptions of datasets (August 2011), and 8583 of these datasets are Open Access. From the other datasets, 1743 have some form of restricted access, 9135 are related to the restricted-access "Archaeology" group and 439 have another form of restriction. Major concern of DANS is that all datasets are described in EASY. The policy remains, however, "open if possible, protected if necessary".

Summarizing the findings from the case studies, it could be said that generally speaking researchers seem positive to Open Access to data. Some reservations, however, could be observed, as discussed below.

4.1.1 Personal data

Researchers are clearly extremely cautious in handling records containing personal data in whatever form. This is both for ethical and legal reasons. The European laws are very strict, in some research fields still stricter than in others. In particular, medical data can only handled with extreme care. In other research fields there is a grey legal area. The privacy laws are not always perfectly clear: what is still allowed to do and what is not allowed anymore? Specific examples mentioned are oral history interviews on sensitive topics like war memories and deaf-mute children (not being patients) filmed for research purposes. Making these kinds of personal data available on the internet, even restricted for research reasons, creates all kinds of challenges of safety and security. There is a common understanding amongst researchers that these data have to be handled, disseminated and archived with great care. Handling personal data is, however, unavoidable in research and should certainly not be made impossible or deterred by stricter regulation. Clearer regulations on some points should certainly help.

4.1.2 Copyright issues

This barrier to a free and open use of material that in some disciplines (linguistics, social science, economics) is indispensable for research is often mentioned. There are clearly huge hurdles here. It could, however, also be concluded that there seem to be flexible and creative ways to go around this problem. A more generic legal solution would always be preferable, like a general exemption for research purposes. This exemption exists more or less for privacy data. A European directive would help here.

4.1.3 The reluctance of scholars to share data

This is an attitude that could be found in particular amongst social scientists, in particular in social psychology it seems. The main reason here is protecting research data on which research teams worked for years in which a lot of money is invested. The great fear is that other researchers will gain an advantage, in particular too soon, and steal the show. Another, maybe even stronger fear is that other, non-competent researchers or amateurs will misuse the data and willingly or unwillingly misinterpret them.

Regarding the last point the only concession that would be acceptable from the perspective of Open Access to data is temporary protection in the form of a short embargo. This is acceptable for most researchers, especially when it concerns PhD research. Opinions differ on the duration: 1 or 2 years are mostly mentioned. In humanities (history, archaeology), researchers do favour on the whole longer embargo periods than in the social sciences and in particular economics. Especially in the social sciences, data may become outdated within a relatively short period. In this field, a long embargo period may be a serious hinder for research progress.

For the rest it seems a question of mentality, maybe of generations of researchers. There are indications that younger researchers are more inclined toward Open Access generally. On the other hand, as some interviewees hinted at, younger researchers could be more protective as well. They still have to make a scientific career with their data. It should be mentioned that funding organizations are becoming stricter on this point. The largest research funder in the Netherlands, NWO, has recently adopted a stricter policy on Open Access to data, which will make it impossible for researchers to keep their data locked away from anybody else for years.

Implications for OpenAIRE

– Especially in disciplines like health sciences, social sciences and history researchers are dealing with personal data. There is a need for a European regulation that will allow the handling of these data for research purposes.
– Apart of the settlements for Open Access to publications and created datasets (datasets that are produced during the research), a copyright regulation for existing resources like text corpora or financial data that would allow free access for academic research would benefit the progress of research in disciplines like linguistics and economics.
– Advocacy in favour of Open Access to data is needed to overcome the reluctance of scholars to share their data. The readiness to share date could be augmented by allowing an embargo period of 2 years as a maximum, possibly dependent upon discipline.

4.2 Publications

As already explained, the differences between the disciplines within the social sciences and the humanities are broad. Nevertheless, one can conclude that there is a tendency in favour of Open Access, although this impression coming from the interviews will have to be confirmed by additional questionnaires amongst the researchers in this field.

Traditionally, the book plays a more important role – especially in the humanities – than compared to the MST fields. Open Access to books has just begun, as publishers were reluctant to give up their traditional business model. An interesting project is OAPEN[11] in which European publishers cooperate in producing Open Access books in the fields of Humanities and Social Sciences. Meanwhile, the OAPEN site offers access to 832 books (August 2011).

Journals are less important than in the MST field, partly because of the big differences in editorial policy. Peer review is not always possible as in for instance the medical sciences. Nevertheless, the Directory of Open Access Journals (DOAJ)[12] counts the number of journals per (sub-)discipline, as shown in Table E.1. Double counts in DOAJ are possible, but one can see the vast amount of Open Access journals.

The number of available journals in DOAJ is given in combination with the figures from Thomson's Citation indexes.

Table E.1 Number of journals per DOAJ category

DOAJ category	Soc.Sci.Cit. Index	Art and Hum Index	DOAJ
History		273	181
Archaeology		81	32
Religion		128	78
Philosophy		165	163
Linguistics, language and literature		310	428
Social sciences	2475		1445

Other resources for Open Access publications are the repositories Social Science Research Network (SSRN[13]) and Research Papers in Economics, RePEc.[14] These very popular repositories (with a high rank on the Ranking Web of World Repositories[15]) are comparable with PubMedCentral in the biomedical field.

The interviews revealed an important new aspect: the major role of data in the production of publications. This role is that strong, that one could see a shift from traditional publications (articles, reports, books) to enhanced publications (publications that are a combination of the traditional publication and the underlying datasets). This new way of publishing will stimulate the re-use of datasets. Unfortunately, Open Access to these enhanced publi-

[11] http://www.oapen.org.

[12] http://www.doaj.org.

[13] http://ssrn.com.

[14] http://repec.org.

[15] http://repositories.webometrics.info/toprep.asp.

cations is somewhat cumbersome. This is caused by the fact that the Open Access status of the different components of an enhanced publication will vary.

In comparison with other disciplines, Open Access has not reached the same level of importance, but the trend is positive.

An alternative for the traditional publisher is the publishing of research results via institutional repositories. The instability of the URLs is, in this respect, a source of concern. A system assigning PIDs to object could resolve this problem.

Finally, researchers do accept that, in some situations, an embargo period will be necessary.

Implications for OpenAIRE

- Although there is a tendency in favour of Open Access to publications, the fields of humanities and social sciences are too broad to come to definite conclusions. A more detailed questionnaire is needed to validate this preliminary conclusion.
- In the humanities and social sciences, there is a very close relationship between the (traditional) publications and the resources that have been used in writing these publications. Therefore, there is a growing need for so-called "enhanced publications". In these, the relationships between the different resources can be made clear to its readers. This important new development will only be successful if all components of an enhanced publication will have the same level of accessibility (as open as possible).
- Related to both enhanced publications and the re-use of objects deposited in (institutional) repositories is the problem of sustainable access to these objects. Sustainable access could be reached by introducing PIDs to any type of object (text, dataset, image and so on), in combination with a European resolver services that will lead a user from a PID to an actual URL. The next stage would be the connecting of future continental resolvers to realize a world-wide system of resolver nodes. Concurrently, there should be developed a policy of discouragement for using non-persistent URLs to identify objects.

5 Current research infrastructure projects

DANS is involved in a number national and international research infrastructure projects both in preparation and development. These projects may include the archive facility or infrastructure in a particular research area. They may also relate to availability, to the software used or to technical aspects of the desired data infrastructure. The goals of the projects are the stimulation of collaboration and the sustainable access to data to serve data-intensive research.

5.1 AlfaLab[16]

AlfaLab is a 2-year project (2009–2011) that aims to redress the lack of co-operation and the absence of a technical information infrastructure which forms an obstacle to humanities research in the battle for resources to successfully compete with other sciences. The project initiators of AlfaLab aim to develop a digital infrastructure that may take the form of a digital portal and a virtual laboratory (research area) for alphas, or in short AlfaLab.

AlfaLab is an initiative of the KNAW (Royal Dutch Academy of Sciences), where five scientific institutes which have the objective is to cooperate and promote the use of digital methods within humanities research. AlfaLab disseminates knowledge about digital tools and data. AlfaLab creates and develops digital tools for the humanities community and investigates the use of digital tools in the humanities and how these (virtual) tools support and encourage partnerships.

The goal is to create a digital infrastructure for the humanities, consisting of the following components:

- laboratory: modifying, linking, providing and implementing innovative ICT tools for manipulating and analysing digital resources for the humanities;
- portal: developing a common access to these tools and to the available Dutch data files (a digital portal for the humanities);
- disseminarium: spreading knowledge about new research opportunities by using this infrastructure in larger groups of researchers.

The project focuses on three elements: (i) the Tekstlab (focusing on textual sources); (ii) the Spacelab (focused on geo-data); and (iii) Lifelab (focusing on lifelong population data).

The ultimate goal of AlfaLab is to develop a significant contribution to the humanities infrastructure within the Netherlands. The role of DANS in AlfaLab is to design, develop and implement of the portal environment (the ICT infrastructure where applications and data are being made accessible).

[16] http://www.dans.knaw.nl/en/content/categorieen/projecten/alfalab.

5.2 Connecting ARchaeology and ARchitecture to Europeana (CARARE)[17]

The ambition of CARARE is to ensure that digital content for Europe's unique archaeological monuments, architecturally important buildings, historic town centres and industrial monuments of heritage importance is interoperable with Europeana[18] and accessible alongside contents from national libraries, archives, museums and other content providers. CARARE aims to enable spatial and virtual reality content for heritage places to be brought together in Europeana and new services for users.

CARARE will add value to Europeana and its users by:

- promoting and enabling participation in Europeana by heritage agencies and organizations, archaeological museums and research institutions and specialist digital archives, and raising awareness of Europeana in the domain;
- establishing an aggregation services which contributes on a practical level to enabling interoperability, promoting best practices and standards to heritage organizations, taking account of the particular needs of content for archaeology and architecture. It will bring 2 million items (images, maps, plans, aerial photographs and 3D models) for Europe's unique archaeological monuments, historic buildings and heritage places into Europeana;
- implementing Europeana compatible infrastructures, standards and tools so as to make available millions of digital items for heritage places across Europe, thus contributing to the growth of Europeana;
- acting as a test bed for Europeana's APIs that are intended to make content available for other service providers to use, for example in the areas of tourism, education and humanities research;
- establishing the methodology for 3D and virtual reality content to be made accessible to Europeana's users.

CARARE runs from 1 February 2010 until January 2013 and is funded under the European Commission's ICT Policy Support Programme.

[17] http://www.carare.eu.

[18] The Europeana service offers access to millions of digital items provided by Europe's museums, galleries, archives, libraries and audio-visual organizations. Some of these are world famous; others are as yet hidden treasures. Europeana will deliver public access to over 15 million digital objects by 2011. http://www.europeana.eu.

5.3 Council of European Social Science Data Archives (CESSDA)[19]

The Consortium of European Social Science Data Archives (CESSDA) has been an umbrella organization for social science data archives across Europe since the 1970s. The member organizations have worked together to improve access to data for researchers and students. CESSDA research and development projects and Expert Seminars have enhanced exchange of data and technologies among data organizations.

In 2011, CESSDA is working to become a truly integrated European data infrastructure, with legal personality and full legal capacity, preferably with the legal status of a European Research Infrastructure Consortium. CESSDA is one of the 44 projects listed on the ESFRI Roadmap. The Netherlands is one of the countries taking part in CESSDA-ERIC, with DANS as the Dutch service provider.

CESSDA-ERIC will provide a distributed research infrastructure expressing the principal tasks as follows:

- facilitate access for European social science and humanities researchers to the data resources they require in order to conduct research of the highest quality, irrespective of the location of either researcher or data within the European Research Area (ERA), and beyond;
- coordinate and develop access practices, agreements, licensing models and any other legal and organizational measures that enable and extend such access to distributed data resources;
- coordinate and support the installation and maintenance of a technical infrastructure that allows such access to distributed data resources;
- actively contribute to the development, promotion and adoption of best practice for data distribution and data management, thereby enhancing the quality of infrastructural services;
- work continuously for the inclusion of further data sources, from Europe and beyond, into the infrastructure;
- provide training within the CESSDA-ERIC and beyond on best practise in operational processes and data management.

5.4 Common Language Resources and Technology Infrastructure (CLARIN)[20]

The project “Common LAnguage Resources and technology INfrastructure” (CLARIN) provides researchers access to language, text and speech resources

[19] http://www.cessda.org.

[20] http://www.clarin.eu.

and research tools across Europe. The services include archiving and re-use of data, as well as advice on metadata and standards. This way, CLARIN stimulates the exchange of knowledge and data between linguists, historians, speech technologists, communication researchers and many others. An important aim of CLARIN is the availability and usability of resources such as lexicons and text corpora for each language within the European Union.

DANS participates in a number of research projects within this framework. Once the partners in these projects have achieved all goals, they will become CLARIN centres. DANS is one of five Dutch organizations with this ambition.

Other CLARIN projects that DANS participates in have a linguistic nature. They aim at automatically extracting and/or curating linguistic data, as well as developing demonstrators. DANS is involved in the projects because of questions regarding software archiving and persistent identification of small text fragments.

Finally, the Netherlands Organisation for Scientific Research (NWO), CLARIN, and DANS have agreed that for certain applications for NWO grants for humanities research DANS informs the researchers about CLARIN standards that the research should live up to.

CLARIN stipulates that researchers and research groups participating in CLARIN make their resources freely available to other researchers. For this, CLARIN-NL refers to the *NWO Open Access Initiative* for publications but extends it to research data and tools. CLARIN demands from all projects that the deliverables will be accessible for the CLARIN community on a (future) CLARIN central server.

5.5 CLIO-INFRA[21]

CLIO-INFRA is embedded within the European Commission Initiative Digital Research Infrastructure for the Arts and Humanities (DARIAH). Within this NWO-funded project, the goal of the CLIO-INFRA project is: “On a global scale, bringing different and sometimes fragmented data sources together through alliances in an open access model disclosed with the use of a central portal”.

The purpose of CLIO-INFRA is the systematic mapping of the available quantitative information on the development of the world economy in the last 500 years. The goal is to provide a solid basis for the systematic study of the causes of global inequality. By using CLIO-INFRA as a foundation one is able to design and test crucial economic and social development theories. CLIO-INFRA shall interconnect a number of databases (hubs) consisting of

[21] http://www.clio-infra.eu.

data on global social, economic and institutional indicators over the past five centuries, with special attention for the past 200 years.

In CLIO-INFRA datasets will be created or improved, for example on living standards, human capital and cultural and political institutions. Economic and social historians from around the world will work together in thematic collaborations, increasing their knowledge of the relevant indicators of economic performance and its causes to collect and share.

The data sets are accessible via a central portal, which creates opportunities for the visualization of data. The long term goal of the project – as developed by the International Economic History Association – to the academic rules as to provide that international cooperation and to facilitate data exchange.

DANS will seek to establish and setup a long-term storage solution by creating a database archive consisting of aggregated data from Excel sheets.

5.6 Digital Collaboratory for Cultural Dendrochronology (DCCD)[22]

Within the dendrochronology field, the exchange of information about tree-ring analysis and samples is of the utmost importance. Dating of wood can only be made possible by exchanging finds and constructing benchmarking timelines based on pre-dated material.

DANS has built an international repository system for dendrochronological material as part of the Digital Collaboratory for Cultural Dendrochronology project. The repository facilitates dendro researchers with a system to safely store their measurements and object descriptions, and exchange this information with their colleagues.

The repository functions as the central hub of the dendro infrastructure, through which all European dendro materials can be uploaded, shared, searched, examined and downloaded. Its backbone is the TRiDaS data format, which is compatible with most of the laboratory software suites and has been defined especially for data exchange.

5.7 Digital Research Infrastructure for the Arts and Humanities (DARIAH)[23]

The mission of DARIAH is to enhance and support digitally enabled research across the humanities and arts. DARIAH aims to develop and maintain an infrastructure in support of ICT-based research practices. It brings together

[22] http://dendro.dans.knaw.nl/about.

[23] http://www.dariah.eu.

researchers, information managers and information providers and it gives them a technical framework that enables enhanced data-sharing among research communities.

DARIAH is working with communities of practice to:

- explore and apply ICT-based methods and tools to enable new research questions to be asked and old questions to be posed in new ways;
- improve research opportunities and outcomes through linking distributed digital source materials of many kinds;
- exchange knowledge, expertise, methodologies and practices across domains and disciplines.

DARIAH has now entered a transitional phase that will end in 2011. DANS was one of the initiators of DARIAH and coordinated the Preparing DARIAH project. DANS also contributed its expertise to technical and dissemination work packages.

In the current transition and future construction phases, DANS will be jointly leading the content element of DARIAH (along with Centre National de la Recherche Scientifique (CNRS), France) and will be contributing expertise in PIDs and some of its research infrastructure technologies to the e-Infrastructure.

5.8 European Social Survey (ESS)[24]

The European Social Survey (ESS) is an international database that has become a prime instrument for innovative comparative research on social and political attitudes. Up until now ESS has collected four biannual rounds of survey data (2002, 2004, 2006, 2008). For each round, the project interviews approximately 50,000 people in Europe (about 1800 in each country) on a broad scope of their social and political attitudes and social backgrounds in a strictly replicated format, which creates rich opportunities for internationally comparative research with a longitudinal perspective. Currently available download statistics reveal over 17,000 users of ESS data in more than 170 countries: users in the Netherlands rank in the top ten. In the first 4 years of their availability (2003–2007), research based on ESS data has yielded more than 400 scientific publications.[25] The data of ESS are disseminated by the Norwegian data archive NSD.[26] The data are available free of charge and without restrictions, for not-for-profit purposes.

ESS may be depicted as a research infrastructure. The infrastructural nature of ESS lies first of all in the rich and high-quality data resource it con-

[24] See http://www.europeansocialsurvey.org. The information in this paragraph is abstracted from the ESS website.

[25] See http://www.europeansocialsurvey.org. ESS Bibliography.

[26] See http://ess.nsd.uib.no.

stitutes for comparative research on social and political attitudes. Secondly, ESS also collects a rich array of social background variables, and thirdly, ESS constitutes an important methodological infrastructure, due to its innovative procedures for question formulation, translation, measurement, sampling and data access.

In the Netherlands, ESS is a project on the National roadmap and is funded by the National funding of the European Strategy Forum on Research Infrastructures (ESFRI).

DANS is responsible within the ESS-NL Task Group together with the consecutive National Coordinators to disseminate the use of the ESS data among both junior and senior researchers in the Netherlands.

5.9 European Holocaust Research Infrastructure (EHRI)[27]

The aim of European Holocaust Research Infrastructure (EHRI) is to "create a sustainable world-class Holocaust Research Infrastructure of European dimensions, which will bring together virtual resources from dispersed archives". Archives containing Holocaust-related materials are fragmented and scattered across the world, therefore making access to resources both complicated and time-consuming.

EHRI was launched in Brussels in November 2010 and will run from October 2010 for four years. This project is financed by the European Union under the 7th Framework Programme for Research and Technology Development. EHRI's main objective is to support the European Holocaust research community by providing an online portal that will give access as open as possible to dispersed sources relating to the Holocaust from all over Europe and Israel, making data available for Holocaust research around Europe and elsewhere into a cohesive body of resources. EHRI will also be encouraging collaborative research through the development of tools.

DANS role in EHRI is to contribute to the Standards and Guidelines[28] for participating archives and for EHRI itself, and to the Data Integration Infrastructure of tools, metadata and multilingual thesauri.

5.10 PersID[29]

The PersID initiative provides PIDs as well as a transparent policy and technical framework for using all kinds of scientific, cultural and other resources in the internet.

[27] http://www.ehri-project.eu.

[28] http://www.ehri-project.eu/partners-organisation.

[29] http://www.persid.org.

Eight national libraries and research institutions in six European countries have worked successfully together in the PersID project (from October 2009 to March 2011). The project was funded by the Dutch SURFShare programme, with a grant from the Knowledge Exchange programme.

The identifier system chosen in the PersID initiative is that of uniform resource names (URNs). The URN system encompasses the traditional bibliographic identifiers such as ISBN and ISSN, but also national bibliographic number (NBN). National libraries administrate the identifiers, which may be assigned to a wide variety of digital objects.

The PersID project partners agree in a letter of intent to use the results from the project for future cooperation, aiming minimally to keep the achieved situation up and running.

DANS is responsible for developing and building the Dutch URN:NBN resolver (http://www.persistent-identifier.nl/). A recent update of the resolver for the German-speaking countries (http://www.nbn-resolving.org/) successfully demonstrated how a URN:NBN identifier entered there was redirected to and resolved by the Dutch resolver.

5.11 Verteld Verleden, spoken testimonies online[30]

The goal of the Verteld Verleden (spoken testimonies) project is to make a start with a distributed approach for shared access to oral history collections and, in relation to that, formulating clear guidelines and best practices for owners of collections in order to get started with new technology.

As a basic principle, the participating institutions offer the metadata in XML format according to the Dublin Core metadata model. The metadata have been made accessible to the harvester of Verteld Verleden via the Open Archives Initiative (OAI) Protocol for Metadata Harvesting.

In the field of technology, existing standards are used within Verteld Verleden and open source components are re-used which have been developed within national and European research programmes (among which Catch, MultimediaN, MultiMATCH). The components include voice recognition, search features, thesaurus audiovisual archives and an interface component for visualizing words from a transcript in a cloud.

The project, which gets its funding from the regulation *Digitalisering met Beleid* (Digitizing with Policy), has a 2-year span. A web portal will be launched in 2012 which initially will make the oral history collections of the project partners accessible for the general audience.

[30] http://www.verteldverleden.org.

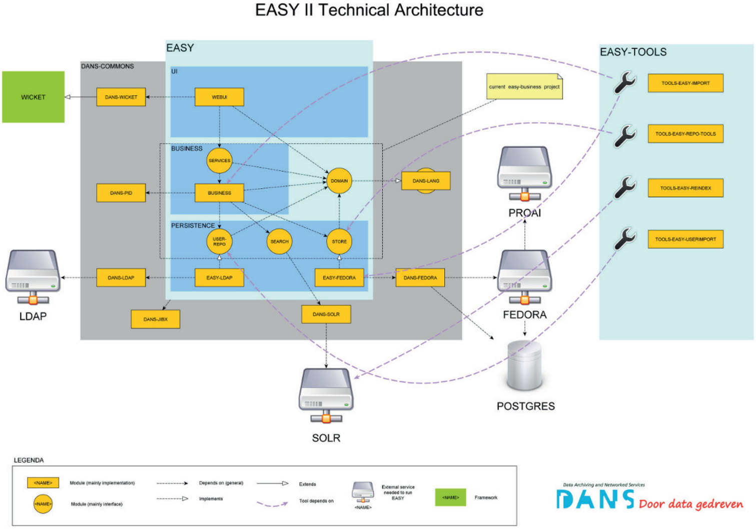


Figure E.1 The EASY system

6 DANS data research infrastructure

The data infrastructure at DANS is centred on its digital repository plus tools and additional support services that help researchers deposit and maintain research data. Self-archiving services targeted to specific communities of researchers, such as archaeologists, are built on top of the of the repository infrastructure.

6.1 EASY[31]

The Electronic Archiving SYstem (EASY) is a pivotal component within the repository infrastructure of DANS (Figure E.1). EASY is based on the repository system Fedora and it has been developed to:

- facilitate self-archiving by researchers;
- enable data curation and management by in-house data experts;
- facilitate publication of data.

Registration is mandatory, but open to anyone interested. When registered, users can archive their data by entering metadata and uploading the accompanying raw data-files. After review by one of the data experts, the data are published on the EASY website from where they can be downloaded.

[31] https://easy.dans.knaw.nl.

Although EASY advocates Open Access, one has to keep in mind that researchers may have reservations with making their data available to just anyone. For this reason, those who are hesitant to release their data into the public domain can either impose an embargo period, during which access is restricted, or take full control over who can download their data by granting access based on individual permission requests.

Metadata are always publicly available, even to those who have not registered as a user in EASY. They are published through the search and browse interface available in the web application, as well as through an OAI interface. Anyone is free to harvest this OAI data and do with the metadata whatever he likes. In order to enable more control of what is actually harvested, service providers can limit their queries to specific disciplines, collections or metadata formats.

Future plans with regard to the dissemination of data include providing (programmatic) access to the data itself, for instance by publishing an API to inspect the data as RDF triples and connecting EASY to Open Linked Data initiatives. This would require a more format/discipline-specific set of content models, that will tell the system how a dataset is actually structured.

A new version of EASY (EASY-II) has been released in September 2011.

6.2 e-Depot Dutch Archaeology (EDNA)[32]

The e-Depot Nederlandse Archeologie (EDNA) pilot project was initiated in September 2004 and ran until February 2006 with funding from SURFfoundation.[33] In 2007, the setting up of EDNA was followed up by the retrospective archiving project EDNA II, which is collaboration between DANS and the Rijksdienst voor het Cultureel Erfgoed.[34] The Dutch archaeology e-depot of digital grey literature and research data is located at DANS, using the EASY self-archiving system. There are currently over 15,000 datasets deposited in EDNA, with some 12,000 being publications only.

The aim of EDNA is to highlight, for Dutch archaeologists, the importance of durable archiving of digital data generated through archaeological research. There is a legal requirement for all archaeological finds and analogue documentation to be deposited with a provincial depot after the completion of the research.

The deposition of data and literature is primarily a process of self-deposition with the depositors adding the metadata themselves. However, there are

[32] http://www.dans.knaw.nl/en/content/categorieen/projecten/edna-e-depot-dutch-archeology.

[33] http://www.surffoundation.nl/en.

[34] http://www.racm.nl.

additional checks made by archivists at DANS. Additional documentation about the methodology used in the research is also deposited.

A PID is assigned to the research project as a whole, but currently no PIDs are being assigned to each individual document.[35]

EDNA has more levels of restriction to access the data archived than is normally part of EASY. These levels of restriction, from the most open to the most restricted are:

- Open Access, but not anonymous as the data user must log on. The registration to use EDNA is minimal and consists of name, password and e-mail address;
- professional archaeologists, including archaeologists working for companies, government agencies and students (for educational use only) and validity of membership is checked;
- personal access, for a researcher first requesting access, which must be agreed and confirmed by the depositor. This provides the depositor with control over their data and who might have access to it. The professional archaeology community is small in the Netherlands and therefore this control over access is easy to manage. It is believed that this control may help to limit access to important information about archaeological monuments to treasure seekers;
- no access (to private data), for example data that have been deposited during an excavation but the literature has yet to be published.

6.3 Persistent identifier services[36]

As organizations and researcher more and more tend to add PIDs to datasets and publications they have produced, the need for services based on these PIDs are growing. The main service, developed for the Netherlands by DANS, is the availability of a so-called resolver (`http://persistent-identifier.nl/`), which can be used to detect the actual URI of a resource with a specific PI. Of course, combining the national resolver on a European level can augment the value of this service. DANS is cooperating with both an Italian and a German meta-resolver.

[35] This level of granularity of PIDs will occur in the imminent migration to the next release of EASY.

[36] `http://www.persistent-identifier.nl`.

6.4 Migration to Intermediate XML for Electronic Data (MIXED)[37]

Migration to Intermediate XML for Electronic Data (MIXED) contributes to digital preservation by dealing with the problem of file formats. Over time, file formats become obsolete. When that happens, the information in such file types is no longer accessible. MIXED follows the strategy of converting files to XML as soon as possible, preferably when data are ingested into an archive, such as EASY. As a service, MIXED can convert files with tabular data (spread sheets and databases) to softwareindependent XML for long-term preservation. The XML can be converted back to the original software dependent format, or to formats of other suppliers of software, or to new formats in the future. This approach solves or alleviates two problems of ordinary migration: (i) it diminishes the need for repeated migrations considerably, because it migrates out of the version sequence of application-bound file formats; and (ii) it facilitates interoperability of data that have been created in different file formats, because they all will be translated into application independent XML.

MIXED consists of a framework plus plug ins. Plug ins take care of the conversions between application file formats and application independent XML formats. At present MIXED can handle these file formats:

– Data Perfect;
– Access 2000 and 2002;
– dBase III and IV;
– Excel 2003.

MIXED is used in the DANS ingest and dissemination workflow, but it is public software. Parts of it are already published in open source repositories and other parts will follow.

6.5 National Academic Research and Collaborations Information System (NARCIS)[38]

National Academic Research and Collaborations Information System (NARCIS) is the national Dutch portal for information about researchers and their work. NARCIS has been a service of DANS as from February 2011. Based on a user survey conducted in 2009,[39] most of its users come from universities and scientific institutions. The number of users is about 1 million per year.

[37] https://sites.google.com/a/datanetworkservice.nl/mixed.

[38] http://www.narcis.nl.

[39] http://depot.knaw.nl/5662/2/What_are_your_information_needs_Elpub_2010.pdf.

As a portal, NARCIS is collecting information from different types of data providers: metadata from repositories, metadata from EASY and descriptions of research institutions, researchers with their expertise and research projects. Besides, NARCIS acts as an access point to scientific news feeds.

Harvesting is the main aspect of the NARCIS system. NARCIS acts as a service provider but can also act as *a data provider to internationally operating services providers like DRIVER and WorldScientific (links!). In fact, you could see NARCIS like a kind of a national aggregator*

One of the most important developments in NARCIS is the implementation of identifiers for objects and researchers. In the Netherlands, digital author identifiers (DAIs) are assigned to researchers (authors). OCLC maintains the central database with DAIs. The Dutch scientific institutions are using a special name space for the DAIs: the eu-repo namespace, to identify information assets used in the European Research Libraries. Information on the eu-repo name space can be found at http://info-uri.info/registry/OAIHandler?verb=GetRecord&metadataPrefix=reg&identifier=info:eu-repo/

Apart from the DAI, NARCIS is also showing PIDs of the objects (publications and datasets) in the repositories of the scientific institutions. Although the institutions are free in choosing the system to add PIDs to objects, the decision has made that at least the URNs will be used (see http://tools.ietf.org/html/rfc3188 and 5.10 PersID).

An elaboration of the existing model is the inclusion of description of enhanced publications in NARCIS. An enhanced publication is a composed object, for instance a publication enhanced with other information like the dataset that has been used in writing this very publication. The metadata of these enhanced publications are described in so-called resource maps, using OAI Object Reuse and Exchange.[40]

6.6 DANS EASY online analysis tool[41]

DANS EASY online analysis tool offers its users additional features in comparison with the standard EASY. The tool is to be used within the social sciences. This service allows the searching, browsing, analysing and downloading of social science data. Major difference with EASY is the possibility to create online tables based on the well-documented datasets. With DANS EASY online analysis tool, data may be analysed online, for instance using regression analysis. This feature is typical for this service that as a matter of facts has been derived from Nesstar[42]

[40] http://www.openarchives.org/ore.

[41] http://194.171.144.69/webview.

[42] http://www.nesstar.com/

DANS EASY online analysis tool is available to a small subset of the social sciences datasets in EASY, namely:

- Cultural Changes in the Netherlands Studies (CV);
- Dutch Parliamentary Election Studies (NKO);
- Social and cultural trends in the Netherlands (SOCON);
- National survey pupils secondary schools (NSO);
- Facilities use survey (AVO);
- Time-budget survey (TBO).

6.7 Netherlands' Geographical Information System (NLGis)[43]

Netherlands' Geographical Information System (NLGis) is a DANS service in which historians can reproduce and visualize regional variation in Dutch historical municipal data, based on data from the last two centuries.

NLGis is a web application that supports the spatial component in historical research. GIS plays an important role in this kind of research. Researchers may upload, display and download the map with historical municipal data.

6.8 DANS data support services and policies, standards and guidelines

DANS promotes the permanent storage and traceability of research data. To this end it provides, among others, practical services to researchers and research groups. Data from numerous resources is made freely available by or through the mediation of DANS. Besides, DANS organizes on request symposia, subsidizes small data projects and carries out ICT activities on behalf of various research projects.

DANS can guarantee permanent access for all standard and preferred formats; in the case of irregular data formats, access is dependent on future developments in software.

Researchers may use consultancy services (data consultancy) with regard to scientific data processing. Merely offering an archiving system is not enough. Therefore, consultancy services have been developed in order to encourage the use of EASY or improve the quality of research data. Examples are:

- data deposit guidebooks for Archaeology, Social Sciences And History;
- help texts in EASY;
- courses and presentations (text material and powerpoints stored on internal DANS website, accessible to DANS employees only);

[43] http://www.dans.knaw.nl/en/content/categorieen/diensten/dutch-geographic-information-system-nlgis.

- communication with depositors.

DANS puts a lot of care and effort into ensuring that data deposits come with good metadata. We discern between three types of metadata: (i) project-specific metadata (Dublin Core); (ii) file-specific metadata; and (iii) metadata on the level of variables (codebooks).

File-specific metadata needs to be sent as a file list together with the dataset. The data deposit guidebook contains a reference table for other variables the depositor can choose to describe the files with. The guidebook texts will be revised in 2011.

With regard to archaeological files, we have agreed on a hybrid division of tasks. We have made arrangements with the archaeological field that archaeological data must be made accessible at DANS.

The guidebooks state that DANS archivists do not change the contents of the file. Archivists only convert files to preferred formats, check the (meta)-data and assign file and dataset rights (publish files). There is no general policy for archivists at DANS on converting files; however, these are common migrations that DANS currently performs:

- word processor documents to PDF/A;
- images to jpeg and tiff;
- vector images to PDF/A and SVG;
- Geographical Information System files to Mid/Mif (MapInfo export format);
- Computer Aided Design (CAD) to DXF version r12 (AutoCAD);
- spreadsheets to CSV (datatables) or PDF/A (reports);
- databases and data tables (dbf) to CSV;
- video to MPEG-4.

6.9 Data seal of approval[44]

Within DANS, a seal of approval for data has been developed to ensure that archived data can still be found, understood and used in the future.[45] In 2008, the first edition of the Data Seal of Approval, written by Laurents Sesink, René van Horik and Henk Harmsen, was presented in an international conference. In spring 2009, the Data Seal of Approval was handed over to an international Board.

[44] http://www.datasealofapproval.org.

[45] DANS – Data Archiving and Networked Services – is an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW), and is also supported by the Netherlands Organisation for Scientific Research (NWO). Since its establishment in 2005, DANS has been providing storage of and continuous access to research data in the social sciences and humanities.

The Data Seal of Approval and its quality guidelines are of interest to research institutions, organizations that archive data and to users of that data. It can be granted to any repository that applies for it via the online assessment procedure.

The criteria for assigning the Data Seal of Approval to data repositories are in accordance with and fit in with national and international guidelines for digital data archiving such as Kriterienkatalog vertrauenswürdige digitale Langzeitarchive, as developed by NESTOR;[46] Digital Repository Audit Method Based on Risk Assessment (DRAMBORA), published by the Digital Curation Centre (DCC) and DigitalPreservationEurope (DPE);[47] and Trustworthy Repositories Audit and Certification (TRAC): Criteria and Checklist of the Research Library Group (RLG).[48] Furthermore the following has been taken into account: Foundations of Modern Language Resource Archives of the Max Planck Institute[49] and Stewardship of Digital Research Data: A Framework of Principles and Guidelines published by the Research Information Network.[50] The guidelines in this document can be seen as a minimum set distilled from the above proposals.

Fundamental to the guidelines are five criteria for digital research data, which together determine whether or not it may be qualified as sustainably archived:

- available on the internet;
- accessible, while taking into account relevant legislation with regard to personal information and intellectual property of the data;
- available in a usable format;
- reliable;
- citable.

The associated guidelines relate to the implementation of these criteria and focus on three stakeholders:

- the data producer is responsible for the quality of the digital research data;
- the data repository is responsible for the quality of storage and availability of the data: data management;
- the data consumer is responsible for the quality of use of the digital research data.

[46] http://edoc.hu-berlin.de/docviews/abstract.php?id=27249.

[47] http://www.digitalpreservationeurope.eu/announcements/drambora.

[48] http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=91.

[49] Peter Wittenburg, Daan Broeder, Wolfgang Klein, Stephen Levinson and Laurent Romary. http://www.lat-mpi.eu/papers/papers-2006/general-archive-paper-v4.pdf.

[50] http://www.rin.ac.uk/data-principles.

More information can be found on the Data Seal of Approval website: www.datasealofapproval.org.

6.10 Repository audit and certification (trustworthy digital repository)

On a European level there many data repositories that can cooperate in a network. But these data repositories may differ in technical level. However, users need common guidelines. Important developments are going on. Apart from the ESA European LTDP Common Guidelines (earth.esa.int/gscb/ltdp/EuropeanLTDPCommonGuidelines_Issue1.1.pdf), a similar approach is proposed in the European Framework for Audit and Certification of Digital Repositories, which was outlined in a Memorandum of Understanding between CCSDS, DANS and DIN.[51] This framework defines three levels of trustworthiness:

- **basic certification:** granted to repositories which obtain Data Seal of Approval (DSA) certification;
- **extended certification:** granted to Basic Certification repositories which in addition perform a structured, externally reviewed and publicly available self-audit based on ISO 16363 or DIN 31644;
- **formal certification:** granted to repositories which in addition to Basic Certification obtain full external audit and certification based on ISO 16363 or equivalent DIN 31644.

The granting of these certificates will allow repositories to show one of three symbols (to be agreed) on their web pages and other documentation, in addition to any other DSA, DIN or ISO certification marks.

6.11 DANS literature publishing infrastructure

DANS participates in the electronic newsletters “Archive Letter” and the quarterly journal “e-Data and Research”. Notable recent acquisitions/published datasets are brought to extra attention in these newsletters. Links to these newsletters and their archives are given on the DANS homepage.

[51] Giaretta, D, Harmsen, H, and Keitel, C. “Memorandum of understanding to create a european framework for audit and certification of digital repositories”. 2010. Available at http://www.datasealofapproval.org/sites/default/files/20100709_020_signedMoUtocreate
aEuropeanFrameworkforAuditandCertificationofDigitalRepositories.pdf.

7 Lifecycles and scholarly primitives in the humanities and social sciences

DANS is involved in a wide variety of research projects in the humanities and social sciences. Therefore, to identify points of commonality in disparate research practices and methodologies where Open Access infrastructures are potentially utilized by researchers to support their practice, we reviewed a number of data and research lifecycles to identify one that could be utilized. The aim was to develop a framework for structured interviews of scholars and identify phases within the lifecycle where data and literature are produced and consumed. These phases may appear obvious until one considers that in some humanities disciplines research literature is the source material for further research questions, and data from Social Science surveys in one discipline can be source data in another.

The Scholarly Communication Life Cycle[52] (Microsoft, 2008) identifies four phases to a research lifecycle, plus two activities common to all. The phases are:

- data collection, research and analysis;
- authoring;
- publication and dissemination;
- storage, archiving and preservation.

Collaboration and discoverability, additional cross-cutting activities, augment all phases of the lifecycle. The addition of collaboration as a feature and need across the lifecycle is not seen in other data or research lifecycles. Although the concept of the "lone humanities scholar" is often quoted,[53][54] the activities of collaboration and communication, throughout the whole of the research process, must be considered central to any Open Access infrastructure.

The British Library also present a similar four-phase research lifecycle, consisting of:

- idea, discovery, design;
- obtain funding;
- experiment, collaborate, analyse;
- disseminate findings.[55]

Although essential, "obtain funding" maybe considered as out-of-scope for the purpose of an Open Access infrastructure for the humanities and social sciences. Furthermore it seems somewhat incongruous that "collaborate" is

[52] http://research.microsoft.com/en-us/about/msr_scholarlycom.pdf.

[53] http://openreflections.files.wordpress.com/2008/10/talk-communia-20102.doc.

[54] www.ahds.ac.uk/e-science/documents/Robinson-report.pdf.

[55] Newbold, 2008,

only in a single phase of the lifecycle and that there is no mention of archiving, preservation or publication.[56]

The DARIAH research lifecycle model (DARIAH, 2010), based primarily upon the previous two examples but also others, provides a simplified combination of both research and data lifecycles (Figure E.2). Search and discovery for research resources is a key feature of this lifecycle as is gathering (or collecting) these resources into an environment where analysis and experimentation can take place (DARIAH, 2010). The addition of "share" at the centre of the lifecycle, in addition to collaborate, implies that data and literature is made available (publicly or to a select group during early research phases) in every phase. This should be considered as an important feature in an Open Access infrastructure. Although *archiving* of a research data and literature should be an integral phase in the lifecycle, occurring as it does after *publishing*, non-permanent *storing* of collected research material and data created occurs at all stages of the lifecycle.

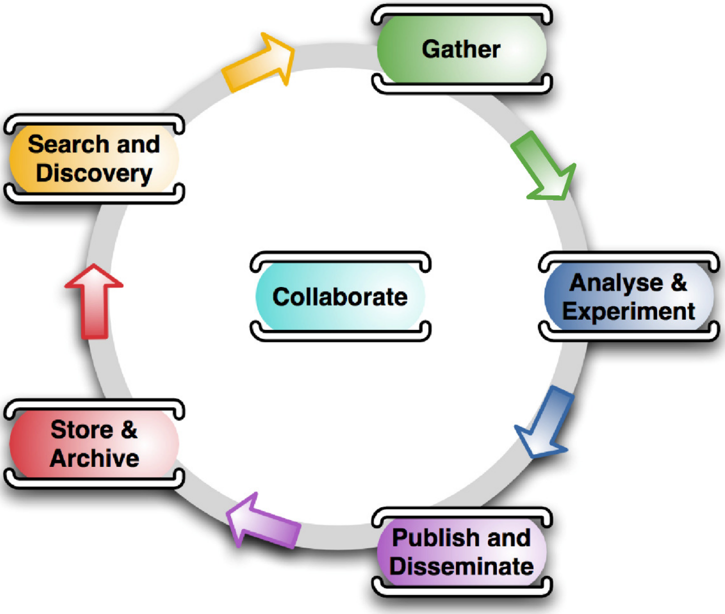


Figure E.2 The DARIAH research and data lifecycle

[56] Publication can be inferred as a form of dissemination.

8 Challenges, opportunities and trends

As an organization promoting storage, curation and access to datasets, DANS will be confronted with some major changes in the next decade.

8.1 E-research

One of the biggest challenges is the rise of e-research: data-intensive research. In this type of research, researchers will rely on the quality of and the access to data. The role of DANS is clear: on the one hand it may start to serve as a central access point to datasets from all kinds of disciplines. On the other hand it plays an important role in the development of regulation of the deposit of data, so that researchers are confident that the data accessed via DANS are valuable. Here one can see a close relationship with data curation.

In e-research, the boundaries between the different disciplines tend to blur. e-Scholars often want to combine data from the humanities and social sciences with data from for instance biology and geography. It certainly will be a major challenge to DANS to realize data curation and data access to (for DANS) non-traditional disciplines (such as the natural sciences) as well. By close cooperation with the Dutch technical universities, the scope may be broadened to the technical disciplines as well.

8.2 DANS as a research organization

DANS could continue to act as a pure service-oriented organization. By doing so, it will take the risk to miss important development in the field of e-research. It would be better to change the scope of the institute in such a way that participating in research will become possible. In other words, for the improvement of the data infrastructure, it will be advantageous to have (e-)scholars working in the institute. Such a group could for instance be involved in research in standardization of (meta-)data and in a kind of trend watching in the data-intensive research field.

8.3 Grid

Apart from this, within DANS a research focus will have to be on giving access to data. Developments in Grid and Cloud computing have hardly been studied with this respect. What is for instance the influence of Cloud computing on costs, trust and quality of data? How will researchers couple data from different sources to each other using the cloud and will these couplings be sustainable?

8.4 Software for data access

Another aspect of data preservation is the fact that more and more datasets cannot be used without special applications that have been developed for these very datasets. What could be the role of DANS in giving access to these data? Would it be necessary to preserve the application software as well or would it be possible to develop an application independent archival system?

8.5 Enhanced publications

An institute as DANS has to be prepared to cope with the developments in the fields of semantic web, "deep access", linked data and RDF. Unlike the examples above, these development will make it possible to combine traditional publications with datasets, audio fragments, software and so on: enhanced publications. Enhanced publications will influence the scholarly communication process in a dramatic way, on the condition that all their components will be Open Access. Otherwise, a frustrating, partially closed, data infrastructure will be developed.

8.6 Scientometrics

Giving access to datasets will also change impact measurements for institutes and researchers. Until now, scientometrics was only based on traditional publications. When re-use of data(sets) will be measured, this will give an additional impact to measure the impact of a specific research.

8.7 Linking of datasets

As already is common in a field like astronomy, linking of datasets available in different data centres (in different countries) will give scholars the opportunity to combine these datasets in, for instance, secondary analysis, data mining and visualization.

9 List of figures

10 List of tables

F Climate Research

Ilse Hamann

1 Climate science

The general public awareness that Earth's climate is an important conditioning factor for the quality of human life has grown because earth science has in the past decades produced and publicised many new and fascinating facts about the dynamics of our complex climate system. Practically everybody experiences, reflects upon, communicates about and depends on the weather and climatic conditions in earning his/her livelihood. Access to often vital information is a prerequisite for strategic planning and taking economically sound decisions.

Climate change challenges the stability of human and environmental systems alike, therefore the interest in a reliable database of global and regional earth system data is great, since such data are needed to find out how resilient these systems are, how climate shifts impact different sectors of environment and society and how to optimally adapt to the increased variability of our climate.

In Germany the National Committee on Global Change Research (NKGCF)[1] has for the past 15 years coordinated German contributions to global change research, with the consideration of the global change-related decisions of the senate commissions of the German Science Foundation (DFG) and of the advisory board of the German Federal Ministry of Education and Research (BMBF). The NKGCF is also the national contact point for the international global environmental change programmes IGBP,[2] WCRP,[3] DI-

[1] http://www.nkgcf.org.

[2] International Geosphere-Biosphere Programme, http://igbp.sv.internetborder.se.

[3] World Climate Research Programme, http://www.wcrp-climate.org.

VERSITAS[4] and IHDP[5] as well as for the Earth System Science Partnership (ESSP).[6] The NKGCF's 15 voting members[7] are working in natural and social science areas with relevance to global change.

Climate science derives quantitative information from observations of key climate variables and from simulations using mathematical, numerical models. Many different types of data are useful for statistical and dynamical approaches in climate research. Observational data from in-situ measurements in the atmosphere, on land and in the oceans form one strong pillar of the database. They are, however, irregularly distributed in space and time, and the global coverage of such measurements of the key variables is insufficient. Remotely sensed data of near-surface properties at Earth's surface augment the in-situ data by highly improved spatial and sometimes temporal sampling density.

On the basis of these data and well-established physical principles, mathematical models of the dynamical behaviour of the atmosphere, the oceans and the cryosphere have been developed and numerical experiments are being carried out with coupled models that involve not only these physical realms, but also biogeochemical processes in and between them. A detailed discussion of the capacity of global climate models to reproduce observed features of the recent climate, investigate past climate changes and produce credible quantitative estimates of the climate development in the future, i.e. possible climatic "futures", is given, for example, by Randall, Wood, Bony et al. (2007) in chapter 8 of Solomon et al. (2007)[8]

The findings described in the following sections of this chapter are to a large extent based on an analysis of documents available to me at the in-

[4] Integrating biodiversity Science for human well-being, http://www.diversitas-international.org.

[5] International Human Dimensions Programme on Global Environmental Change, http://www.ihdp.unu.edu.

[6] http://www.essp.org/index.php?id=10.

[7] http://www.nkgcf.org/committee_members.php?year=2009-2011.

[8] Much of this chapter describes the research infrastructure in climate modelling, on the basis of which the last so-called "IPCC Report" or Fourth Assessment Report (AR4) of the Intergovernmental Panel on Climate Change (IPCC) was written. This Status Report was widely publicised when the IPCC together with Al Gore was awarded the Nobel Peace Prize 2007 "for their efforts to build up and disseminate greater knowledge about man-made climate change, and to lay the foundations for the measures that are needed to counteract such change". The contribution of Working Group I to this Assessment Report, i.e. "Climate Change 2007: The Physical Science Basis", provides extensive descriptions of the complex processes of Earth's climate system (Solomon et al. 2007). The IPCC was established by the World Meteorological Organization (WMO) and the United Nations Environmental Program (UNEP) to assess scientific information on climate change. The IPCC regularly publishes reports that summarise the state of climate science: http://www.ipcc.ch/publications_and_data/publications_and_data_reports.shtml.

stitution where I work, i.e. in the Data Management department (DM)[9] of the German Climate Computing Centre (DKRZ, Deutsches Klimarechenzentrum),[10] or are part of websites of other institutions in the climate research community. Additional information about aspects of the research infrastructure in climate science comes from what I could glean from the literature and by receiving answers to a questionnaire (Spohr, 2010, see Appendix 1) from senior scientists working at six different institutes in teams conducting research or providing service(s) in this field. In the next sections, frequent reference is made to these information sources, and many of the diagrams and figures shown have been prepared by my colleagues and by the principal investigators of large research projects for a variety of purposes. The figures include depictions of workflows, organisational diagrams highlighting collaborative aspects and easy-to-grasp displays of pathways of information, data and scientific results within the research cycle.

When planning the scope of my work for this study, i.e. writing the climate science subject-specific chapter, I felt that I needed to strike a balance between a broad survey of climate research programmes and institutions worldwide and a highly selective case study of the research infrastructure at DM/DKRZ and at the World Data Center for Climate (WDCC,[11] maintained by DM/DKRZ), for which I would be able to provide much more detail (WDCC is partner in this Task 7.1 in OpenAIRE). I opted for an eclectic analysis of what I think are important entities in Germany operating at different locations/stages of the climate research cycles and in the climate data web. Job titles of climate scientists working in these organisations and institutions have many facets as there are linkages of climate science with many other disciplines.

First, in section 2, descriptions of the workflows adopted for the presently worldwide largest international climate change research project illustrate relevant climate modelling research infrastructure. In section 3, an overview is given of the major data centres that curate data from observational climate science programmes in order to describe the infrastructure that is in place in Germany for this second large branch of climate science. Section 4 contains the results of a survey of selected climate researchers with differing scientific foci. The survey intended to determine current practices (workflows) at institutions in Germany regarding climate research, climate data management and the kinds of services provided on behalf of the climate science community for a variety of end users, as well as Open Access policies and incentives with respect to literature and data in the respective organisation.

[9] http://www.dkrz.de/daten-en.

[10] DKRZ is a non-profit and non-commercial limited company with four shareholders.

[11] http://www.dkrz.de/daten-en/wdcc?set_language=en.

Open Access aspects, practices and policies with respect to literature and data, whether already implemented or being presently developed in the climate research groups and institutions which I have surveyed, are described in sections 5 and 6, respectively. At the chapter's end, challenges for climate science research infrastructures are mentioned, and an outlook towards upcoming developments is given. Finally, some pointers are given that relate to the Open Access infrastructure in the field of climate science.

2 Current status of climate modelling research infrastructure

2.1 Data management and WDCC at the German Climate Computing Centre (DM/DKRZ and WDCC/DKRZ)

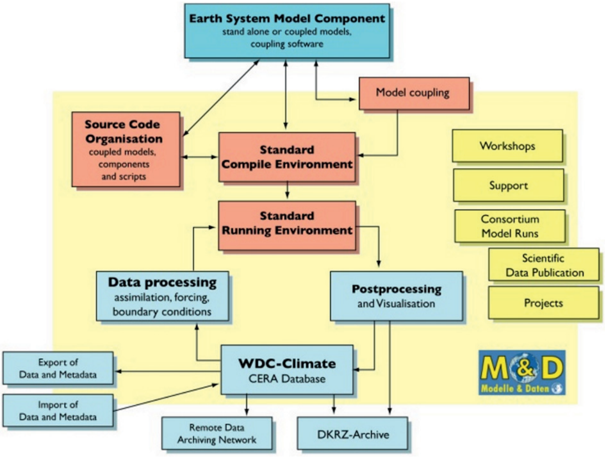

Earth System Model Component
stand alone or coupled models, coupling software
Model coupling
Source Code Organisation
coupled models, components and scripts
Standard Compile Environment
Standard Running Environment
Workshops
Support
Consortium Model Runs
Scientific Data Publication
Projects
Data processing
assimilation, forcing, boundary conditions
Postprocessing and Visualisation
Export of Data and Metadata
Import of Data and Metadata
WDC-Climate
CERA Database
Remote Data Archiving Network
DKRZ-Archive
M&D
Modelle & Daten

Figure F.1 M&D services (the diagram was taken from the static mirror of the original M&D website)[12]

Since the time when DM/DKRZ was still called "Model & Data" (M&D, during the decade 2000–2009[13]) and when it was part of the Max Planck Institute for Meteorology in Hamburg, it has availed itself of the highly developed research infrastructure characteristic for the Max Planck Society, possibly the most renowned research organisation in Germany.[14] The scien-

[12] http://www.mad.zmaw.de/service-support/index.html.

[13] http://www.mad.zmaw.de.

[14] http://www.mpg.de/183251/portrait.

tists at the Max Planck Institute for Meteorology in Hamburg (founded in 1975) "have been studying how physical, chemical and biological processes and human behaviour contribute to global and regional climate changes. The scientists develop numerical models and measurement methods to explain the natural variability of the atmosphere, the oceans and the biosphere and to assess the influence of land use changes, industrial development, urbanisation and other human influences. Together with the Max Planck Institutes for Biogeochemistry (Jena) and for Chemistry (Mainz), they strive to provide a better understanding of the chemical and biological factors that determine the concentrations of greenhouse and other trace gases in the atmosphere and how they interact with the terrestrial and marine biospheres".[15]

The scientists who worked in the M&D group and now belong to DM/DKRZ provide central support for the German and European climate research community, with an emphasis on development and implementation of best practice methods for Earth System modelling and related data management. Figure F.1 summarises the support services that M&D offered.

The DKRZ is a national service provider, providing high performance computing platforms, sophisticated and high capacity data management, and superior service for premium climate science.[16] DKRZ operates a fully scalable supercomputing system designed for and dedicated to earth system modelling.

Today the DKRZ supports the whole data life cycle of climate (model) data,[17] i.e. data creation, diagnostics, visualisation, archiving and dissemination to scientists all over the world. Therefore the day-to-day work of earth system scientists and the long-term archiving of scientific results is supported within a virtual research environment (see Figure F.2).

While this data cycle management corresponds in principle to the mission of the WDCC/DKRZ,[18] the WDCC restricts itself to curating mostly *data products* from climate modelling, since storage of *raw data* from satellites or climate models, for example, on a global basis is beyond the scope of the available facilities. In section 3, thematically corresponding data centres like Earth observation, meteorology, oceanography, paleoclimate and environment are described. WDCC is enhancing its cooperation with these centres with the aim to establish a complete network for climate data.

The WDCC's development during the past two decades:

- Scientific data management at DKRZ started in the 1990s, with the aim of collecting, scrutinising and disseminating data related to climate change on all time scales

[15] http://www.mpg.de/155345/meteorologie?section=all.

[16] http://www.dkrz.de/about-en/aufgaben.

[17] http://www.dkrz.de/daten-en.

[18] http://www.dkrz.de/daten-en/wdcc.

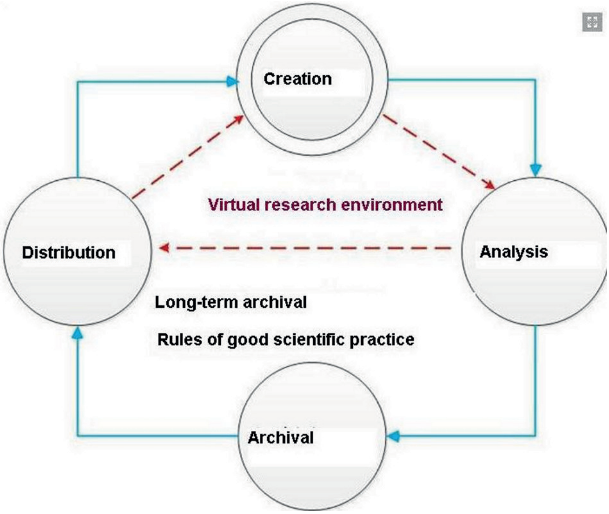

Creation
Virtual research environment
Distribution
Analysis
Long-term archival
Rules of good scientific practice
Archival

Figure F.2 Data life cycle and virtual research environment

- In 2001 DKRZ began the development of the Climate and Environmental Retrieval and Archiving (CERA-2) data model, which was a collaboration of M&D, the Alfred Wegener Institute for Polar and Marine Research (AWI)[19] and the Potsdam Institute for Climate Impact Research (PIK)[20]
- In 2003 the CERA database was accepted by the WDC for Climate (WDCC) of the International Council of Scientific Unions (ICSU, today called International Council for Science).[21] In August 2011 the WDCC was re-evaluated and accepted as a member into the new International Council for Science World Data System (WDS)
- In 2004/2005 the "consortium experiments" were established, i.e. "capacity computing" necessary for highly relevant national or international research.
- Since 2010 the WDCC has been integrated in the Data Management department of DKRZ.

The geographical distribution of WDCC users by continent and by country is shown on the WDCC "Statistics" web page[22] and in Figures F.3 and F.4.

[19] http://www.awi.de/en/institute.

[20] http://www.pik-potsdam.de.

[21] http://www.icsu.org.

[22] http://www.dkrz.de/daten-en/wdcc/statistics/wdcc-statistics-2010.

[23] Country Group 1 (1 user):

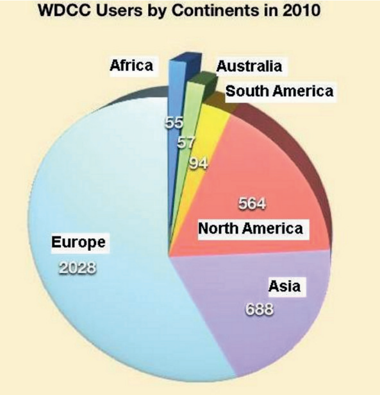


Figure F.3 Number of WDCC users by continent in 2010

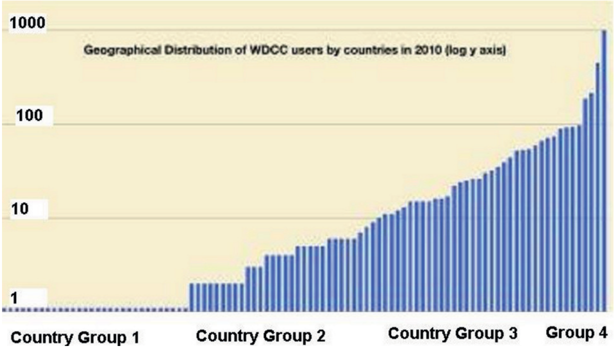


Figure F.4 Geographical distribution of WDCC users by country groups[23]in 2010 (logarithmic y-axis)

Tansania, Solomon Islands, Georgia, Jordan, Uganda, Cyprus, Venezuela, Madagascar, British Indian Ocean Territory, Latvia, Mali, Lesotho, Saudi Arabia, Serbia and Montenegro, Macao, Lebanon, Myanmar, Oman, Cuba, Estonia, Slovenia, Iceland, Armenia, Croatia, Luxembourg, South Georgia, Trinidad and Tobago, Paraguay, North Korea, Congo.
Country Group 2 (2–10 users):
Cameroon, Sudan, Senegal, Bulgaria, Zambia, Nicaragua, Ukraine, Mongolia, Costa Rica, Bangladesh, Bolivia, Jamaica, Nepal, Niger, Lithuania, Colombia, Hungary, Romania, Philippines, Hong Kong, Nigeria, Ethiopia, Egypt, Pakistan, Kenya, Singapore, Israel, Viet Nam, Czech Republic, Chile.
Country Group 3 (11–100 users):
Malaysia, Peru, Argentina, New Zealand, Mexico, Poland, South Africa, Portugal, Finland, Ireland, Turkey, Denmark, Norway, Greece, Russian Federation, Indonesia,

2.2 The Climate and Environmental Retrieval and Archive (CERA) database

The georeferenced data held at WDCC, i.e. climate model results from global and regional climate model experiments performed at DKRZ, are available from the WDCC via the operational CERA data and information system. Input is accepted in electronic form. This includes present-day climate, paleoclimate simulations and IPCC scenario[24] runs for the future. The WDCC archives and disseminates more than 341 TB climate model data and related observations. In Figure F.5, the increase of the volume of data content since the first version of CERA is shown.

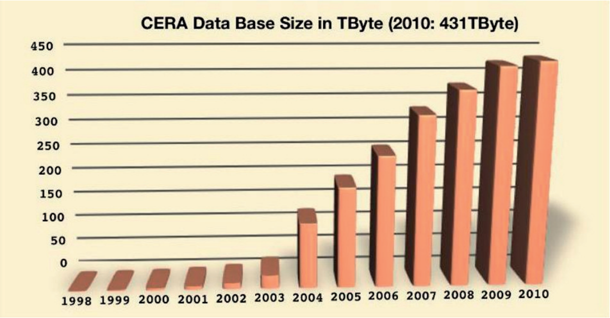


Figure F.5 Development of the CERA database size (terabytes) since 1998[25]

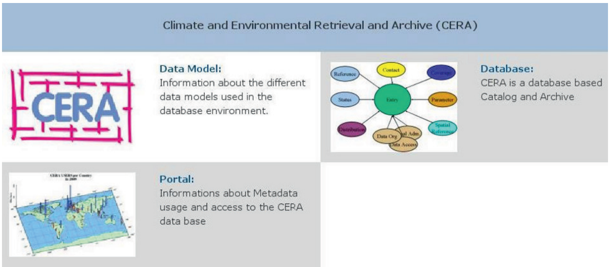
The graphical user interface of the CERA Portal guides the user to search functions for the data sets in the database, and to utilities and tools for data download, processing and visualisation[26] (Figure F.6). More information on the data models used in CERA, the modules of the catalogue, details about the technological implementation and metadata usage and access to the CERA database can be found at http://www.dkrz.de/daten-en/cera.

Belgium, Thailand, Taiwan, Austria, Sweden, Australia, Netherlands, Iran, Brazil, Switzerland, South Korea, Spain, Italy, Japan, Canada, France, India.
Country Group 4 (more than 100 to over 1000 users):
Great Britain, China, USA, Germany.

[24] http://www.ipcc.ch/pdf/special-reports/spm/sres-en.pdf.

[25] http://www.dkrz.de/daten-en/wdcc/statistics/wdcc-statistics-2010.

[26] http://cera-www.dkrz.de/WDCC/ui/Index.jsp.



Climate and Environmental Retrieval and Archive (CERA)

Data Model: Information about the different data models used in the database environment.

Database: CERA is a database based Catalog and Archive

Portal: Informations about Metadata usage and access to the CERA data base

Figure F.6 Components of the CERA archive: data model, database and data portal

2.3 The Coupled Model Intercomparison Projects (CMIP)

M&D was and DM/DKRZ is now a vital component of the Coupled Model Comparison Projects CMIP3[27] and CMIP5,[28] respectively. In these projects, particularly large data volumes were/are generated, which require storage and archiving and need to be distributed among and maintained for (re)-use worldwide. WDCC holds data from model simulations for the Fourth Assessment Report (AR4, CMIP3),[29] and in CMIP5, WDCC/DKRZ is one of the four gateways to which a subset of the data that project participants are delivering to the Program for Climate Model Diagnosis and Intercomparison at the Lawrence Livermore National Laboratory in California (PCMDI, USA) is replicated. The data management for CMIP5 will be shared between the British Atmospheric Data Centre (BADC, UK), the PCMDI and WDCC/DKRZ.

2.3.1 The Coupled Model Intercomparison Project, phase 3 (CMIP3)

During the years 2005 and 2006, climate model output from simulations of the past, present and future climate was collected by the PCMDI. The joint Working Group on Coupled Modelling (WGCM) of the World Climate Research Programme (WCRP) and of Climate Variability and Predictability (CLIVAR), which are both projects of the World Meteorological Organization (WMO), organised this activity to enable researchers "outside the major modelling centres to perform research of relevance to climate scientists preparing the AR4 of IPCC".[30] This so-called "WCRP CMIP3 multimodel data set" was to serve IPCC's Working Group 1, which focuses on the physi-

[27] The Coupled Model Intercomparison Project, phase 3.

[28] The Coupled Model Intercomparison Project, phase 5.

[29] http://www.dkrz.de/daten-en/wdcc/projects_cooperations/ipcc-data-1.

[30] http://cmip-pcmdi.llnl.gov/cmip3_overview.html?submenuheader=1.

cal climate system (atmosphere, land surface, ocean and sea ice). The size of the data set amounts to 36 TB in 83,000 files. By the end of 30 January 2009, 2570 users had downloaded 536 TB in 1,781,000 files[31] More comprehensive sets of output for a specific model may be available from the modelling centre that produced it.

With the consent of participating climate modelling groups, the WGCM has declared the CMIP3 multimodel data set open and free for non-commercial purposes. After registering and agreeing to the “terms of use”,[32] anyone can obtain model output[33]

2.3.2 The Coupled Model Intercomparison Project, phase 5 (CMIP5)

After publication of the AR4 in 2007, the WGCM had agreed on a new set of coordinated climate model experiments in phase 5 of the CMIP. In a Memorandum of Understanding (MoU) the PCMDI, BADC and WDCC outlined how they would deliver a model intercomparison archive for CMIP5 and the IPCC Data Distribution Centre (DDC)[34] (this MoU was signed in December 2008). Subsequently a description of the design of the experiments to be done in CMIP5 was given by Taylor, Stouffer and Meehl, in 2009[35] (updated in Taylor et al., 2011). The results of CMIP5 are expected to be useful not only to the IPCC’s Working Group 1, but also to those considering possible consequences of climate change, e.g. the IPCC Working Groups 2 “Impacts, Adaptation, and Vulnerability”[36,37] (Parry, Canziani, Palutikof, van der Linden and Hanson, 2007) and 3 “Mitigation of Climate Change”[38,39] (Metz, Davidson, Bosch, Dave and Meyer, 2007).

According to the CMIP5 web page[40] (which includes a schedule) the objectives of the planned standard set of model simulations of CMIP5 are:

- to evaluate how realistic the models are in simulating the recent past,
- to provide projections of future climate change on two time scales (near term to about 2035 and long term to 2100 and beyond),

[31] http://www-pcmdi.llnl.gov/ipcc/usage_statistics.php.

[32] http://www-pcmdi.llnl.gov/ipcc/info_for_analysts.php.

[33] https://esg.llnl.gov:8443/index.jsp.

[34] http://home.badc.rl.ac.uk/lawrence/static/2008/12/03/cmip5_archive_mou_final.pdf.

[35] http://cmip-pcmdi.llnl.gov/cmip5/docs/Taylor_CMIP5_design.pdf.

[36] http://www.ipcc-wg2.gov/AR4/website/fi.pdf.

[37] http://www.ipcc-wg2.gov.

[38] http://www.ipcc.ch/publications_and_data/ar4/wg3/en/contents.html.

[39] http://www.ipcc-wg3.de/publications/assessment-reports/ar4.

[40] http://cmip-pcmdi.llnl.gov/cmip5/index.html.

- to reach a better understanding of some of the factors responsible for differences in model projections, including quantifying some key feedbacks such as those involving clouds and the carbon cycle

There are 21 modelling groups in 12 countries participating in CMIP5 (Table F.1).

In CMIP5, data archiving is again – as it was in CMIP3 – managed by the PCMDI, which also collects the multimodel output data and is responsible for authorisation and authentication. PCMDI and two more data centres constitute the Earth System Grid Federation[41] (ESGF), i.e. the BADC which organises the description of the data (metadata) and the replication of the data sets, and the WDCC which is responsible for the development of some quality control tools and data publication. Information on data access and availability is updated on a daily basis.[42] A check of the CMIP5 archive status page[43] shows that nine of the 21 modelling groups have delivered data so far (groups 1, 2, 4, 6, 12, 17, 18, 19 and 20 in Table F.1)

Compared with CMIP3, CMIP5 model documentation will be made more comprehensive and accessible by a standardised vocabulary for describing models and model simulations, i.e. the data reference syntax (DRS). Model metadata[44] will include global attributes (information about the experiment and the model which originate the data), variable attributes (names and units of output variables) and coordinate variables (bounds of the model region, grid axes; Taylor and Doutriaux, 2010).[45] Furthermore, an interactive web-based questionnaire that makes it easier for modelling groups to provide the model and simulation documentation, is being developed by the (mostly European) consortium of project METAFOR[46] (Common Metadata for Climate Modelling Digital Repositories). WDCC is a project partner in METAFOR, charged with the development of common information model (CIM) creation tools (workpackage 6 of METAFOR; Toussaint and Lautenschlager, 2008).[47] The primarily US Earth System Curator[48] team is providing tools for ingesting the information in the questionnaire, designing web-based discovery tools for interrogating the documentation and integrating these tools into

[41] http://esg-pcmdi.llnl.gov/esgf.

[42] http://cmip-pcmdi.llnl.gov/cmip5/availability.html?submenuheader=3.

[43] http://esgf.org/wiki/Cmip5Status/ArchiveView, 16 August 2011.

[44] The metadata shall be consistent with the CF (Climate and Forecast Metadata) Convention, while data files will be accepted in the Network Common Data Form (NetCDF).

[45] http://cmip-pcmdi.llnl.gov/cmip5/docs/CMIP5_output_metadata_requirements.pdf.

[46] http://metaforclimate.eu.

[47] http://colab.mpdl.mpg.de/mediawiki/images/2/20/ESci08_Sem_3_CERA-2_Toussaint.pdf.

[48] http://www.earthsystemcurator.org.

the ESG framework, whereby this meta-information is put in a searchable database linked to the model output.[49]

The Thematic Real-time Environmental Distributed Data Services[50] (THREDDS) data server is the central distribution unit at WDCC to deliver or publish CMIP5 data to those data centres that are members of the Earth System Grid Federation.[51]

Table F.1 List of groups participating in CMIP5 and terms of use of model output data[52]

No.	Primary group/acronym	Full name	Country	ToU*
1	NCC	Norwegian Climate Center	Norway	ns
2	MOHC	Met Office Hadley Centre	UK	a, ns
3	GFDL	Geophysical Fluid Dynamics Laboratory	USA	a
4	IPSL & LMD	Institut Pierre-Simon Laplace	France	a
5	NIES & U Tokyo	National Institute for Environmental Studies	Japan	nc
6	CCCMA	Canadian Centre for Climate Modelling and Analysis	Canada	nc
7	CSIRO & BMRC	Commonwealth Sci. and Industrial Research Org/Bureau of Meteorology Res. Centre	Australia	nc
8	MPI	Max Planck Institute for Meteorology	Germany	a
9	INGV, CEMCC	Istituto Nazionale di Geofisica e Vulcanologia	Italy	nc
10	EC-Earth Consortium	EC-Earth model based on European Centre for Medium-Range Weather Forecasting's seasonal forecasting system	Europe	nc
11	NASA GSFC	NASA Goddard Space Flight Center	USA	ns
12	CSIRO & QCCCE	CSIRO/Queensland Climate Change Centre of Excellence	Australia	ns
13	NCAR	National Center for Atmospheric Research	USA	ns
14	MRI	Meteorological Research Institute	Japan	nc
15	METRI (with MOHC)	National Institute of Meteorological Research	Korea	ns

[49] http://www.earthsystemcurator.org/projects/end2end.shtml.

[50] THREDDS is middleware facilitating the supply of data from provider to users, http://www.unidata.ucar.edu/publications/factsheets/2007sheets/threddsFactSheet-1.doc.

[51] http://esgf.org/wiki/ESGF%20Members.

[52] https://is.enes.org/documents/Taylor_CMIP5_update_pub.pdf.

16	LASG IAP	LASG, Institute of Atmospheric Physics (IAP), Chinese Academy of Sciences (CAS)	China	ns
17	NASA GISS	NASA Goddard Institute for Space Studies	USA	ns
18	BCC	(National) Beijing Climate Center, China Meteorological Administration	China	ns
19	INM	Institute for Numerical Mathematics	Russia	a
20	CERFACS & CNRM	Centre Europeen de Recherche et Formation Avancees en Calcul Scientifique	France	ns
21	U. Reading	University of Reading	UK	ns

* ToU: Terms of use of output data: nc: non-commercial; a: unrestricted; ns: not specified

2.4 The Earth System Grid data infrastructure

A lesson from CMIP3 has been that there lies great value in archiving multi-model output in a structured and uniform way. The user community expects to be able to extract data efficiently and in a uniform way across all models. The modelling centres that are contributing the data are responsible for writing the data in that desired way.[53] For the extensive list of model output that is requested in CMIP5 specifications for writing this output are provided. In addition, a software library, the so-called Climate Model Output Rewriter (CMOR2)[54] has been written to facilitate writing model output that conforms to these requirements. CMOR2 fills the gap that ISO metadata standards leave, i.e. information about different modelling grids or the rotation of the earth's pole with respect to the model grid, for example.

Other software developed and distributed by PCMDI includes the ESG data node software for archiving and publishing[55] output and the ESG gateway software to deliver data to end users and provide portal services like registration, security, search and discovery, subsetting and server-side calculations, automated capability to inform users of database withdrawals/additions and use statistics (e.g. number of downloads categorized by model/expt/variable). For data management and analysis for the Earth System Grid, see Williams et al. (2008).

CMIP5 model output will be served by federated centres around the world using different storage architectures, which are as far as possible hidden from the user (Figure F.7).

[53] http://cmip-pcmdi.llnl.gov/cmip5/output_req.html.

[54] http://www2-pcmdi.llnl.gov/cmor/documentation.

[55] Here "publishing" means "provision of data sets to users".

[56] http://cmip-pcmdi.llnl.gov/cmip5/submit.html?submenuheader=2.

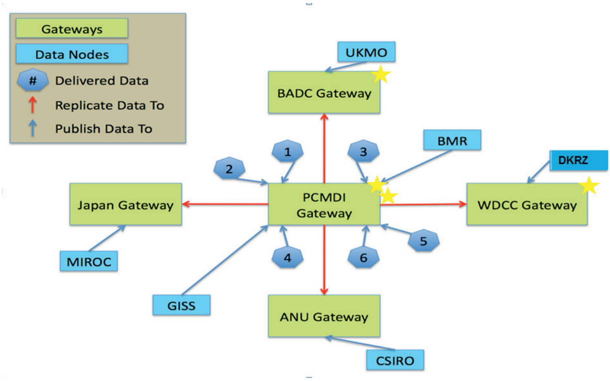


Figure F.7 The Earth System Grid data infrastructure (slightly modified from source)[56]

The CMIP5 archive will be distributed among several centres and will appear to be a single archive. The blue heptagons in Figure F.7 stand for several modelling groups/data nodes publishing data to the PCMDI, while the blue data nodes are publishing data to the respective green gateways. The data centres with the yellow stars will curate the complete CMIP5 data archive, and the second yellow star at PCMDI denotes that here access control is being exercised.

In an ESG Federation wiki on the CMIP5 status page, the operational data nodes and gateways are listed.[57] The size of the CMIP5 archive will be approximately 2 Petabytes published and 1 Petabyte replicated: see, for example, the WDCC gateway realised via the gateway portal software.[58]

The PCMDI also developed the climate data analysis tools (CDAT), whose utility grew from a model behaviour assessment tool to an open-source environment of software which facilitates the analysis of very large data volumes generated by model intercomparison projects and observational programmes that are widely dispersed among many international institutions[59] (Williams, Doutriaux, Drach and McCoy, 2009).

[57] http://esgf.org/wiki/Cmip5Status.

[58] http://ipcc-ar5.dkrz.de/home.htm;jsessionid=47933B7BA7DC201C98803F5D20FF57F2.

[59] http://www.ametsoc.org/meet/annual/annual90shortcourses/1.30pm%20Doutriaux%20II.pdf.

In his IS-ENES[60] Barcelona presentation, Karl Taylor outlined the path of CMIP5 model output through the data infrastructure as follows:[61]

- model output is produced and sent to PCMDI for quality control (QC) checks,
- model output produced and checked for compliance with some output requirements (by CMOR or CMOR-checker),[62]
- METAFOR questionnaire[63] completed generating model and simulation documentation,
- data are made available via the Earth System Grid Federation,
- digital object identifiers (DOIs) are assigned to model output for reference by published literature,[64]
- data are served by ESG gateways via web interfaces.[65]

At DM/DKRZ, the MPI data node (compare with Figure F.7) is being structured as shown in Figure F.8.

2.5 Quality control of CMIP5 model output data[66]

The ESGF partners at PCMDI, BADC and DKRZ which are hosting an ESG gateway and who are producing data replicates of subsets of CMIP5-Data for the AR5[67] carry out distributed quality control. QC occurs at different levels:

- Level 1: CMOR2 and ESG publisher conformance checks are performed at all ESGF partners during ESG publication. The QC1 metadata checks are testing for completeness and execute the technical validation of the questionnaire input.
- Level 2 is performed on requested subsets of CMIP5 data at all ESGF partners. With respect to data, QC2 involves consistency checks, i.e.

[60] "Infrastructure for the European Network for Earth System Modelling" is an FP7-Project funded by the European Commission under the Capacities Programme, Integrating Activities. The project has started on the 1st March 2009 and will finish on the 28th February 2013. IS-ENES promotes the development of a common distributed modelling research infrastructure in Europe in order to facilitate the development and exploitation of climate models and better fulfill the societal needs with regards to climate change issues.

[61] https://is.enes.org/documents/Taylor_CMIP5_update_pub.pdf.

[62] http://cmip-pcmdi.llnl.gov/cmip5/output_req.html?submenuheader=3#cmor.

[63] http://q.cmip5.ceda.ac.uk.

[64] Note that this "publishing" of scientific primary data has a different connotation than that described in section 2.1.5.

[65] ESG data gateways are located at BADC, DKRZ, NASA JPL, NCAR, NCI, NERSC, PCMDI, and ORNL (full names of these institutions can be found in the Acronym section at the end of this chapter). http://www.earthsystemgrid.org/about/overview.htm.

[66] Description of quality control arrangements courtesy of Martina Stockhause, DM/DKRZ.

[67] Fifth Assessment Report of the IPCC.

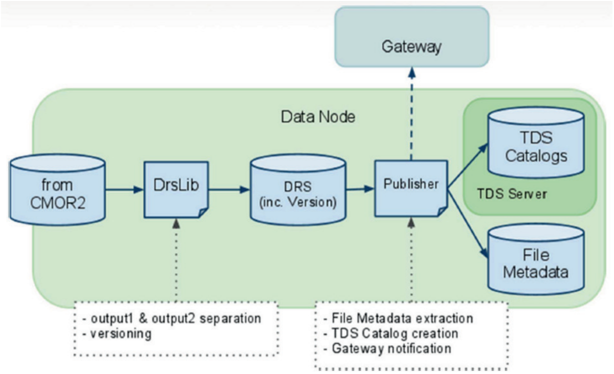


Figure F.8 Processing chain for model output data (courtesy of Estanislao Gonzalez, DM/DKRZ)

check of statistical global values and additional DRS checks (software developed at WDCC),[68] i.e. where exactly in the directory structure the data file is to be found (Taylor et al., 2011).[69] With respect to metadata a subjective QC2 is carried out by the scientist producing the model output data.

- Level 3: technical quality assurance implies double and cross-checking of data and metadata (approval needed from the author); QC Level 3 scientific quality assurance is a check of data and metadata (and approval) by the author.

2.6 Long-term archiving at WDCC

The numbers in circles in Figure F.9 denote steps 1–8 of the following sequence of actions:

1. information gathering and consulting with regards to a request for long-term archiving of data set(s),
2. project specification and cost estimate,
3. defining and including metadata into the WDCC archive,
4. integration and filling of data into the WDCC database,
5. quality assurance,
6. assignment of a DOI (optional),
7. completion of archiving assignment, activation of access permission,
8. maintenance of data archive, possibly adaptation of access right.

[68] QC Level 2 tool developed by Heinz-Dieter Hollweg, DM/DKRZ.

[69] http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_syntax.pdf.

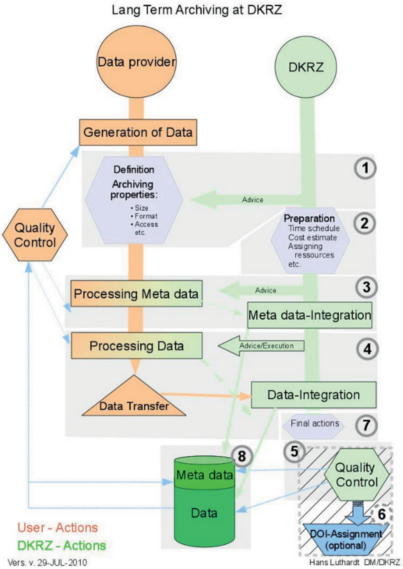


Figure F.9 Work flow of long-term archiving at WDCC (courtesy of Hans Luthardt, DM/DKRZ)

2.7 Publication and citation of scientific primary data

Weather and climate researchers need comprehensive and detailed data in order to arrive at reliable findings. The project Publication and Citation of Scientific Primary Data – Scientific and Technical Data (STD-DOI), for the development of a standard to publish and secure the valuable data for the long term, was supported by the DFG. The aim was to make primary scientific data citeable as publications. In this system, a data set is attributed to its investigators as authors like it is done for a work in the conventional scientific literature. Scientific primary data should therefore not exclusively be understood as part of a scientific publication, but may have its own identity. Since completion of the STD-DOI project, a production service for DOIs has been established. Figure F.10 gives a schematic overview of the network of data publication agents in Germany, i.e. data centres for scientific and technical data in the earth and environmental sciences, the registration agency Technische Informationsbibliothek (TIB Hannover),[70] i.e. the German National

[70] http://www.tib-hannover.de/en/the-tib/doi-registration-agency.

Library of Science and Technology, and the International DOI Foundation (IDF).[71]

Of the archives shown in Figure F.10, the World Data Center for Marine Environmental Sciences (WDC-MARE) was the first one to register data sets as part of a bibliographic citation using the DOI (see also section 3.3). Using the system, scientists worldwide gain access to a web-based platform that enables them to enter and find the data. The system ensures high data quality and long-term use with persistent identifiers (DOI/uniform resource name) for tomorrow's research. With the proposed publication process a method is given by which credit can properly be assigned for the data producers related to a defined citation.

In a new DFG-funded project, KomFor,[72] the TIB cooperates with the four data centres shown in Figure F.10 to establish a competence centre, i.e. the Centre of Expertise for Research Data from the Earth and Environment. In another DFG-funded project, Wikidora,[73] DKRZ and two partners[74] are preparing meteorological research data for persistent identifier registration, which is realised via the web-based workflow application Atarrabi. This new workflow system will be made available as open source software to scientists worldwide[75] (see also Hense, Hense and Lautenschlager, 2010).

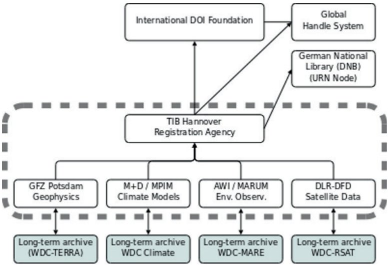

International DOI Foundation
Global Handle System
German National Library (DNB) (URN Node)
TIB Hannover Registration Agency
GFZ Potsdam Geophysics
M+D / MPIM Climate Models
AWI / MARUM Env. Observ.
DLR-DFD Satellite Data
Long-term archive (WDC-TERRA)
Long-term archive WDC Climate
Long-term archive WDC-MARE
Long-term archive WDC-RSAT

Figure F.10 Network of publication agents in Germany for scientific and technical data in the earth sciences, the registration agency and the International DOI Foundation (source: STD-DOI project homepage)[76]

Figure F.11 shows the process followed when the WDCC, as a publication agent, has been asked to publish environmental data. After mutual agreement for publication has been reached between the scientist and WDCC, the

[71] http://www.doi.org.

[72] http://www.tib-hannover.de/en/the-tib/projects/komfor.

[73] http://umwelt.wikidora.com/wikidora.

[74] Bonn-Rhein-Sieg University of Applied Sciences (Department of Computer Science) and the Meteorological Institute of Bonn University.

[75] http://sourceforge.net/projects/atarrabi.

[76] http://www.std-doi.de.

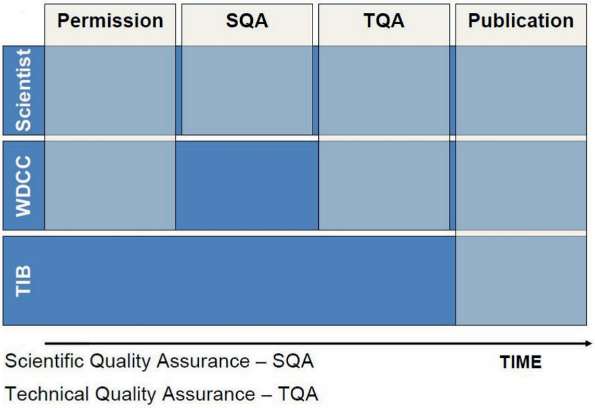

Permission
SQA
TQA
Publication
Scientist
WDCC
TIB
TIME
Scientific Quality Assurance – SQA
Technical Quality Assurance – TQA

Figure F.11 Responsibilities (upper x-axis) of scientist, WDCC and TIB (y-axis) in the data publishing process (lower time x-axis), indicated as shaded cells in this tabular diagram (courtesy of Heinke Hoeck, DM/DKRZ; Hoeck, 2010)

former takes on the scientific quality assurance (SQA), whereupon WDCC gets involved with technical quality assurance (TQA). This includes checks whether:

- number of data sets is correct and not equal 0,
- size of every data set is not equal 0,
- data sets and corresponding metadata are all accessible via the internet,
- data size is controlled and correct,
- time description (metadata) and existence of data are consistent, complete, start/ stop date consistent, continuous time steps are correct,
- format is correct,
- variable description and data are consistent.

After TQA follows the quality control of the descriptive metadata set by the author and WDCC (Figure F.12), after which the registration agency TIB assigns a DOI (Figure F.13). The DOI consists of two parts. The prefix is assigned by the registration agency. The suffix is provided by the data centre, i.e. the agency which is responsible for the contents. Resolution occurs via the resolver[77] or directly.[78]

TIB was the first DOI registration agency for primary data worldwide (since 2005) and is one of the founding members of DataCite[79] (a not-for-

[77] `http://dx.doi.org`.

[78] `http://dx.doi.org/10.1594/WDCC/CCSRNIES_SRES_B2`.

[79] `http://www.datacite.org/whatisdatacite`.

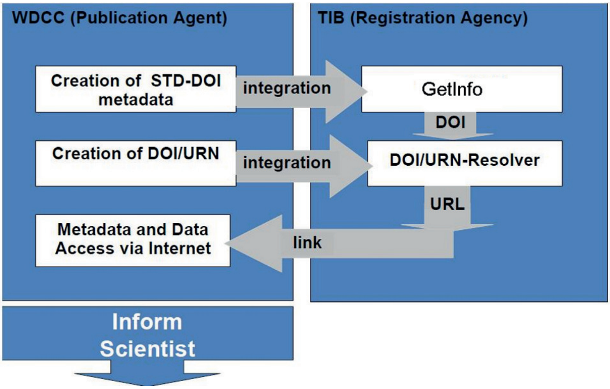


Figure F.12 Workflow for publication of environmental data sets (courtesy of Heinke Hoeck, DM/DKRZ; Hoeck, 2010)

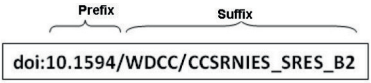


Figure F.13 Composition of the DOI (courtesy of Heinke Hoeck of DM/DKRZ; Hoeck, 2010)

profit organisation formed in London on 1 December 2009). In fact, not DataCite itself but its members are the national institutions who function as registration agencies. WDCC/DKRZ in Germany, for example, has a contract with the TIB as their registration agency. For a university or research institute in the USA, for example, the (contract) partner would be a US registration agency (like the California Digital Library as a member of DataCite).

Meanwhile three further registration agencies in Germany are offering DOI registration services, i.e. the German National Library of Medicine (ZB MED), the Leibniz Institute for Social Science (GESIS) and the German National Library of Economics (ZBW). By doing this, the registration agencies facilitate and promote non-profit online publications:[80]

[80] ZB MED lists these types of digital content for which it can assign DOIs to: publications such as journal articles, research reports, websites with scientific/academic contents, congress publications, posters, and Research data such as image data, videos, audio data, statistical data, sequence data, interview data.

- TIB for all German data centres in the fields of science and engineering, architecture, information technology, mathematics,
- ZB MED for the fields of medicine, health, nutrition, the environment and agriculture,[81]
- GESIS for social science data in Germany,[82]
- ZBW for economic literature and data.[83]

3 Current status of the research infrastructure of observational climate science programmes in Germany

In section 2, earth system and climate modelling projects and their data management arrangements have been described. The complex mathematical models producing the primary data need to be validated, however, by and their results compared with observations done either *in situ* within the compartments of the earth system or remotely from space, for example. Therefore observational earth science is indispensable in climate research. Since a comprehensive treatment of the research infrastructure for climate scientists developing sensors and deploying these in the field, for example, is beyond the scope of this study, merely the main data archives for climate-relevant observation programmes and projects are briefly described in the following six sections. Since the WDCC also holds some observational data sets, it is included (section 3.6).

3.1 German Weather Service (DWD, Deutscher Wetterdienst), Offenbach

The DWD, a public institution under the German Federal Ministry of Transport, Building and Urban Development,[84] is responsible for meeting meteorological requirements arising from all areas of economy and society in Germany, The DWD, founded in 1952 as National Meteorological Service of the Federal Republic of Germany, provides services in the form of weather and climate information. These include meteorological safeguarding of aviation and marine shipping as well as issuing warnings of meteorological events that could endanger public safety and order.

[81] http://www.zbmed.de/en/about-us/who-we-are/doi-service.html.

[82] http://www.gesis.org/dara/en/home/?lang=en.

[83] http://www.zbw.eu/e_services/e_publication_services.htm.

[84] http://www.dwd.de/bvbw/appmanager/bvbw/dwdwwwDesktop?_nfpb=true&_pageLabel=dwdwww_wir_ueberuns&_nfls=false.

DWD operates Germany's densest meteorological and climatological observing network, in which data have been collected for many decades for further processing and archiving. Approximately 100 billion climate data entries were gathered, some time series dating back to the 18th century. Data stem from surface weather stations, upper-air stations and ships, as recorded every day at the synoptic hours, and are disseminated or released in encoded form and archived and reusable in synoptic ordering. They are archived as collected during a single day in the various observing networks, e.g. the main synoptic-climatological network, the secondary climate and precipitation network etc., for climatological purposes, and are stored in various formats, stages of validation and sorting orders.[85] In the future, phenological data[86] will be used more and more for trend analyses in climate diagnosis, as the dates of the beginning of many phenological phases can be shown to correspond to trends in temperature. Besides being a curator of these observational data, the DWD is hosting several transnational and global data centres (Table F.2).

Table F.2 Data centres at the Deutscher Wetterdienst (DWD) (table modified from source)[87]

Acronym	Purpose/task
NKDZ	The National Climate Data Centre makes climatological data collected by the DWD available for users. Development of methods and applications for quality and data management
CM-SAF	Satellite Application Facility on Climate Monitoring
GCC	Global Collecting Center: international centre to receive and to distribute the non-real-time data under the Marine Climatological Summaries Scheme
GZS	Global Center for Weather Reports from Ships: national centre to archive the worldwide data of maritime meteorological platforms
ACD	Archive of the worldwide CLIMAT data: national archive of monthly and annual climate data (monthly and annual means or totals), which are provided by about 2500 stations worldwide on a monthly base

[85] http://www.dwd.de/bvbw/appmanager/bvbw/dwdwwwDesktop?_nfpb=true&_windowLabel=dwdwww_main_book&switchLang=en&_pageLabel=dwdwww_book.

[86] Phenology deals with the periodically recurring growth and development phenomena of plants during the course of a year. The starting time of characteristic vegetation stages (phases) is observed and recorded. These beginnings are closely connected to the weather and climate and are thus suited for the most varied areas of application and for manifold scientific studies.

[87] http://www.dwd.de/bvbw/appmanager/bvbw/dwdwwwDesktop?_nfpb=true&_pageLabel=_dwdwww_klima_umwelt_datenzentren&T21400353661157011331648gsbDocumentPath=BEA__Navigation%2FKlima__Umwelt%2FKlimadatenzentren.html%3F_nnn%3Dtrue&lastPageLabel=_dwdwww_klima_umwelt_klimadaten_deutschland.

GPCC	The Global Precipitation Climatology Centre analyses the monthly precipitation on earth's land surface based on rain gauge station data. It supports global and regional climate monitoring and research and is a German contribution to the World Climate Research Programme (WCRP) and to the Global Climate Observation System (GCOS)
GSNMC	The Global Climate Observing System Surface Network Monitoring Center monitors the availability and quality of CLIMAT reports from stations of the GCOS surface network exchanged via the Global Telecommunication System of WMO

To stay current with technological developments and ensure continuity in its services as National Meteorological Service for the protection of life, the DWD needs to replace its technical infrastructure at least every 10 years. By participating in projects such as VGISC, SIMDAT and C3-Grid,[88] the DWD contributes to the development of new tools and infrastructure for improved services.

In VGISC, the DWD together with Meteo France and the UK Met Office are creating global information system centres (GISCs), data collection and production centres (DCPCs) and national centres virtually tied together according to the WIS standard of the World Meteorological Organisation (WMO-Information System). In the EU-funded project SIMDAT (Data Grids for Process and Product Development using Numerical Simulation and Knowledge Discovery) and with the additional partners of the European Centre for Medium-Range Weather Forecasts (ECMWF) and the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT), these three meteorological services are developing generic Grid technology for the solution of complex application problems.[89]

In the project Collaborative Climate Community Data and Processing Grid (C3-Grid), an effective grid-based environment for earth system research in Germany was created to enable distributed data processing and inter-institutional exchange of large-volume model and observational data.

[88] http://www.c3grid.de/index.php?id=44&L=1.

[89] http://www.dwd.de/bvbw/appmanager/bvbw/dwdwwwDesktop?_nfpb=true&_pageLabel=dwdwww_zusammenarbeit&T17401110631149743806488gsbDocumentPath=Navigation%2FOeffentlichkeit%2FZusammenarbeit%2FTechnikprojekte%2FHome__node.html%3F__nnn%3Dtrue.

3.2 The German Remote Sensing Data Center (DFD, Deutsches Fernerkundungsdatenzentrum), Oberpfaffenhofen[90]

DFD is part of Germany's national research centre for aeronautics and space (DLR, a chartered non-profit organisation). DFD and DLR's Remote Sensing Technology Institute (IMF) together comprise the Earth Observation Center (EOC),[91] whose institutional funding is governed by the research programme of the Helmholtz Association, Germany's largest scientific organisation.[92] From the Earth Observation Center website:[93] "IMF and DFD are the leading national earth observation research and development institutions with public funding. DFD's expertise is in operational tasks (Center for Satellite-based Crisis Information ZKI, National Remote Sensing Data Library NRSDL, international data reception facilities) as well as in the application of remote sensing to obtain information about the land surface, the atmosphere and civil crisis situations. IMF focuses on scientific research related to sensor specific algorithms and methodology development, image processing, and data product development."

While IMF develops methodologies for processing radar data and sophisticated image analysis, specifically for marine remote sensing, DFD's focus lies in the development of user-oriented products and services. With DFD's national and international receiving stations, direct access to data from earth observation missions is possible and information products from the raw data are being derived. Dissemination of these products to users and curating all data in the National Remote Sensing Data Library for long-term use are further tasks regularly performed by DFD. Applications focus on the land surface, civil security and the atmosphere. For core competences and seven departments of DFD, see the website.[94]

3.2.1 Atmosphere (AT)

Research and development in the AT department entails basic research, new applications and data products. The department combines satellite measurements with numerical models to develop innovative, demand-driven services for future operational implementation, as well as technologies, and data products relating to the atmosphere. Research and development extends also to aerosols, radiation, trace gases and atmospheric dynamics. By these activi-

[90] http://www.dlr.de/caf/en/desktopdefault.aspx/tabid-5278/8856_read-15911.
[91] http://www.dlr.de/caf.
[92] http://www.helmholtz.de/en.
[93] http://www.dlr.de/caf/en/desktopdefault.aspx/tabid-5277/8858_read-15912.
[94] http://www.dlr.de/caf/en/desktopdefault.aspx/tabid-5278/8856_read-15911.

ties, DLR contributes to the Global Earth Observation System of Systems[95] (GEOSS), offering access to remote sensing, geospatial static and in-situ data, information and services via the GEO Portal[96] (operated by the European Space Agency and the Food and Agriculture Organization of the United Nations).

3.2.2 Land surface (LS)

Natural processes and human activities are constantly influencing the Earth's surface. These changes can be detected and analysed using remote sensing methodologies. The LS department of the DFD defines systems and mission parameters for new optical earth observation missions, supervises missions and customises information derived from optical and synthetic aperture radar (SAR) sensor systems to meet the needs of investigators from geology, soil science, geography, agriculture and forestry. The LS department thus enables the applied geosciences to benefit from the engineering-related achievements of the EOC.

3.2.3 World Data Center for Remote Sensing of the Atmosphere (WDC-RSAT)[97]

DFD has hosted and operated the WDC-RSAT since 2003, which holds and offers free access to atmosphere-related satellite-based data sets (raw as well as value added data), information products and services.[98] Data on atmospheric trace gases, aerosols, dynamics, solar radiation, cloud physical parameters and surface parameters (land and sea), such as the vegetation index for the northern and southern hemisphere and surface temperatures are available. Data may either be directly accessed if they are stored at the WDC-RSAT or found through the WDC-RSAT portal if they are safeguarded by other providers (compare with section 6.2).

The WDC-RSAT is a member of the WDC cluster Earth System Research and is a publication agent for digital data (see section 6.2 and Figure F.10).

[95] http://www.earthobservations.org/geoss.shtml.

[96] http://www.geoportal.org/web/guest/geo_home.

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

[98] http://wdc.dlr.de/about/index.php.

3.3 World Data Center for Marine Environmental Sciences (WDC-MARE, Biogeochemistry, Circulation, and Life of Present and Past Oceans), Bremen[99]

WDC-MARE (founded in 2001) curates data from marine environmental research and facilitates the international collection and exchange of all forms of marine data. WDC-MARE collects, critically reviews and disseminates data related to global change and earth system research in the fields of environmental oceanography, marine geosciences, and marine biology. Its focus is on georeferenced data, and the PANGAEA information system is used as WDC-MARE's long-term archive and publication unit (Publishing Network for Geoscientific and Environmental Data).[100]

WDC-MARE is maintained by AWI, a research centre of the Helmholtz Association, and the Center for Marine Environmental Sciences (MARUM), University of Bremen, with additional support from the DFG Research Center Ocean Margins. On behalf of Germany's participation in the Integrated Ocean Drilling Program[101] (IODP), Bremen University operates the international core repository[102] (Bremen Core Repository, BCR), i.e. 1100 m^2 refrigerated storage area, Since 1994 about 142 km of deep-sea cores from 83 ocean drilling cruise legs in around 210,000 boxes have been collected. The core repository is visited by approximately 200 scientists per year for sampling and around 50,000 samples are taken by guests and by the repository staff. The BCR also houses 142 km of core taken in the North and South Atlantic and Arctic Oceans, and the Mediterranean and the Black Seas, while the other two core repositories in the world maintain cores from the Pacific Ocean plate, the Southern Ocean south of 60řS latitude (except Kerguelan Plateau), the Gulf of Mexico, the Caribbean Sea (at the Gulf Coast Repository in Texas, USA, more than 116 km of core), and from the Indian Ocean and marginal seas, the western and northern marginal seas of the Pacific region, defined by the plate boundaries that extend from the Aleutian trench to the Macquarie Ridge (Kochi Core Repository, Japan, 91 km of core; source: BCR brochure).[103]

WDC-MARE was the first publication agent in Germany centre using the DOI to automatically register scientific and technical data sets as part of a full bibliographic citation (see Figure F.10).

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

[100] http://www.pangaea.de/about.

[101] http://www.oceandrilling.org.

[102] http://www.marum.de/en/IODP_Core_Repository.html.

[103] http://www.ecord.org/pub/BCR.pdf.

3.4 National Oceanographic Data Centre for Germany (NODC), Hamburg[104]

The NODC is the focal point of the national and international exchange of oceanographic data. It acquires the marine data sampled by German institutes and agencies, archives it and promotes data exchange on a national and international level. Both NODC and WDC-MARE participate in the IOC/UNESCO International Oceanographic Data and Information Exchange (IODE). NODC also curates the Baltic and North Sea/North-East Atlantic monitoring data according to the resolutions of the Oslo/Paris[105] and Helsinki Conventions,[106] respectively. The Marine Environmental Monitoring Network in the North Sea and Baltic Sea[107] (MARNET) presently comprises ten automated measuring stations. The NODC is hosted by the Federal Maritime and Hydrographic Agency (BSH) in Hamburg. Data are curated in the Marine Environmental Database[108] (Meeresumweltdatenbank MUDAB) which was developed jointly by BSH and the Federal Environmental Agency (Umweltbundesamt UBA), Dessau.

The oceanographic data, which generally are relevant for climate research, are based on in-situ hydrographic measurements from regular surveys as well as from several large oceanographic research programmes. The database covers about 5500 cruises, with data from 250,000 stations (more than 13 million records). The data are quantitative information about the environmental status of the North and Baltic Seas, i.e. values of physical variables like temperature and salinity, chemical variables like nutrients and organic, inorganic and radiochemical components and biological data (distribution of benthos species, for example).

Data originators, external experts and members of the public may access the MUDAB via a Web client.[109] The data are categorised into:

1. water and suspended matter: samples at individual or repeatedly visited stations well as from light vessels and buoys,
2. sediment and pore water: samples at individually or repeatedly visited stations,
3. biota: organisms living in the water body,
4. benthos: organisms living on the ocean bottom

[104] http://www.bsh.de/en/Marine_data/Observations/DOD_Data_Centre/index.jsp.

[105] http://www.ospar.org/content/content.asp?menu=00170301000000_000000_000000.

[106] http://www.itameriportaali.fi/en/tietoa/helcom_seuranta/en_GB/helcom_seuranta.

[107] http://www.bsh.de/en/Marine_data/Observations/MARNET_monitoring_network/index.jsp.

[108] http://www.informus.de:8080/mudab/welcome.faces.

[109] http://www.informus.de:8080/mudab/welcome.faces.

The inventory[110] for these categories of MUDAB (based on a query in May 2007) is given as:

1. vertical profiles: about 22 million data sets for 345 parameters,
2. time-series of weather data, water levels, temperature, salinity and sometimes nutrients and oxygen: about 13 million data sets for 20 parameters, and a historical climate time series for temperature and salinity from four stations in the Baltic and North Seas,[111]
3. about 619,000 data sets for 293 parameters,
4. about 33,000 data sets for 43 parameters,
5. about 1000 data sets for five parameters.

3.5 National Bathymetric Data Centre, Rostock

As part of the BSH, the Bathymetric Data Centre archives bathymetric data from German marine research missions at the BSH branch in Rostock, Germany. Bathymetric data sets from all oceans except the Southern Ocean are available. Sea depths are important marginal information in earth system and climate research. By delivering bathymetric data sets to the International Hydrographic Office[112] (specifically to the IHO Data Center for Digital Bathymetry, IHO-DCDB), BSH supports indirectly the development and updating of global and regional bathymetric charts. International marine research institutions have central access to worldwide bathymetric data through the IHO-DCDB.

A non-exhaustive list of marine geophysical data and information, data compilations and data holders was published by the Commission on the Limits of the Continental Shelf (CLCS) of the United Nations Division for Ocean Affairs and Law of the Sea[113]

3.6 World Data Center for Climate (WDCC)[114]

While the vast majority of the data in the CERA database at WDCC results from global or regional climate modelling experiments, some data sets from observational climate research programmes and projects have also been

[110] http://www.informus.de:8080/mudab/documents/070530_mudab_webclient_faltblatt.pdf.

[111] http://www.bsh.de/de/Meeresdaten/Beobachtungen/MARNET-Messnetz/Klima_MARNET/Klima.jsp.

[112] http://88.208.211.37/srv1.

[113] http://www.un.org/Depts/los/clcs_new/sources/data_portals_holders_websites.pdf.

[114] http://www.bsh.de/en/Marine_data/Hydrographic_surveys_and_wreck_search/Bathymetry/index.jsp.

archived and are being re-used. The high values for data set size and number of downloads per project are associated with data from numerical model experiments (Figure F.14).

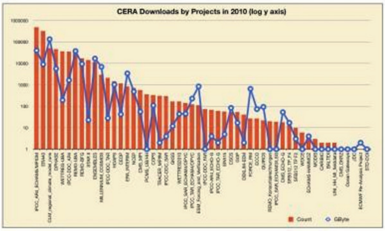


Figure F.14 Number of (red bars) and size of (blue line) data sets downloaded from the CERA database in 2010[115](note the logarithmic scale of the y-axis). Along the x-axis the projects are sorted such that the project data sets which were downloaded most often come first (left side), followed by downloads of project data sets that were less frequent (towards the right).

Data sets from some important observational measurement campaigns aiming to better understand climate aspects of the earth system appear on the right of the x-axis of Figure F.14. These and other observational projects are listed in Table F.3. Most are from field experiments and monitoring efforts of the last three decades, but there are also some historical compilations, such as the "Global Land Cover Reconstruction AD 800 to 1992" covering over 1000 years and observational set ups that are still ongoing (e.g. at the Wettermast Hamburg).[116] The land-based observation areas are located in Germany and Austria, the oceanic ones in the North Atlantic and Arctic Ocean and adjacent seas. Satellite data sets often have a global coverage, as do those from the WOCE (World Ocean Circulation Experiment) Hydrographic Programme. The experiments listed in the first column of Table F.3 as well data sets from numerical experiments can be found, for example, by "browsing by experiment" after accessing the database through the CERA portal[117] (cf. lower left in Figure F.6).

Table F.3 Data from observational programmes at the WDCC

Project name, purpose	Geographic region	Variables measured, processes observed	Instruments, carriers, platforms

[115] http://www.dkrz.de/daten-en/wdcc/statistics/wdcc-statistics-2010.

[116] http://wettermast-hamburg.zmaw.de.

[117] http://cera-www.dkrz.de/WDCC/ui/Index.jsp.

ACSYS, air mass modification in on-ice air flows (1998), arctic atmospheric boundary layer and sea ice interaction study (2003)	Arctic	Turbulent fluxes, standard meteorological parameters, radiosonde measurements[118]	Research aircraft, buoys, ship
AQUA_AMSRE:[119] to better understand the Earth's water cycle and determine if the water cycle is accelerating as a result of climate change	Global ocean	Geophysical parameters, including SST, wind speed, atmospheric water vapour, cloud water, and rain rate, local copy of NASA data	Advanced Microwave Scanning Radiometer (AMSRE), Satellite Aqua
AVHRR Pathfinder SST v5,[120] a more accurate, consistent land mask, higher spatial resolution, inclusion of sea ice information, better flagging of aerosol-contaminated data retrieval	global	4 km AVHRR Pathfinder version 5, SST monthly means data set (daytime measurements), local copy of NASA data	Advanced Very High Resolution Radiometer (AVHRR), NOAA satellite
ALKOR, BASIS, eight field experiments (1998, 2000, 2001) to collect a comprehensive data set to validate the coupled model system BALTIMOS[121] for the Baltic	Central Baltic Sea	Atmospheric boundary layer structure and processes and air-sea-ice interaction over areas with inhomogeneous sea ice cover; atmospheric boundary layer structure over open water under different synoptic conditions such as cold-air advection, warm-air advection or frontal passages, radiosonde	Various, RV *Alkor*

[118] http://www.erh.noaa.gov/gyx/weather_balloons.htm.

[119] http://ssmi.com/amsr/amsr_data_description.html.

[120] http://podaac.jpl.nasa.gov/SeaSurfaceTemperature/AVHRR-Pathfinder.

[121] http://www.borenv.net/BER/pdfs/ber7/ber7-371.pdf.

ARKTIS 1988, atmospheric boundary layer in the marginal ice zone, investigation of boundary layer modification and certain cloud structures in cases of off-ice and on-ice air flows	Fram Strait	Mean structures, variances and covariances at different distances from the ice edge	ships, aircraft, Icebreaker *Polarstern*, RV *Valdivia*, several aircraft operating from the airport at Longyearbyen on Spitsbergen
ARKTIS 1991, cellular convection, investigation of cold air outbreaks from the surrounding Arctic ice sheets	Norwegian Sea	Observation of newly formed boundary layer: its depth, mean temperature, moisture with increasing distance from the ice edge	RV *Valdivia*, research aircraft FALCON-20 and DO-128
ARKTIS 1993, air mass modification in off-ice air flows, investigation of cold air outbreaks from the Arctic sea ice onto the open water	West Spitsbergen current	Aerological data collected at three land stations (Bear Island, Danmarkshavn, NyAlesund), radiosonde	RV *Polarstern*, RV *Valdivia*, RV *Prof. Multanovsky*, aircraft Falcon and DO-128
ASTEX 1992, Atlantic Stratocumulus Transition Experiment, observations with modelling activities to investigate the consequences to the atmosphere and ocean of marine stratocumulus clouds and their life-cycle variations, including the important broken cloud regimes	Azores and Madeira Islands, north-eastern Atlantic	156 radiosonde ascents	Satellite, airborne, island, buoy, RV *Valdivia*

BALTEX[122] Baltic Sea Experiment, meteorological, hydrological and oceanographic research to explore and model the various mechanisms determining the space and time variability of energy and water budgets of the BALTEX region and this region's interactions with surrounding regions	Baltic Sea, Danish Straits	Lateral exchange with the atmosphere outside the BALTEX region, wind stress at the sea surface, evaporation and precipitation over land and sea, heat and energy flux at the air–sea and air–land interfaces, including radiation, river runoff, in- and outflow through the Danish Straits (each country providing its own set of meteorological parameters)[123]	Radiances, OVS atmospheric temperature-humidity and ice/snow correlative data sets,[124] satellites
BASIS 1998, 2001, Baltic Air Sea Ice Study	Gulf of Bothnia, Baltic Sea	Standard meteorological measurements, surface measurements at four land stations (Marjaniemi, Oulu, Kuivaniemi, Haparanda), radiosonde ascents	RV *Aranda*, research aircraft DO-128
COPS[125] (2007), Convective and Orographically induced Precipitation Study, to advance the quality of forecasts of orographically induced convective precipitation by 4D observations and modelling of its life cycle	Black Forest, Germany	Atmospheric Radiation Measurement (ARM), tropospheric profiles of water vapour and wind, and many more,[126] radiosonde ascents	Soil moisture network, research aircraft, mobile meteorological masts

[122] http://www.baltex-research.eu/background/bp1.html.

[123] http://www.baltex-research.eu/data.

[124] http://eosweb.larc.nasa.gov/GUIDE/dataset_documents/base_isccp_d1_d2_dataset.html#overview.

[125] http://www.cops2007.de.

[126] http://www.meteo.uni-bonn.de/messdaten/passive-microwave-radiometer-admirari/cops-measurements-2.

AVISO,[127] Archiving, Validation and Interpretation of Satellite Oceanographic data DUACS, Data Unification and Altimeter Combination System, processing data from altimeter missions to provide a consistent and homogeneous catalogue of products for varied applications, both for near-real-time applications and offline studies	Global	Altimeter, monthly gridded sea surface heights computed with respect to a 7-year mean (averaged sea surface heights averaging month by month), monthly means created from weekly sea level anomaly maps	Satellites Jason1, Topex/ Poseidon, Envisat, GFO, ERS1 and 2, Geosat
DAMOCLES 2007–2008, Developing Arctic Modeling and Observing Capabilities for Long-term Environmental Studies, Hamburg Arctic Ocean Buoy Drift Experiment	Central Arctic Ocean	Ice drift, sea ice	Array of 16 drifting autonomous buoys
FGGE 1979, First GARP Global Experiment	Central equatorial Atlantic Ocean	Near-surface oceanographic and surface meteorological data,[128] 291 radiosondes	RV *Meteor* and 40 other ships
CARIBIC (1997–2002) and CARIBIC–LH (2007), Civil Aircraft for the Regular Investigation of the atmosphere Based on an Instrumented Container, to study and monitor important chemical and physical processes in the Earth's atmosphere	Along intercontinental flight tracks	Suite of variables collected during each flight and analysed in-flight on board or later in the laboratory[129]	Commercial aircraft

[127] http://www.aviso.oceanobs.com/en/data/products/sea-surface-height-products/global/index.html.

[128] http://www.sciencedirect.com/science/article/pii/007966118690008X.

[129] http://www.caribic-atmospheric.com.

FRAMZY (1999, 2002, 2007, 2008, 2009), five field experiments to investigate the properties of Fram Strait cyclones, their cyclogenetic conditions on the large- and meso-scale, and their local effects on sea ice drift and sea ice distribution and, thus, on the freshwater flow through the Fram Strait. The data were used for validation of cyclone simulations with coupled mesoscale models of the atmosphere-ice-ocean system	Fram Strait, Greenland Sea	Meteorological data, ice, ice drift	14 autonomous ice buoys, RV *Aranda*, research aircraft Falcon, satellites (NOAA-AVHRR, RADARSAT, DMSP-SSM/I)
FRONTEX 1989, atmospheric fronts, to investigate cold fronts moving in from the North Sea and reaching the coastal area with high temporal and spatial resolution	Coastal area of northern Germany, Heligoland, Schleswig, Hanover, Emden, Berlin	Ground-based remote sensing and in-situ measurements, physical properties of sea and land surface (roughness, humidity, temperature, heat conduction and heat capacity), radiosonde ascents	Research vessel, research aircraft POLAR-2, POLAR-4, DO-128
GEBCO,[130] General Bathymetric Chart of the Oceans	Global	Depth soundings	Ships, various others
Glacier monitoring data of Austria[131]	Austria[132]	Hydrological parameters, glacier mass balance	

[130] http://www.bodc.ac.uk.

[131] http://imgi.uibk.ac.at.

[132] http://imgi.uibk.ac.at/iceclim/glacierinventory.

GOP (2007), General Observation Period of Priority Program on Quantitative Precipitation Forecasting	German Weather Service networks	Rain gauges, weather radar, Light Detection And Ranging (Raman Lidar), ground-based Global Positioning System (GPS), lightning, satellite data,[133] radiation, microwave radiometer, ceilometer, cloud radar, wind profiler, radiosonde ascents	Ground stations, satellites (Meteosat Second Generation MSG, MODIS and MERIS)
HADEX,[134] global climate extremes indices	Global	Land-based climate extremes data set, 27 indices of temperature and precipitation on a 2.5 × 3.75řgrid from 1951 to 2003. Indices represent seasonal and/or annual values derived from daily station data	

[133] http://www.geo.fu-berlin.de/met/ag/sat/satdaten/index.html.

[134] http://www.hadobs.org.

KONTROL (1984, 1985), experiment on convection and turbulence with the objectives (1) to observe the formation and time variation of regularly organised convection in the lower troposphere as a function of the mean atmospheric flow and the lower boundary condition and to quantify the dependence of the vertical transports of momentum, heat and water mass on various scales of motion, and (2) to determine the mean and turbulent quantities within the marine atmospheric boundary layer, including the large-scale horizontal and vertical advection of momentum, heat and water vapour, cloud microphysics and the radiation field. Goal: to test existing convection models and to provide an observational background for the extension of theoretical concepts	German Bight, south-eastern North Sea	Continuous aerological and surface observations at fixed stations (island Heligoland, Borkumriff, moored platform, ships), detailed observations during special periods done by aircraft, supporting observations, such as satellite images, cloud photography, surface and upper air large-scale fields from routine data	RV *Valdivia*, RV *Meteor*, RV *Gauss/ Poseidon*, research platforms Nordsee and Elbe 1, research aircraft (Falcon 20, DO-28 Skyservant, Hercules C-130)

LOFZY 2005, first field experiment on cyclones over the Norwegian Sea, low-pressure systems (cyclones) and the climate system of the North Atlantic	Lofoten archipelago	From ship: meteorological observations (radiosondes, standard parameters), oceanographic CTD measurements. From aircraft: observation of synoptic conditions with high spatial and temporal resolution. Additionally deployment of 23 autonomous marine buoys in advance of the campaign to measure drift, air-temperature and -pressure and water-temperature	RV *Celtic Explorer*, research aircraft Falcon
MODIS_ACDNC, adiabatic cloud droplet number concentration daily value	Global	Cloud droplet number concentration is derived from MODerate Resolution Imager Spectroradiometer (MODIS)[135]	NASA satellite Terra (EOS AM)[136]
Reconstruction of global land use and land cover AD 800 to 1992[137]	Global, 30 minute resolution	Population data,[138] three human land use types (crop, pasture) and 11 natural vegetation types[139]	
SeaWinds on QuikSCAT[140] Level 3 Daily Gridded Ocean Wind Vectors[141]	Global	Gridded values of scalar wind speed, meridional and zonal components of wind velocity, wind speed squared and time given in fraction of a day	NASA satellite QuikSCAT (Quick Scatterometer)

[135] http://modis.gsfc.nasa.gov/about.

[136] http://terra.nasa.gov.

[137] http://www.mpimet.mpg.de/fileadmin/publikationen/Reports/WEB_Bze_51.pdf.

[138] Atlas of World Population History (McEvedy and Jones 1978)

[139] http://www.sage.wisc.edu/pubs/abstracts/ramankuttyGBC1999.html.

[140] http://winds.jpl.nasa.gov/missions/quikscat/index.cfm.

[141] http://idn.ceos.org/KeywordSearch/Metadata.do?Portal=idn_daacs&KeywordPath=[Source_Name%3A+Short_Name%3D%27QUIKSCAT%27]&NumericId=27759&MetadataView=Text&MetadataType=0&lbnode=mdlb2.

Wettermast Hamburg	Hamburg-Billwerder	Ground-based continuous measurement of weather data since 1995 at several height levels: 2, 10, 50, 70, 110, 175, 250 m above the ground	Broadcasting tower of 300 m height
WOCE Hydrographic data, Onetime and Repeat Survey,[142] carried out mostly between 1990 and 1998	World ocean along trans-oceanic sections	Full-depth CTD profiles of temperature, salinity, oxygen, from water bottle samples chemical properties were analysed, including nutrients, chemical oxygen demand, chlorofluorocarbons, tritium, helium and other tracers	Many research vessels and volunteer observing ships worldwide

CTD: conductivity-temperature-depth; SSS: sea surface temperature.

[142] http://woce.nodc.noaa.gov/wdiu/diu_summaries/whp/index.htm.

4 Results of a survey concerning climate research practices in six German institutions

In addition to the climate modelling and climate data resources available in German data centres that have been already described in sections 2 and 3, respectively, I used a sample questionnaire to learn directly perceptions of the infrastructure from eight researchers working in representative climate research departments of major research institutions in Germany (see Table 4). The questionnaire was designed by Dennis Spohr of the Centre of Excellence Cognitive Interaction Technology (CITEC, Bielefeld) and supplied to all subject-specific chapter authors (Spohr, 2010).

The aim of the survey was to hear first-hand from climate scientists the details regarding the infrastructure with respect to data and literature which they experience on a day-to-day basis. After explaining first the group's work focus, they described the characteristic data lifecycles, data processing data formats, data management, access to data and the ways that publication and exchange of research data are customary in their institution. The third set of questions dealt with the organisation of literature, to what extent publication of literature and research data jointly was established and whether Open Access was customary. Finally the group's specific outlook was queried for the future developments in data and literature infrastructure in climate science. The interviewed scientists are affiliated with institutions/organisations and departments focusing on either research, service providing and infrastructure development (Table F.4).

Table F.4 German research institutions participating in the survey regarding climate research practices

	Names of institution, department and/or group of researchers interviewed	Personnel and equipment		
		People	PCs	Other
1 r	Max Planck Institute for Meteorology MPI-M, Hamburg (a) Dept. "Land in the Earth System", group "Terrestrial Remote Sensing" (b) Dept. "Atmosphere in the Earth System", group "Observations & Process Studies"	(a) 8 (b) 15	35 for both groups	(a) 1 data server (b) 1 computer server
2 sp (r)	Climate Service Center (CSC), Hamburg, (an Institution at the Helmholtz-Zentrum Geesthacht),[143] Dept. "Climate System"	8	12	–
3 r	Karlsruhe Institute of Technology/Institute for Meteorology and Climate Research, Atmospheric Environmental Research (KIT/IMK-IFU), Garmisch-Partenkirchen a) Regional climate and hydrological modelling b) Collection and analysis of observational data	18	30	1 data server 1 powerful computer (Linux cluster)
4 id	Collaborative Climate Community, Data and Processing Grid (C3-Grid) at Alfred Wegener Institute for Polar and Marine Research (AWI), Bremerhaven	30		Distributed data processing capacity at various WDCs
5 r	Helmholtz-Zentrum Geesthacht (HZG), Geesthacht, Department "System Analysis", Paleoclimatology group	4	4	–
6 sp	Federal Maritime and Hydrographic Agency (BSH), Hamburg, Dept. "Marine Sciences", Division "Data and Interpretation Systems"	28	28	6 servers

id: infrastructure development ; r: research; sp: service providing.

[143] http://www.flyhy.eu/HZG.html.

4.1 Max Planck Institute for Meteorology (MPI-M), Hamburg

4.1.1 General information

Two researchers were interviewed, one (R1) speaking for the group "Terrestrial Remote Sensing" (department "Land in the Earth System" of MPI-Met), the other (R2) for the group "Observations & Process Studies" (department "Atmosphere in the Earth System" of MPI-Met). R2 is responsible for a new working group "observational data", installed to organise the whole suite of observations done by scientists at MPI-M. The institute has a third department, i.e. "Ocean in the Earth System".

In R1's group, the primary research objective is atmospheric observation by airplane, with instruments at the earth surface, e.g. at the Cloud Observatory on Barbados,[144] and remotely sensed data from satellites. Inside MPI-M there are numerous cooperations with the model developers, and outside with colleagues at the Meteorological Institute of the University of Hamburg and other national and international teams.

4.1.2 Data infrastructure

Both R1 and R2 collect a variety of atmospheric observational data and receive the level-1 satellite data. Processing of data includes quality control and the derivation of level-2 satellite data. The data are enriched by the calculation of geophysical parameters from the level-1 satellite data. The data are archived and re-used. The data types include those taken with a camera from an airplane or collected using a stationary webcam (GB range). Collected binary and ASCII data are in the GB to TB range. Some proprietary software is used in order to collect primary data (the instrumentation manufacturer may require this). Data products, i.e. "exploited" level-2 data are distributed, for example the HOAPS Climatology[145] (Hamburg Ocean Atmosphere Parameters and Fluxes from Satellite Data).

The data are further annotated with metadata, and frequently re-formatted according to the CF-convention (NetCDF Climate and Forecast (CF) Metadata Convention).[146] In some rare cases the scientists have to deviate from standard metadata formats, for example where there is no representation possible for soil moisture. As satellite data formats are quite heterogeneous, the scientists in this group cannot very often rely on conventional formats and software for data representation and processing. Developed software as

[144] http://www.mpimet.mpg.de/en/wissenschaft/atmosphaere-im-erdsystem/initiativen/barbadosstation.html.

[145] http://www.hoaps.zmaw.de.

[146] http://cf-pcmdi.llnl.gov.

well as primary and secondary data are stored and archived in a "repository" (definition: versioning is possible, using, for example, SVN[147]) within the group on a central file server. In addition copies of the software are kept on own storage devices on the hard disk of the office PC. R1 and R2 believe this to be a representative practice in climate science. Having seen the need, the MPI-M recently created the "Observation Steering Group" for observational data which is responsible for data management issues.

Data access is ensured by dissemination via the CERA database maintained at WDCC (see section 2.2), and also via the Integrated Climate Data Center[148] (ICDC) of CLISAP[149] (Integrated Climate System Analysis and Prediction), a "Cluster of Excellence" at the University of Hamburg which is funded by the German Research Foundation (Stockhause and Hoeck;[150] in Curdt and Bareth, 2010).

Research data and software are made available to close colleagues and other research projects, e.g. for use in publications. The general public may receive secondary data only. This restriction exists in order to give the group and its collaborators priority for publishing first results based on the data they collected and processed. Furthermore, there are only limited resources available in the group to guarantee maintenance and user support (there is, for example, no help desk). What is said here for observational data is transferable to modelling software: without additional support not all earth system models are usable as "community models", and this is generally true in climate science. Sporadically, however, software is made available to other institutions. This typically includes source code as software/models need to be compiled locally. The principal investigators in the group delivering the data must have priority with respect to analysis and publication, however. Data exchange happens via file transfer protocol or by shipping of a hard disk. For internal exchange within the MPI-M, the data server is used.

4.1.3 Literature

Online access is available to most subject-specific journals (cf. section 5). Tools that are used include JabRef, an open source bibliography reference manager using the file format BibTeX, a standard LaTeX bibliography format and others that individual researchers choose. Similar to other researchers in the climate community, publications both as print medium (e.g. article or book) and as electronic publication online or offline are preferred and

[147] http://svnbook.red-bean.com.

[148] http://icdc.zmaw.de/icdc.html?&L=1.

[149] http://www.klimacampus.de/clisap0.html?&L=1.

[150] http://icdc.zmaw.de/397+M59fd2f2bea8.html.

established. Publishing via a combination (e.g. book/CD-ROM or proceedings/website) is also preferable but is done rather infrequently.

The following scientific journals are often chosen for publication:

- *Hydrology and Earth System Sciences:* interactive Open Access journal (European Geosciences Union)
- *Remote Sensing of the Environment:* interdisciplinary journal for results on theory, science, applications and technology of remote sensing of Earth resources and environment (Elsevier)
- *Journal of Climate:* online journal (American Meteorological Society)
- *Remote Sensing:* online journal (Yale's Center for Earth Observation)
- *Biogeosciences:* interactive Open Access journal (European Geosciences Union)
- *Journal of Geophysical Research (Atmospheres):* journal (American Geophysical Union)
- *Geophysical Research Letters:* journal (American Geophysical Union).

Some publishers enable the exchange of data and/or literature and it is currently possible to publish both together. It is/would be desirable to be able to also publish a movie together with an online accessible reference. This would happen if, for example, the funding agency required this. However, a data server would be required to guarantee long-term storage.

As for other climate research groups, these interviewees confirm that Open Access practices are supported by having a dedicated database and data server. Access to literature is well established at large research facilities through, for example, the national licences of the German Research Foundation.

4.1.4 Outlook

Satellite data sets already now have large volumes. In the coming 5–20 years, satellite data volumes will increase exponentially. It will not be possible any longer to move such large data sets from A to B via ftp (TB to PB range). The EU funds sentinel satellites,[151] for which 1 TB of data per sensor per day are expected! The data to be returned from this mission is tractable for high spatial resolution of a small area only for a short time of interest. However, climate researchers who are more interested in a global coverage need to go elsewhere, i.e. to hosting data centres with computing facilities like ESA/ESRIN (European Space Agency/European Space Research Institute of the European Space Agency, near Rome), ECMWF and DKRZ. The user participates via cloud computing in the analysis of such data sets. For example, it would take 8 weeks to download 3-hourly global data at a horizontal

[151] http://www.esa.int/esaLP/SEMZHM0DU8E_LPgmes_0.html.

resolution of 6 km for the past 25 years. A workshop took place in Hamburg from 30 March to 1 April 2011 on Climate Knowledge Discovery. The goal is to find new fields of application for new technologies (e.g. pattern recognition software).

4.2 Climate Service Center (CSC, Helmholtz-Zentrum Geesthacht), Hamburg

The CSC is predominantly a service providing institution, i.e. a national agent brokering climate information to aid the dialogue between climate science and politics. The five departments of the CSC all focus their work on four sectors, i.e. agriculture, forestry, energy and health. The CSC prepares the knowledge derived from climate research in a practice-oriented way and conveys it to decision makers. Besides from the information on its own website,[152] a brief profile of the CSC can be found on the website of the Regional Science Service Center in the Southern African subregion (RSSC).[153] The CSC is one of the German institutions involved in this joint initiative of Angola, Botswana, Namibia, South Africa, Zambia and Germany, responding to the challenges of global change.

4.2.1 General information

The two senior scientists interviewed are meteorologists and belong to the CSC's “Climate Science” department.

4.2.2 Data infrastructure

Only binary data in the terabyte range are collected from the CERA database, processed and enriched, i.e. new quantities like, for example, precipitation are computed. Data are archived, re-used and distributed. No proprietary software is used, but software that was developed at the DKRZ.

Regarding data processing, metadata standards exist, but specification of all important information is not yet possible with them, e.g. the type of model grid that a used. Some software is being developed but also proprietary software is used (e.g. Aquacis, ARCGIS). Software and secondary data are mainly stored and archived on own storage devices. No person in group is specifically dealing with data management issues, although this is deemed necessary in order to make secondary data available and (re)usable for customers. Access to data is offered via a ftp server.

[152] http://www.climate-service-center.de/index.html.en.
[153] http://www.sasscal.org/.

Software and primary data are made available only among close colleagues, whereas secondary data are available also for members of the general public. The scientist should have priority for publishing using the data of which he/she was the originator, and the user/customer may be unable to use software necessary for data visualisation, analysis etc. due to lack of technical possibilities. If the group had the appropriate technological equipment, data could be exchanged with other groups within the centre. No software is given to other institutions, because the mission of the Climate Service Center is to deliver only products to customers. At present, general terms and conditions for data delivery are under development at the centre. Data exchange is supported by maintaining a ftp server.

4.2.3 Literature

As far as internal and external publications are concerned, the rules of the HZG (i.e. the umbrella organisation of the Climate Service Center) apply: all publications and lectures need to be registered before submission.

Publications as print medium as well as electronic publication online or offline are preferred and established at the CSC. Scientists use a broad spectrum of journals to publish papers, because climate and climate change impacts occur in many other disciplines. Some publishers enable the exchange of data and/or literature together. Open Access is to some extent established in the group, which is common in the discipline of climate research. Good networks facilitate this practice.

The interviewees envision for the future the creation of a network of climate service centres at the national and the international level and implementation of cooperation contracts.

4.3 Karlsruhe Institute of Technology/Institute for Meteorology and Climate Research, Atmospheric Environmental Research (KIT/IMK-IFU), Garmisch-Partenkirchen

4.3.1 General information

The interviewees are a senior scientist, whose research objective is regional climate and hydrological modelling, and a PhD candidate who concentrates on the collection and analysis of observational data.

4.3.2 Data infrastructure

Data collected include the forcing data of/for a global climate model, e.g. a data set generated by a modern, consistent and invariant data assimilation system such as the ERA set of reanalysis data.[154] The climate model output data then may serve, for example, as input data for regional model experiments. The data volume is in the terabyte range.

Observational data may come from X band radars detecting small particles in the atmosphere. Precipitation intensity can be measured with this device over areas of 50 km extent, data are in binary format and volume is. Another instrument, with which the observational group at this institution collects data, is a disdrometer, i.e. an instrument used to measure the drop size distribution and velocity of falling hydrometeors. Some disdrometers can distinguish between rain, graupel and hail. Thirdly, microwave transmission line integrated precipitation and humidity observations are carried out and deliver data. Data are stored on a universal mobile telecommunications system (UMTS) data server.

Software is developed to derive statistical quantities, such as occurrence frequencies for specific climate indicators. In the numerical modelling group, the software has been extended to some extent, e.g. the coupling software enabling a hydrological model to be linked to an atmospheric climate model.

Annotation of the data sets with metadata happens mostly on a bilateral basis, although metadata standards exist. Sometimes software is developed in-house (e.g. after completion of a numerical experiment) to derive further secondary data. Usage of proprietary software is not required for the formats of data sets, but when analysing data sets, such software products like ARCGIS and MATLAB are used.

Data sets (primary and secondary) as well as software are generally stored in a repository within the group and in a repository shared with other institutes or institute wide. This situation is comparable with another regional modelling group at the Max Planck Institute for Meteorology (the REMO group) in Hamburg. No particular person in the group at KIT/IFM-IFU is dealing with data management issues, but it would be desirable. Somebody is needed who has a combined IT knowledge and also knows what the climate scientists need.

In the project DEKLIM-QUIRCS[155] (Quantification of uncertainties in regional climate and climate change simulations), the partners decided on an exchange formats for model comparisons, the data exchange being achieved on a bilateral basis via shell access to a ftp data server.

[154] http://www.ecmwf.int/research/era/do/get/index.

[155] http://imk-ifu.fzk.de/441.php.

As far as publication and exchange of research data is concerned, the group makes software available only among close colleagues, and primary and secondary data also to other research projects. This restriction (no free delivery to the general public) is done for the well-known reasons that scientific judgment is required to handle data and software reasonably. Incentives for data exchange are good agreement within research projects, for example, functioning on a give-and-take basis. This is considered to reflect the general attitude within the discipline of climate science. If one should decide to also make software available to other institutions, this would, in most cases, also include the source code.

There are no special rules within the group regarding time frames for exchanging data. Raw data are kept as long as possible (several years). The time frame for archiving is guided by users needs. The ftp tool is used for model input data, but for model-produced data a special selection is agreed on and the subset is placed on the ftp server for distribution.

4.3.3 Literature

The group uses Zotero[156] as a free tool, i.e. a plug-in to the Firefox web browser, to collect, organise, cite and share their research resources. The preferred and established modi of publication are the print medium (e.g. article or book) and electronic publication online or offline. At times, but rather infrequently, a combination (e.g. book/CD-ROM or proceedings/website) is chosen for publication of research results. Again, this practice is felt to be representative for climate science.

Scientific journals that members of the group use are:

- *Hydrology and Earth System Sciences*: interactive Open Access journal (European Geosciences Union)
- Journal of Geophysical Research (Atmospheres): journal (American Geophysical Union)
- *ScienceDirect*:[157] SciVerse ScienceDirect scientific database contains more than 10 million journal articles and book chapters. Peer-reviewed full-text articles can be accessed.
- Comptes Rendus Geoscience: journal (Elsevier)
- *IEEE Geoscience and Remote Sensing*: journal (IEEE Geoscience and Remote Sensing Society).

The answer to the question as to whether there are publishers which enable the exchange of data and/or literature is affirmative.

Open Access according to the Berlin Declaration is to some extent established in the group, but is implemented after communication and not

[156] http://www.zotero.org.

[157] http://www.sciencedirect.com.

automatic. This is felt to be similar to the policies of other climate science institutions. The interviewees point out that Open Access practices need the technological backing of good and fast networks (internet, ftp, mailing via the post office).

4.3.4 Outlook

Metadata are found to be extremely important. For literature, abstracts need to be preserved and a good indexing scheme needs to be in place.

4.4 Collaborative Climate Community Data and Processing Grid (C3-Grid), AWI, Bremerhaven[158]

4.4.1 General information

The interviewee is coordinator of the project C3-Grid which has created a unified and transparent access to several large geographically distributed data archives.

C3-Grid's objectives are to provide a service mainly to the German climate science community and to new and emerging scientific communities such as the one primarily concerned with "climate impact", agencies concerned with strategies to adapt to climate change, but also biostatisticians, photon physicists, and others. C3-Grid is supported by the German Federal Ministry of Education and Research (BMBF).

An examples of the data management facilities adopted by C3-Grid and innovative developments in the climate community to alleviate the metadata generation, extraction and management is Fedora Enabled Repository with Cocoon (Federico),[159] which is a state-of-the-art AJAX front end for the Fedora Commons Repository developed in the scope of the Work Package 3 of WissGrid,[160] for the long-term preservation of research archives.

In their presentation "A Collaborative Environment for Climate Data Handling"[161] at the "Geoinformatics 2008 – Data to Knowledge" conference in Potsdam, Germany, Kindermann and Stockhause described some problems encountered in the C3-Grid project. The infrastructure for tracking the whole data cycle from discovery of input data to publication and archiving of the results is designed in three layers: a common data discovery layer, a data

[158] http://www.c3grid.de/index.php?id=32&L=1.

[159] http://www.wissgrid.de/publikationen/deliverables/wp3/WissGrid-D3.5.3-grid-repository-Federico.pdf.

[160] http://www.wissgrid.de/index_en.html.

[161] http://www.c3grid.de/fileadmin/c3outreach/material/Kindermann_C3-geoinf.pdf.

access layer, and a data manipulation layer. In the extended abstract published on page 31 in the proceedings of this conference[162] (Brady, Sinha and Gundersen, 2008), the authors summarise the challenges:

"In general, a major challenge in the project is to find or develop legal agreements that reflect an elaborate balance between technical progress and manageable effort. The established data and computing-service providers want to re-use their current implementations in order to minimise the maintenance of their software and the labor required to adapt to changes that are necessary when building the infrastructure. Yet, integrating collaborative environments always requires the creation of prototypes and the adoption of not-yet-established technologies. Different technological pathways have to be merged with respect to the specific needs of the existing scientific community and the future needs of intercommunity cyber infrastructures."

4.5 Coastal Research of the Helmholtz-Zentrum Geesthacht (HZG), Geesthacht

4.5.1 General information

The interviewed researcher is a senior scientist and his group's objective is "Climate simulation of the past millennia". His role in the group includes climate research, supervision of PhD students, and responsibilities as a member of the editorial board in some journals. He collaborates with one other group within the institute and about ten external groups.

4.5.2 Data infrastructure

Research data are generated with the help of climate models, typically within one calendar year. Archiving occurs almost simultaneously to data generation, while data analysis may take between 2 and 5 years. The data are collected in binary form and in the TB range. No own software development takes place in primary data generation nor is proprietary software being used (community models produce the data). On the basis of the primary model, output data summaries of the data and spatial and temporal coverage are stored. The data are annotated with metadata, whereby all steps in post processing are automatically generated and attached to the resulting secondary data. Almost always are the standard metadata used. For the derivation of secondary data, software is being developed; it is not necessary to use proprietary formats or software for secondary data representation.

Software and secondary data are stored and archived in a repository within the group while primary data are stored and archived on a supercomputer at

[162] http://pubs.usgs.gov/sir/2008/5172/sir2008-5172.pdf.

DKRZ. In addition, copies of the software are kept on own storage devices on the hard disk of the office PC. This situation is thought to be representative for climate research. There is no person specifically responsible in the group for data management issues, which the researcher finds satisfactory, because data management is an integral and closely related part of the scientific activity. It should be integrated in the research projects proper. The data are stored on servers accessible to every member of the group. The local computer network is the support tool for this.

Software and secondary data of this group are made available to the general public, but because of their large volume the primary data is only available to other research projects. Another reason for this restriction is that some sort of guidelines and interpretations from the originating group are needed, but is true for secondary data as well. It seems that the situation is representative in climate science and no further incentives for data exchange of research data and software seem to be necessary. Also as a matter of principle, software including its source code is made available to other institutions. The reason is that when software is requested to exactly understand the data post processing, software sharing is the most accurate and convenient way. Only when data sharing conflicts with current research projects occur, an exception/restriction is made. Data exchange is enabled by writing the data in a commonly used format and placing them on a public server from where they may be downloaded via ftp.

4.5.3 Literature

For this climate science group, publication of results in print media is still preferred but infrequent, whereas publication online or offline is the preferred and established practice, typical for climate science today. The combined publishing as book/CD-ROM or proceedings plus website is not preferred. There are publishers which enable the exchange of data and/or literature, but for this group it is currently not possible to publish data and literature together. However, this is not viewed as a problem. The reason is that the volume of data may be very large and the same data may be used in several publications, e.g. showing different types of analysis for recent research targets.

Open Access is to some extent established in the group, and this seems to be representative for the discipline, Open Access journals offering their own technologies for online publication.

4.5.4 Outlook

Data infrastructure: It is envisioned that centralised sites would be used as repositories and would manage the data sharing, taking into account proprietary issues and the need of scientific replication. These sites would accept data from different groups and make them available to other groups conditional on certain use (e.g. replication of published results), or extended analysis where coauthorship should be required.

Literature infrastructure: This researcher does not foresee a lot of further developments. Some journals will remain restricted and others would go for full Open Access (Golden Route). Journals will increasingly offer the possibility to publish graphic material, such as videos, and perhaps some journals will adopt wiki technologies. But since journals are commercial products after all, their success depends on marketing factors among scientists which are very difficult to predict.

4.6 Federal Maritime and Hydrographic Agency (BSH), Hamburg

4.6.1 General information

The interviewed researcher is a senior scientist in the group "Marine Data and Interpretation Systems", which is mainly concerned with national and international data management projects. The 28 employees in this group have 28 PCs at their disposal and access to six servers. The members of the group cooperate with six other groups within the institution and 72 external groups.

4.6.2 Data infrastructure

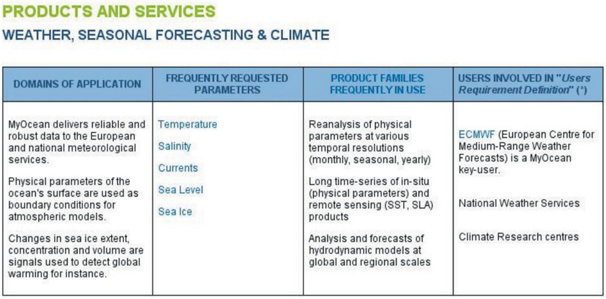
Data collection focuses on oceanographic real-time data, e.g. hydrographic, ocean current and wave height measurements, from the North West European Shelf and the Baltic Sea. The gathered data are undergoing real-time quality control, and are enriched by adding metadata and combining the data sets with data from other disciplines. Data are archived in the German National Oceanographic Data Centre, which is part of the BSH. The principal (re-)use of these data occurs within the spatial data infrastructure of the institution. The storage requirements for such oceanographic real-time data sets lie in the gigabyte range (ASCII and binary data).

From the primary data the following secondary data are typically derived: oceanographic products, maps, statistics, trends and some others, using also software that was developed within the institution. The secondary data are published in the internet and in the scientific literature. For metadata annotation, the standard ISO19115 is mostly used. If some partners in European projects do not use ISO19115, however, deviations from this standard may occur. When collecting, representing and processing the primary data, the group relies on proprietary formats and software to a great extent.

Primary and secondary data as well as developed software are all archived on an in-house data server. The interviewee feels that having an in-house data server is typical or “standard” for comparable groups in climate/ocean science and also that certain staff members are responsible for data management.

To assure access to their data, this BSH group enters into so-called service level agreements, for example in the EU project MyOcean[163] (ocean monitoring and forecasting), a 3-year project that started on 1 April 2009. In the module “Weather, seasonal forecasting and climate”, interested parties may register to view the full catalogue of products and services. A non-exhaustive list is shown in Figure F.15.

This interviewee’s partner’s group has also signed memoranda of understanding with the Baltic Operational Oceanographic System (BOOS)[164] and the North West European Shelf Operational Oceanographic System (NOOS).[165] The principal vision of both organisations is to develop and implement online operational marine data and information services. One of the objectives of BOOS is to contribute to ocean climate variability studies and seasonal climate prediction, and NOOS wants to establish a marine database from which time series and statistical analyses can be obtained, including trends and changes in the marine environment and the economic, environmental and social impacts.



PRODUCTS AND SERVICES

WEATHER, SEASONAL FORECASTING & CLIMATE

DOMAINS OF APPLICATION	FREQUENTLY REQUESTED PARAMETERS	PRODUCT FAMILIES FREQUENTLY IN USE	USERS INVOLVED IN "*Users Requirement Definition*" (*)
MyOcean delivers reliable and robust data to the European and national meteorological services. Physical parameters of the ocean's surface are used as boundary conditions for atmospheric models. Changes in sea ice extent, concentration and volume are signals used to detect global warming for instance.	Temperature Salinity Currents Sea Level Sea Ice	Reanalysis of physical parameters at various temporal resolutions (monthly, seasonal, yearly) Long time-series of in-situ (physical parameters) and remote sensing (SST, SLA) products Analysis and forecasts of hydrodynamic models at global and regional scales	ECMWF (European Centre for Medium-Range Weather Forecasts) is a MyOcean key-user. National Weather Services Climate Research centres

Figure F.15 List of parameters, their domains of application, products and user groups that typically request data for the module “Weather, seasonal forecasting and climate” via the MyOcean website[166]

[163] http://www.myocean.eu.org.

[164] http://www.boos.org.

[165] http://www.noos.cc.

[166] http://www.myocean.eu/web/19-products-and-services.php?domain=forecast.

The BSH group makes available research products, i.e. software, to close colleagues and other research projects, and primary data (some of which are output data from numerical models) both to these users and to the general public as well. Sporadically, software is also made available to other institutions (including the source code).

The main rule regarding the use of data provided by this group is that real-time data are handled as fast as possible. Data exchange is supported by ftp, http and web portals. Apart from these tools, spatial data infrastructures enable the re-use of data, while foreign data policies may hamper data re-use.

4.6.3 Literature

For the preparation of internal or external publications, software like MS-Office, LateX, FrameMaker and Adobe programs are used. Publications are done preferably on- and offline, albeit rather infrequently. Print media are not preferred as publication medium. However, this is viewed as being atypical for the discipline, where the preferred publishing medium is printed papers. This group does not foster Open Access practices with respect to literature.

4.6.4 Outlook

The way that the EU project INSPIRE is developing a data infrastructure is seen as a promising way for the future.[167]

5 Current status of Open Access in climate research literature

The implementation of Open Access publishing in science in general is contingent on the success of new ways of financing the system[168] and on a common understanding of intellectual property[169] rights.[170] Besides these aspects, in climate science the strong international interdependence on each other's data calls for agreements between partners which may be subjected to differing legal frameworks. Such agreements should be formulated at the beginning of the scientific workflows between potential users of the research results. The utilisation of climate science's large data sets, in particular, should be and is being backed by appropriate data policies and the development of technolo-

[167] http://www.inspire-geoportal.eu.

[168] http://www.ercim.eu/publication/Ercim_News/enw64/velterop.html.

[169] http://en.wikipedia.org/wiki/Intellectual_property.

[170] http://www.w3.org/IPR.

gies that enable transfer and access of the voluminous data sets in climate research. This aspect is dealt with in section 6.

In this section, first the organisation of scholarly information in general is described, followed by an overview of library resources in Germany that are of specific interest to climate scientists. As has been described in the previous sections, climate researchers operate in the given infrastructure their institutions, data repositories and data centres are embedded in. In order to make the outcome of their work known, to collaborate with colleagues and to advance their careers, scientists publish write ups of their research results and also, increasingly, data that their work is based on or that may be a useful starting point for further research by others. Some aspects concerning journals in which climate researchers frequently publish results of their work and some information on Open Access are given in section 5.3

5.1 Library management in Germany

Libraries as part of "Education, Culture and Science" are subject to regulations at the state level (*Länder*), and library legislation is a political statement in concrete terms that a state intends to guide, configure, cultivate and fund libraries. However, there exists no law in Germany in any of its 16 states that makes the operation and maintenance of public libraries mandatory. Discussions have been going on for over 50 years how this can be achieved, and in recent years the German Library Association (DeutscherBibliotheksverband DBV) has designed a "sample national library law",[171] which the states in Germany's federal system are encouraged to use, adjust and incorporate into their *Länder* laws. The DBV was inspired by the best practice examples observed in Denmark, Finland and Great Britain which all include mandatory library services, highly topical holdings taking new media and information technology developments into account, free-of-charge usage by anybody, sufficient funding by the municipality, financial provision for infrastructures and networks by the government and integration of the libraries into educational concepts

This lack of a firm legislative backbone notwithstanding, library services are supported financially by the DFG in its "Nationwide Library Services and National Licenses" programme to "facilitate the provision of a comprehensive range of highly specialised literature collections and digital sources of information for use in scientific research in Germany",[172] with the aim to allow access to specialised scientific information that cannot be had at/through

[171] http://www.bibliotheksportal.de/bibliotheken/bibliotheken-in-deutschland/bibliotheksgesetz.html.

[172] http://www.dfg.de/en/research_funding/programmes/infrastructure/lis/digital_information/library_licenses/index.html.

individual university libraries. All scientists and academics in Germany may use "internet-based services for bibliographical research, interlibrary loans, document delivery" and access digital collections directly online.

The DFG also offers funding opportunities for the acquisition of national licences for nationwide access to literature in digital form. All members of universities, (technical) colleges and research institutions located in Germany which have secured a national licence for certain publications may access these free of charge from the campus networks and the catalogues of German state and university libraries, while others may register for free individual use of many databases and text collections at www.nationallizenzen.de.

5.2 Libraries, literature databases and search tools

The DFG maintains a system of literature supply called "special subject collections"[173] (SSG, Sondersammelgebiete). Besides the libraries listed there, other specialist libraries allow scientists to find literature from the field of geosciences and specifically about climate topics.

5.2.1 Library of the Center for Marine and Atmospheric Sciences (ZMAW), Hamburg

The ZMAW[174] is a cooperative centre of several institutes of the University of Hamburg and the Max Planck Institute for Meteorology promoting research in the fields of marine, climate and earth system science in Hamburg. The Institute of Coastal Research of the Helmholtz Centre Geesthacht (HZG) has been an associated member of the ZMAW since 2005.

The ZMAW Library[175] is a special library for the earth sciences. It is also a member of the German Association of Marine Science Libraries and Information Centers (GAMSLIC), whose catalogue[176] contains the holdings of the libraries of 12 other institutions:

- Alfred Wegener Institute Foundation for Polar and Marine Research,[177] Bremerhaven (including the library of the former Biologische Anstalt Helgoland)

[173] http://dispatch.opac.ddb.de/DB=1.1/LNG=EN/SID=a4a010ab-17/SSG.

[174] http://www.zmaw.de/The-ZMAW.4.0.html?&L=1.

[175] http://www.zmaw.de/index.php?id=5&L=1.

[176] http://gso.gbv.de/DB=2.910/LNG=EN/?COOKIE=U999,K999,D2.910,E874afaef-192,I0,B9994++++++,SY,A\delimiter"026E30F9008+*,H13-15,,17-23,,30,,73-78,,88-90,NGAST,R136.172.96.229,FN&UCLOAD=Y&COOKIE=U999,K999,D2.910,E874afaef-192,I0,B9994++++++,SY,A\delimiter"026E30F9008+*,H13-15,,17-23,,30,,50,,60-61,,73-78,,88-90,NGAST,R136.172.122.54,FN.

[177] http://www.awi.de/en/infrastructure/library.

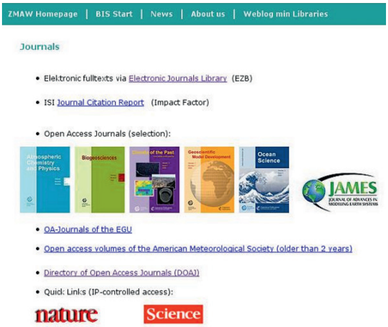
- Bundesamt für Seeschifffahrt und Hydrographie,[178] Hamburg
- Forschungsanstalt der Bundeswehr für Wasserschall und Geophysik,[179] Kiel
- German Maritime Museum,[180] Bremerhaven
- HZG: Institute of Coastal Research[181] (journals only), Geesthacht
- Johann Heinrich von Thünen-Institute, Federal Research Institute for Rural Areas, Forestry and Fisheries,[182] Hamburg
- Leibniz Institute for Baltic Sea Research,[183] Warnemünde
- Leibniz Institute of Marine Sciences[184] (IFM-GEOMAR), Kiel
- Leibniz Center for Tropical Marine Ecology,[185] Bremen
- Max Planck Institute for Evolutionary Biology,[186] Plön
- Deutsches Meeresmuseum,[187] Stralsund
- Terramare Research Centre,[188] Wilhelmshaven.

While these libraries have a marine focus, the ZMAW's Library and Information Service (LIS) has its roots in the Department of Earth Sciences of the University of Hamburg and the Max Planck-Institute for Meteorology and therefore has a special focus on these research fields. The LIS on its website[189] provides entry points to catalogues, databases, journals, the LIS service and ZMAW publications.

A search often starts with a click on “Journals”, resulting in the display shown in Figure F.16.[190] Some Open Access journal groups are indicated directly, while many more may be found in the Electronics Journal Library[191] (Elektronische Zeitschriftenbibliothek; EZB).

The EZB provides information about electronic journals (not the article texts themselves). Partner institutions of the EZB are University Libraries in Germany, institutes belonging to the Max Planck Society or the Helmholtz Association, State and Regional Libraries, some university libraries in Austria and Switzerland, and other institutions. Researchers at these institutions have free full-text access to those journals/articles that are marked green

[178] http://www.bsh.de/en/The_BSH/Organisation/Library/index.jsp.
[179] http://www.fwg-kiel.de.
[180] http://www.dsm.museum/bibliothek.33.de.html.
[181] http://www.hzg.de/central_departments/library/index.html.en.
[182] http://vzopc4.gbv.de:8080/DB=19.2/LNG=DU.
[183] http://www.io-warnemuende.de/library-and-it-group.html.
[184] http://www.ifm-geomar.de/index.php?id=bibliothek_home&L=1.
[185] http://www.zmt-bremen.de/en/Library.html.
[186] http://www.evolbio.mpg.de/english/bibliothek/index.html.
[187] http://www.meeresmuseum.de/wissenschaft/bibliothek.html.
[188] http://www.icbm.de/32114.html.
[189] http://www.zmaw.de/index.php?id=5&L=1.
[190] http://www.zmaw.de/Journals.46.0.html?&L=1.
[191] http://rzblx1.uni-regensburg.de/ezeit/index.phtml?bibid=DM&colors=7&lang=en.


ZMAW Homepage | BIS Start | News | About us | Weblog min Libraries
Atmospheric Chemistry and Physics
Biogeosciences
Climate of the Past
Geoscientific Model Development
Ocean Science
JAMES
Journal of Advances in Modeling Earth Systems
nature
Science

Journals

- Elektronic fulltexts via Electronic Journals Library (EZB)
- ISI Journal Citation Report (Impact Factor)
- Open Access Journals (selection):
- OA-Journals of the EGU
- Open access volumes of the American Meteorological Society (older than 2 years)
- Directory of Open Access Journals (DOAJ)
- Quick Links (IP-controlled access):

Figure F.16 The "Journals" web page of ZMAW's Library and Information Service

(indication of complete Open Access) or yellow (indication that a licence fee has been paid) at the URL that is produced when doing an alphabetic journal title search on the EZB page, for example "Journal of Geophysical Research".[192]

The EZB, which used to be funded by the BMBF, the DFG and the Bavarian Ministry of Science, Research and Art, provides links to journals, newspapers and databases in German and Austrian libraries and is a cooperative effort of 4300 libraries. The EZB is one of the largest databases worldwide for finding journals, newspapers, reports and other periodicals from any country and language in electronic format.

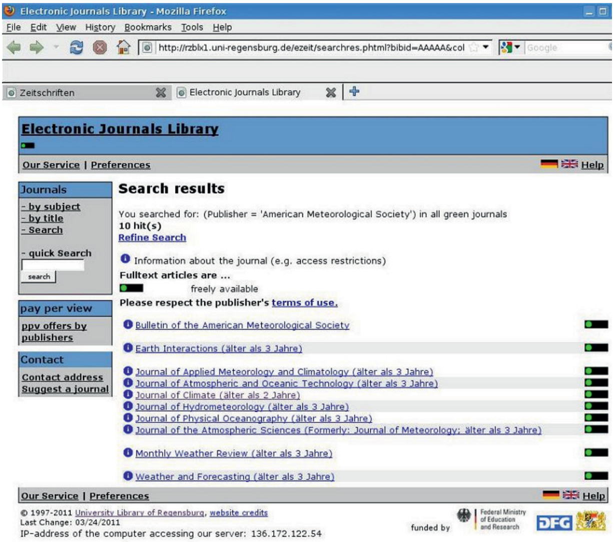
The result of a search for all journals that are available online from the American Meteorological Society, for example, is shown in Figure F.17.[193] All titles shown are Open Access, some with restrictions as to the actuality, i.e. only older issues are fully Open Access. A search in the EZB by subject "geosciences" lists journal titles in alphabetical order, as shown in Figure F.18.

For each journal, the access mode is indicated by the (coloured) dots along the right hand margin: Green (OXX): full text is freely accessible (Open Access); Yellow (XOX): full text can be accessed within the Campus-Net and for university members also from outside the campus; Yellow/red (XOO): full text access only for parts of all published issues; Red (XXO: no free access

[192] http://rzblx1.uni-regensburg.de/ezeit/fl.phtml?bibid=MPIM&colors=7&lang=de&notation=ALL&sc=J&lc=K&sindex=3650#jumpto.

[193] http://rzblx1.uni-regensburg.de/ezeit/searchres.phtml?bibid=MPIM&colors=7&lang=de&jq_type1=KT&jq_term1=&jq_bool2=AND&jq_not2=+&jq_type2=KS&jq_term2=&jq_bool3=AND&jq_not3=+&jq_type3=PU&jq_term3=american+meteorological+society&jq_bool4=AND&%20jq_not4=+&jq_type4=IS&jq_term4=&offset=1&hits_per_page=50&search_journal=Suche*starten&Notations[]=all&selected_colors[]=1&selected_colors[]=2&selected_colors[]=4.

possible for the location. Sometimes access is possible to abstracts or tables of content.


Electronic Journals Library - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://rzblx1.uni-regensburg.de/ezeit/searchres.phtml?bibid=AAAAA&col
Google
Zeitschriften
Electronic Journals Library

Electronic Journals Library

Our Service | Preferences — Help

Journals
- by subject
- by title
- Search

- quick Search
search

pay per view
ppv offers by publishers

Contact
Contact address
Suggest a journal

Search results

You searched for: (Publisher = 'American Meteorological Society') in all green journals
10 hit(s)
Refine Search

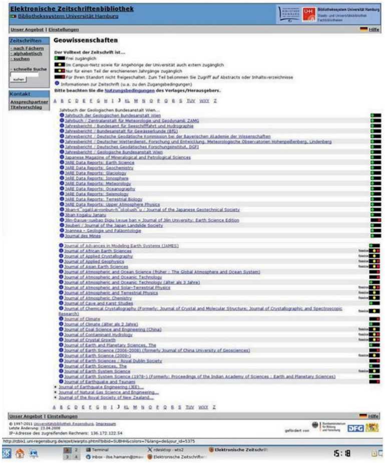
Information about the journal (e.g. access restrictions)
Fulltext articles are ...
🟩 freely available
Please respect the publisher's terms of use.

- Bulletin of the American Meteorological Society
- Earth Interactions (älter als 3 Jahre)
- Journal of Applied Meteorology and Climatology (älter als 3 Jahre)
- Journal of Atmospheric and Oceanic Technology (älter als 3 Jahre)
- Journal of Climate (älter als 2 Jahre)
- Journal of Hydrometeorology (älter als 3 Jahre)
- Journal of Physical Oceanography (älter als 3 Jahre)
- Journal of the Atmospheric Sciences (Formerly: Journal of Meteorology; älter als 3 Jahre)
- Monthly Weather Review (älter als 3 Jahre)
- Weather and Forecasting (älter als 3 Jahre)

Our Service | Preferences — Help

© 1997-2011 University Library of Regensburg, website credits
Last Change: 03/24/2011
IP-address of the computer accessing our server: 136.172.122.54
funded by Federal Ministry of Education and Research DFG

Figure F.17 Journals of the American Meteorological Society, available within the library system of the University of Hamburg



Elektronische Zeitschriftenbibliothek
Bibliothekssystem Universität Hamburg
Bibliothekssystem Universität Hamburg
Staats- und Universitätsbibliothek
Fachbibliotheken

Unser Angebot | Einstellungen — Hilfe

Zeitschriften
- nach Fächern
- alphabetisch
- suchen

- schnelle Suche
suchen

Kontakt
Ansprechpartner
Titelvorschlag

Geowissenschaften

Der Volltext der Zeitschrift ist...
🟩 Frei zugänglich
🟨 Im Campus-Netz sowie für Angehörige der Universität auch extern zugänglich
🟨🟥 Nur für einen Teil der erschienenen Jahrgänge zugänglich
🟥 Für Ihren Standort nicht freigeschaltet. Zum Teil bekommen Sie Zugriff auf Abstracts oder Inhaltsverzeichnisse
Informationen zur Zeitschrift (u.a. zu den Zugangsbedingungen)
Bitte beachten Sie die Nutzungsbedingungen des Verlages/Herausgebers.

A B C D E F G H I J KL M N O P Q R S TUV WXY Z

Jahrbuch der Geologischen Bundesanstalt Wien...
- Jahrbuch der Geologischen Bundesanstalt Wien
- Jahrbuch / Zentralanstalt für Meteorologie und Geodynamik: ZAMG
- Jahresbericht / Bundesamt für Seeschifffahrt und Hydrographie
- Jahresbericht / Bundesanstalt für Gewässerkunde (BfG)
- Jahresbericht / Deutsche Geodätische Kommission bei der Bayerischen Akademie der Wissenschaften
- Jahresbericht / Deutscher Wetterdienst, Forschung und Entwicklung, Meteorologische Observatorien Hohenpeißenberg, Lindenberg
- Jahresbericht / Deutsches Geodätisches Forschungsinstitut, DGFI
- Jahresbericht / Geologische Bundesanstalt Wien
- Japanese Magazine of Mineralogical and Petrological Sciences
- JARE Data Reports: Earth Science
- JARE Data Reports: Geochemistry
- JARE Data Reports: Glaciology
- JARE Data Reports: Ionosphere
- JARE Data Reports: Meteorology
- JARE Data Reports: Oceanography
- JARE Data Reports: Seismology
- JARE Data Reports: Terrestrial Biology
- JARE Data Reports: Upper Atmosphere Physics
- Jiban-k¯ogaku-ronbun-h¯okokush¯u / Journal of the Japanese Geotechnical Society
- Jiban Kogaku Janaru
- Jilin-Daxue-xuebao Diqiu kexue ban = Journal of Jilin University: Earth Science Edition
- Jisuberi / Journal of the Japan Landslide Society
- Joannea - Geologie und Paläontologie
- Journal des Mines

- Journal of Advances in Modeling Earth Systems (JAMES)
- Journal of African Earth Sciences
- Journal of Applied Crystallography
- Journal of Applied Geophysics
- Journal of Asian Earth Sciences
- Journal of Atmospheric and Ocean Science (früher : The Global Atmosphere and Ocean System)
- Journal of Atmospheric and Oceanic Technology
- Journal of Atmospheric and Oceanic Technology (älter als 3 Jahre)
- Journal of Atmospheric and Solar-Terrestrial Physics
- Journal of Atmospheric and Terrestrial Physics
- Journal of Atmospheric Chemistry
- Journal of Cave and Karst Studies
- Journal of Chemical Crystallography (Formerly: Journal of Crystal and Molecular Structure; Journal of Crystallographic and Spectroscopic Research)
- Journal of Climate
- Journal of Climate (älter als 2 Jahre)
- Journal of Coal Science and Engineering (China)
- Journal of Contaminant Hydrology
- Journal of Crystal Growth
- Journal of Earth and Planetary Sciences, The
- Journal of Earth Science (2006-2008) (formerly Journal of China University of Geosciences)
- Journal of Earth Science (2009-)
- Journal of Earth Sciences / Royal Dublin Society
- Journal of Earth Sciences, The
- Journal of Earth System Science
- Journal of Earth System Science (1978-) (Formerly: Proceedings of the Indian Academy of Sciences : Earth and Planetary Sciences)
- Journal of Earthquake and Tsunami
- Journal of Earthquake Engineering (JEE)...
- Journal of Natural Gas Science and Engineering...
- Journal of the Royal Society of New Zealand...

A B C D E F G H I J KL M N O P Q R S TUV WXY Z

Unser Angebot | Einstellungen — Hilfe

© 1997-2011 Universitätsbibliothek Regensburg, Impressum
Letzte Änderung: 23.04.2008
IP-Adresse des zugreifenden Rechners: 136.172.122.54
gefördert von Bundesministerium für Bildung und Forschung DFG

http://rzblx1.uni-regensburg.de/ezeit/warpto.phtml?bibid=SUBHH&colors=7&lang=de&jour_id=5375

Figure F.18 EZB listing of journal titles in the field of "geoscience" with access information for University of Hamburg users

5.2.2 National Meteorological Library hosted by the German Weather Service[194]

The stock of this library reaches back to the 15th century, documenting the development of meteorology as an independent science from its origins up to now. This library holds 180,000 volumes and approximately 800 current journal and periodical titles. It is the official special collecting library for meteorology, climatology and meteorological maps and charts.

The library offers various search methods through the Meteorological Literature Information System (METLIS),[195] which draws from entries about practically all specialist publications since the introduction of electronic data processing,[196] and referring also to external databases such as the internationally recognised Meteorological and Geoastrophysical Abstracts (MGA).[197]

5.2.3 Library at the Federal Maritime and Hydrographic Agency, Hamburg and Rostock[198]

This library is the central maritime library in Germany. The initial stock dates back to 1868 and began with nautical charts and books. The collection grew steadily, but in World War II parts of it were irretrievably lost. After the German reunion the library collections of the BSH and the Seehydrographischer Dienst (SHD) of the German Democratic Republic (GDR) were combined. The collection includes approximately 170,000 media and 50,000 nautical charts, openly accessible by all scientists, researchers, historians, students and the general public. The holdings cover the subjects interesting for climate scientists like oceanography (excluding marine biology), marine physics, marine chemistry, marine geophysics, marine geology, marine meteorology, marine environmental protection and various other nautical subjects.

5.3 Climate science literature management

In sections 5.1 and 5.2, the infrastructure of literature resources that researchers have at their disposal in Germany was described. In this section,

[194] http://www.dwd.de/bvbw/appmanager/bvbw/dwdwwwDesktop?_nfpb=true&_windowLabel=dwdwww_main_book&T3420224081166532168092gsbDocumentPath=&switchLang=en&_pageLabel=dwdwww_menu2_bibliothek.

[195] http://oflsd45.dwd.de:8060/alipac/LGDONSVLIKSRQCMRGYBQ-00001/form/find-simple.

[196] http://www.dwd.de/bvbw/generator/DWDWWW/Content/Oeffentlichkeit/PB/PBFB/Bibliothek/Allgemein/en_Bibliotheksflyer,templateId=raw,property=publicationFile.pdf/en_Bibliotheksflyer.pdf.

[197] http://www.csa.com/factsheets/mga-set-c.php.

[198] http://www.bsh.de/en/The_BSH/Organisation/Library/index.jsp.

some aspects of literature management from the viewpoint of the scientists are discussed.

5.3.1 Journal Citation Report (JCR, ISI) and Open Access in climate science journals

From the survey of researchers (section 4), journal titles emerged which were preferred for publishing scientific results. I used the Journal Citation Report[199] (JCR, ISI) to determine the most frequently cited journals in the field of climate science[200] and to see how many of those are Open Access, i.e. are also listed in the Directory of Open Access Journals[201] (DOAJ).

In the JCR (ISI) subject category "Meteorology & Atmospherics Sciences", an impact factor[202] (and other journal metrical quantities) is given for 68 journals. The journal with the highest impact factor, "Atmospheric Chemistry and Physics" is, in fact, also an Open Access title. The other four Open Access journals have satisfactory impact factors because the median impact factor is 1.6.[203]

In the DOAJ, one finds journals by "browsing by subject".[204] For example, in subject "Earth and Environmental Sciences" 26 journals are listed in the subcategory "Meteorology and Climatology"[205]. For five of these 26 journal titles, journal metrics are given in the JCR (ISI) subject category "Meteorology & Atmospheric Sciences".[206] This means that 19% of the relevant journal titles in this subject area are Open Access. Their titles, ISSN and impact factor (IF) are:

- *Atmospheric Chemistry and Physics:* ISSN 1680-7316, IF 5.5
- *Atmospheric Measurement Techniques:* ISSN 1867-1381, IF 2.6
- *Natural Hazards and Earth System Sciences:* ISSN 1561-8633, IF 1.8
- *Journal of the Meteorological Society of Japan:* ISSN 0026-1165, IF 1.1
- *SOLA: Scientific Online Letters on the Atmosphere:* ISSN 1349-6476, IF 1.0.

[199] http://admin-apps.webofknowledge.com/JCR/JCR?SID=U2699KNOGA%406h4HepFl.

[200] http://admin-apps.webofknowledge.com/JCR/help/h_jcrabout.htm.

[201] http://www.doaj.org.

[202] The annual Journal Citation Reports impact factor is a ratio between citations and recent citable items published: a journal's impact factor is calculated by dividing the number of current year citations to the source items published in that journal during the previous two years. http://thomsonreuters.com/products_services/science/academic/impact_factor.

[203] http://admin-apps.webofknowledge.com/JCR/JCR?RQ=LIST_SUMMARY_CATEGORY&category_sort_by=cat_title&cursor=1.

[204] http://www.doaj.org/doaj?func=subject&cpid=78&uiLanguage=en.

[205] http://www.doaj.org/doaj?func=subject&cpid=86&uiLanguage=en.

[206] http://admin-apps.webofknowledge.com/JCR/JCR?RQ=LIST_SUMMARY_JOURNAL&cursor=1.

5.3.2 Open Access mandates of science organisations

In the survey (section 4), the interviewed scientists described how they as authors of papers and data sets organise the publication of the results of their research. There are established rules, either through the agencies funding projects or the institutional policies where climate science produces results.

Via the OpenAIRE Portal, one finds descriptions of the two policies presently in place for Open Access in Europe,[207] i.e. mandates for scientists funded through projects of the European Research Council (ERC)[208] and the Seventh Framework Programme of the European Commission (FP7),[209] and what, where and when should be deposited.

The main research organisations supporting climate science in their institutes provide information on their Open Access policies and requirements on special websites, e.g. the Max Planck Society[210] Helmholtz Association[211] and Fraunhofer-Gesellschaft.[212] The Leibniz Association's Open Access Working Group has been developing a basic position on Open Access since 2005[213] and the Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz (i.e. the Leibniz Association) maintains a distributed Open Access repository[214] More general information and Open Access implementation advice is available through the Open Access information platform[215]

5.3.3 Tools used for literature management

Scientists use a variety of tools that support their literature management und publication in practice. One is Zotero,[216] a plug in to the Firefox web browser as a free tool that is used to collect, organise, cite and share their research resources. Another one is JabRef,[217] an open source bibliography reference manager using the file format BibTeX, a standard LaTeX bibliography format, or other bibliographic software. A widely used tool for literature searches and sorting the discovered information is Elsevier's platform, SciVerse,[218]

[207] http://www.openaire.eu/en/open-access/open-access-in-fp7.
[208] http://www.openaire.eu/en/component/attachments/download/3.
[209] http://www.openaire.eu/en/component/attachments/download/4.html.
[210] http://oa.mpg.de.
[211] http://oa.helmholtz.de/index.php?id=137.
[212] http://www.fraunhofer.de/content/dam/zv/de/publikationen/Fraunhofer_OpenAccessPolicy.pdf.
[213] http://www.wgl.de/?nid=akroa.
[214] http://www.tib-hannover.de/en/the-tib/wgl-repository.
[215] http://open-access.net/de_en/homepage.
[216] http://www.zotero.org.
[217] http://jabref.sourceforge.net.
[218] http://www.info.sciverse.com/UserFiles/resource_library_brochures/sciverse-brochure.pdf.

which allows access to the ScienceDirect database of peer-reviewed articles and books and to the Scopus citation database, among other applications.

5.3.4 Preferred journals for publication in climate research

From the survey of climate scientists, the following journals[219] emerged as the main publications where the research results are being published. Most of them charge a (often moderate) publication fee and offer various Open Access options.

1. *Hydrology and Earth System Sciences*: ISSN 1027-5606, IF 2.5), interactive Open Access journal of the European Geosciences Union, Open Access, public peer-review and interactive public discussion, personalised copyright under a Creative Commons Licence, moderate service charges.[220,221]
2. *Remote Sensing of the Environment:* ISSN 0034-4257, IF 3.95, interdisciplinary journal of Elsevier. Open Access journal offering authors the option of making their article freely available to all via the ScienceDirect platform. The fee of $3,000 excludes taxes and other potential author fees such as colour charges.[222,223]
3. *Journal of Climate:* ISSN 0894-8755, IF 3.5, online journal of the American Meteorological Society (at the Library and Information Service of the ZMAW in Hamburg there is free access to full text for issues older than 2–3 years).[224,225]
4. *Remote Sensing:* ISSN 2072-4292, online journal of Yale's Center for Earth Observation, Open Access journal, is published by the Multidisciplinary Digital Publishing Institute (MDPI) online monthly. Free for readers, with low publishing fees paid by authors or their institutions. Rapid publication: accepted papers are immediately published online.

[219] Where available from the 2010 JCR Science Edition, the ISSN and (impact factor) are also given.

[220] http://www.hydrology-and-earth-system-sciences.net/home.html.

[221] http://www.hydrology-and-earth-system-sciences.net/submission/service_charges.html.

[222] http://www.elsevier.com/wps/find/journaldescription.cws_home/505733/description#description.

[223] http://www.elsevier.com/wps/find/journaldescription.cws_home/505733/authorinstructions.

[224] http://journals.ametsoc.org/toc/clim/current.

[225] http://rzblx1.uni-regensburg.de/ezeit/searchres.phtml?bibid=AAAAA&colors=1&lang=en&jq_type1=KT&jq_term1=&jq_bool2=AND&jq_not2=+&jq_type2=KS&jq_term2=&jq_bool3=AND&jq_not3=+&jq_type3=PU&jq_term3=American+Meteorological+Society&jq_bool4=AND&jq_not4=+&jq_type4=IS&jq_term4=&offset=-1&hits_per_page=50&search_journal=Start+search&Notations[]=all&selected_colors[]=1.

MDPI is a publisher of peer-reviewed, Open Access journals since its establishment in 1996.[226,227]

5. *Biogeosciences:* ISSN 1726-4170, IF 3.6, interactive Open Access journal of the European Geosciences Union. Public peer review and interactive public discussion, personalised copyright under a Creative Commons Licence.[228]
6. *Journal of Geophysical Research (Atmospheres):* ISSN 0148-0227, IF 3.3, journal of the American Geophysical Union.[229]
7. *Geophysical Research Letters:* ISSN 0094-8276, IF 3.5, journal of the American Geophysical Union.[230]
8. *Journal of Hydrology:* ISSN 0022-1694, IF 2.5, journal of Elsevier, with Open Access solutions.[231,232]
9. *Global Environmental Change:* journal by Elsevier, with Open Access solutions.[233,234]
10. *Comptes Rendus Geoscience:* ISSN 1631-0713, IF 1.7, journal by Elsevier.[235]
11. *IEEE Geoscience and Remote Sensing:* ISSN 0196-2892, publication(s) of the IEEE Geoscience and Remote Sensing Society.[236]
12. *ISPRS Journal of Photogrammetry and Remote Sensing:* ISSN 0924-2716, IF 2.2, journal of Elsevier.[237]
13. *Earth-Science Reviews:* ISSN 0012-8252, IF 5.8, journal of Elsevier.[238]

More journals relevant for climate research can be found in a Geographic Information System and Remote Sensing journal list which has been compiled by Prof Giorgos Mountrakis of the State University of New York, College of Environmental Science and Forestry.[239]

[226] http://www.mdpi.com/journal/remotesensing.
[227] http://www.mdpi.com/about.
[228] http://www.biogeosciences.net.
[229] http://www.agu.org/journals/jd.
[230] http://www.agu.org/journals/gl.
[231] http://www.elsevier.com/wps/find/journaldescription.cws_home/503343/description#description.
[232] http://www.elsevier.com/wps/find/authorsview.authors/openaccess.
[233] http://www.elsevier.com/wps/find/journaldescription.cws_home/30425/description#description.
[234] http://www.sciencedirect.com/science/journal/09593780.
[235] http://www.sciencedirect.com/science/journal/16310713.
[236] http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=36.
[237] http://www.elsevier.com/wps/find/journaldescription.cws_home/503340/description#description.
[238] http://www.elsevier.com/wps/find/journaldescription.cws_home/503329/description#description.
[239] http://www.aboutgis.com/gis-and-remote-sensing-journal-list-with-impact-factors.

6 Current status of Open Access in climate research data

The different data collections that climate research produces or uses were the subject of section 3. In the following sections, additional information is given regarding the access to the data held in those centres, including policies and tools for data retrieval and sharing.

A large step forward in the development of climate data infrastructure was the creation of the World Data Center System[240] which came into existence as a result of the International Geophysical Year (IGY) of 1957/58. The purpose of World Data Centers was to ensure that observational data from the IGY programme would be readily available to scientists of all countries. The arrangement became permanent under the auspices of the International Council of Sciences (ICSU) and has remained so. Today the WDC system includes 52 centres in 12 different countries, 15 of which operate in the USA, 7 in Russia, 11 in Europe, 10 in Australia, India and Japan and 9 in China.[241] WDCs are funded and maintained by their host countries. Details regarding rules, responsibilities, data acquisition and usage of the WDC system are described in the World Data Center System guide.[242]

As a major funding agency of climate research in Germany, the DFG, more exactly its Committee on Scientific Library Services and Information Systems, Subcommittee on Information Management, in 2009 published seven recommendations for secure storage and availability of digital primary research data.[243] In the field of climate science these recommendations have already been successfully implemented to a large extent.

6.1 German Weather Service (DWD Deutscher Wetterdienst), Offenbach[244]

This is the DWD's contact point for the provision of data and products (e.g. numerical products or radar data) to special users, i.e. its clients are typically meteorological service providers, universities and research institutions, and authorities of the German Federal Government or the *Länder* (16 German states), who may receive real-time data and products as well as archived data (called "climate data" by the DWD) and products.

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

[241] http://www.icsu-wds.org/wds-members/wds-members.

[242] http://www.wdc.rl.ac.uk/wdc/guide/wdcguide.html.

[243] http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901_en.pdf.

[244] http://www.dwd.de.

A portion of the data are offered for "free", that means free of charge and mostly without any restrictions for the use of these data (where applicable, restrictions of use are specified next to the data set concerned). The list of data that are available online includes climate data for Germany, climate data from satellites, climate data worldwide and precipitation data worldwide. The broad spectrum of data is displayed on the "Climate and Environment" web page of DWD[245] (note: not all descriptions are available on the English version of this page).

Various web applications of the DWD contain extensive information about the climate data and available (metadata), which facilitate access to the data sets. One such application is WebWerdis (Web-based Weather Request and Distribution System) which provides access to free online climate data without the necessity of registration. WebWerdis is aimed at users with some prior subject knowledge, in particular from research and educational institutions and public authorities. Proper registration with WebWerdis, however, allows full access to all contents and many usage options.

Additional metadata are contained in the Climate Catalogue of the DWD's Climate Data Center (CDC). This data catalogue is currently still being extended and will in the end contain nationally and internationally standardised metadata (ISO19115[246] and ISO19139[247]) for all data available at the DWD. Direct access will be possible in many cases.

The Global Data Set (GDS) comprises freely available data and products relating to current weather and weather forecasts as well as freely available climate data. The data and products are usually presented in the same form that is used for their exchange within and between the national meteorological services. Access is possible via ftp. The use of the GDS is free of charge but requires registration. To obtain any of a large list of other data, products, and services, DWD asks potential customers to contact staff directly.

For special data collections in some international data centres hosted by DWD, see also Table F.2 in section 3.1.

[245] http://www.dwd.de/bvbw/appmanager/bvbw/dwdwwwDesktop?_nfpb=true&_windowLabel=dwdwww_main_book&T82002gsbDocumentPath=Navigation%2FOeffentlichkeit%2FKlima__Umwelt%2FKlimadaten%2Fkldaten__kostenfrei%2FAbrufsysteme__Daten__node.html%3F__nnn%3Dtrue&switchLang=de&_pageLabel=_dwdwww_klima_umwelt_klimadaten_deutschland.

[246] ISO 19115:2003 defines the schema required for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data. http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020.

[247] ISO/TS 19139:2007 defines Geographic MetaData XML (gmd) encoding, an XML Schema implementation derived from ISO 19115, http://www.iso.org/iso/catalogue_detail.htm?csnumber=32557.

6.2 World Data Center for Remote Sensing of the Atmosphere (WDC-RSAT), Oberpfaffenhofen[248]

Data and products available through WDC-RSAT cover many subjects and span a wide range of processing levels. Direct access online is provided to atmosphere-related satellite-based data sets if they are either stored at the WDC-RSAT or found through the WDC-RSAT portal, i.e. when are safeguarded by other providers (for data and variable types compare with section 3.2). Furthermore a table of services regarding air quality forecasting and monitoring, Antarctic ozone hole monitoring, solar energy, virtual lab and sunburn time is on display.[249]

The policy of data use[250] requires that acknowledgement/reference is made to the ICSU World Data Center for Remote Sensing of the Atmosphere or to cite specific references where these are provided.

6.3 World Data Center for Marine Environmental Sciences (WDC-MARE), Bremen[251]

Data contained in WDC-MARE are discovered via the information system Publishing Network for Geoscientific and Environmental Data (PANGAEA), the search engine for which online guidance is provided.[252] PANGAEA operates as an Open Access library to archive, publish and distribute georeferenced data from earth system research. The operating institutions have committed themselves to make the content of WDC-MARE available for the long term.[253]

The majority of the data are freely available, usable under the terms of the licence expressed in the data set description. Some data sets are under moratorium from ongoing projects, but metadata are visible, as is the contact information of the principal investigator who may be asked for access.

Data sets at WDC-MARE can be identified, shared, published and cited via a DOI. Data may be archived as supplements to publications or as citable data collections. The German National Library of Science and Technology has a portal through which citations may be obtained.[254]

Data management and archiving at WDC-MARE follows the principles and responsibilities of ICSU World Data Centers[255] and the OECD principles

[248] http://wdc.dlr.de/data_products.

[249] http://wdc.dlr.de/data_products/SERVICES.

[250] http://wdc.dlr.de/data_products/data_use_policy.php.

[251] http://www.wdc-mare.org/data.

[252] http://www.wdc-mare.org/shared/help/help.php/search/index.html.

[253] http://www.pangaea.de/about.

[254] https://getinfo.de/app/?lang=en.

[255] http://www.wdc.rl.ac.uk/wdc/guide/gdsystema.html.

and guidelines for access to research data from public funding.[256] Authors submitting data to the PANGAEA data library for archiving agree that all data are provided under a Creative Commons Licence.[257]

For data set discovery, the data mining tool Advanced Retrieval Tool (ART)[258] and an internet mapper for georeferenced data may be used.[259] Documentation on the usage of these tools is available through a wiki.[260]

6.4 The Bremen Core Repository (BCR)[261]

The procedure for obtaining sediment core data/samples from any ocean drilling programme or deep sea drilling project is to fill out the Integrated Ocean Drilling Program (IODP) online sample request form[262] (for samples from any of the three repositories, see section 3.3). The same holds for samples from the IODP from expeditions with the non-riser vessel (JOIDES Resolution). More information on data access, policies, guidelines, procedures and obligations and a variety of sample request forms is provided on the corresponding IODP web page.[263] Samples can be either taken by repository staff, or by the scientists themselves. MARUM provides a checklist for planning a visit to the repository.[264]

6.5 National Oceanographic Datacentre for Germany (NODC), Hamburg[265]

The NODC of Germany is partner of the SeaDataNet[266] consortium of 35 countries forming a a unique virtual data management system network providing integrated or raw data sets of standardised quality online. This is achieved by the Common Data Index (CDI) service, that gives users detailed insight in the availability and geographical spread of marine data across the different data centres in Europe. The CDI provides an ISO19115 based index (metadatabase) to individual data sets – which may be samples, time series, profiles or trajectories – and it is the interface to online data access. Direct

[256] http://www.oecd.org/document/2/0,3746,en_2649_34293_38500791_1_1_1_1,00.html

[257] http://creativecommons.org/licenses.

[258] http://www.pangaea.de/advanced/ART.php.

[259] http://mapserver.pangaea.de.

[260] http://wiki.pangaea.de/wiki/Main_Page.

[261] http://www.marum.de/en/IODP_Core_Repository.html.

[262] http://www.iodp.tamu.edu/curation/samples.html.

[263] http://www.iodp.org/access-data.

[264] http://www.marum.de/en/Information_for_visitors_to_the_BCR.html.

[265] http://www.bsh.de/en/Marine_data/Observations/DOD_Data_Centre/index.jsp.

[266] http://www.seadatanet.org.

access to all NODC data is possible via this CDI. For submitting data requests and for downloading data, registration is required[267] (see Figure F.19). Such registration requests are managed per country by the specific National Oceanographic Data Centre/Marine Data Centre, which checks the authenticity of the user, and if ok, lets the central user register provide the user with access information.

In addition to offering this direct data access, the NODC of Germany contributes to SeaDataNet's five metadata catalogues and has the leading role for one of them, i.e. the Cruise Summary Reports Database. NODC therefore gathers all Cruise summary reports from the 42 SeaDataNet partners. It includes cruises from 1873 until today from more than 2000 different research vessels amounting to a total of more than 40000 cruises, in all European waters and global oceans. NODC is also integrated in the GeoSeaPortal[268] thereby implementing a national infrastructure for all marine data.

As was said in section 3.4, data may be accessed through the MUDAB web client,[269] for which online information in a brochure[270] and the MUDAB handbook[271] is offered (both texts are only available in German at present). When searching for data, users and customers of NODC are guided by a menu-based interface to the desired information. As shown in Figure F.20, starting points are either at the cruise, at the station or at the data level.

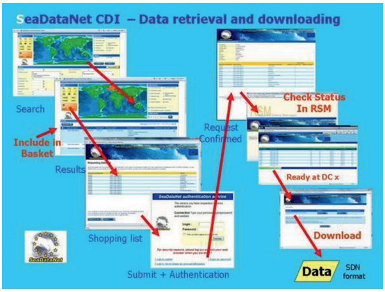


Figure F.19 SeaDataNet's Common Data Index: data retrieval and downloading (source: SeaDataNet newsletter no. 6, March 2011)[272]

[267] http://www.seadatanet.org/Data-Access/User-registration.

[268] http://www.bsh.de/de/Meeresdaten/Geodaten/index.jsp.

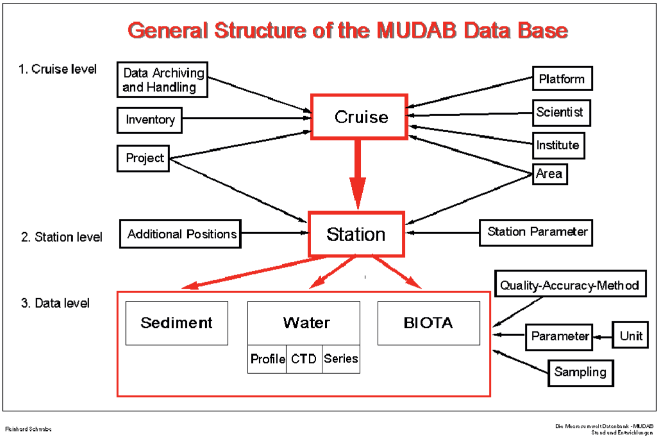
[269] http://www.mudab.de.

[270] http://www.informus.de:8080/mudab/documents/070530_mudab_webclient_faltblatt.pdf.

[271] http://www.informus.de:8080/mudab/documents/mudabws-usage.html.

[272] http://www.seadatanet.org/News/Seadatanet-Newsletter-n-6-March-2011.

[273] http://www.bsh.de/de/Meeresdaten/Umweltschutz/MUDAB-Datenbank/_1493.gif.



General Structure of the MUDAB Data Base

Reinhard Schwabe

Die Meeresumwelt Datenbank - MUDAB
Stand und Entwicklungen

Figure F.20 MUDAB database scheme (courtesy of Reinhard Schwabe, NODC)[273]

It is also possible to retrieve information by SQL. The criteria for database retrieval are dependents of space and time, measured variables and sampling and analysis methods. The retrieved information can be exported for further data processing in a format readable by many PC-based tools. In the near future, the data export facilities will be implemented either into standard formats or into an interface for biological data.

Whereas data requests to the NODC are generally free of costs, for dedicated products and other demands, a price list for digital data including some usage conditions (annexes 2 and 3) is available on the BSH website.[274]

The general terms of use for the MUDAB[275] are based on Open Access principles (commercial companies may be charged for retrieval costs):

- For all datasets, access is granted free of on condition that the user agrees:
  - data are for your scientific use, in particular it is not allowed to use them for any commercial purpose, and shall not be forwarded to third party without our notice;

[274] http://www.bsh.de/de/Produkte/Preise/Entgeltverzeichnis_digitale_Daten.pdf.

[275] http://www.informus.de:8080/mudab/termsofuse.faces.

- to acknowledge the source of the data in all publications and applications;
- to help improve the quality of the data by noting and reporting any errors or omissions discovered;
- to help improve the quality of the Data Service by giving feed back on functionalities and data packaging;
- to help improve the efficiency of our reporting by supplying us with documented digital copies of data and information derived from the data so that it can be re-used by the Agency with reference to the source;
- to supply us with a copy of/URL to all publications and other products based on the datasets.

6.6 World Data Center for Climate (WDCC), Hamburg[276]

WDCC collects, stores and disseminates data related to climate research. It restricts itself to climate data products. Emphasis is spent on climate modelling and related data products. DKRZ as the operating institution of WDCC has committed itself as running WDCC as a long-term archive.

Data can be addressed and cited using a DOI. Metadata collected for data sets are available for the public. Data itself are available under a Creative Commons Licence unless not stated otherwise. To download data, registration at WDCC is required. Most of the data is available for immediate download.

Figure F.21 shows the gateway to the CERA database[277] at WDCC through which data sets can be located and retrieved. Several search options are offered, e.g. finding data sets by the name of an experiment. Entering one of the acronyms of observational projects listed in Table F.3 in section 3.6, will immediately lead to a display of the size of the downloadable data set. However, as most of the data sets in CERA are from numerical experiments, one may first want to browse the experiments and select, for example, IPCC_AR4_ECHAM5/MPIOM,[278] and from the ensuing list further choose the numerical simulation EH5-T63L31_AMIP_1_MM.[279]

Before downloading any digital data, which often have quite large volumes, it is possible to display selected post-processed variables of the model output as shown in Figure F.22. There the variable "EH5_AMIP_1_MM_STP1000", i.e. monthly means of air temperature at 1000 hectopascal" was selected

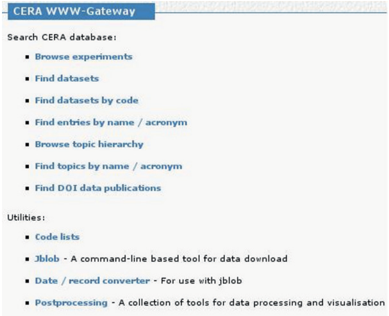
[276] http://www.dkrz.de/daten-en/wdcc.

[277] http://cera-www.dkrz.de/WDCC/ui/Index.jsp.

[278] Coupled numerical experiments carried out during CMIP3 with atmospheric model ECHAM5 and MPIOM of the Max Planck Institute for Meteorology, Hamburg.

[279] http://www-pcmdi.llnl.gov/projects/amip/NEWS/overview.php.

[280] http://cera-www.dkrz.de/WDCC/ui/Index.jsp



CERA WWW-Gateway

Search CERA database:

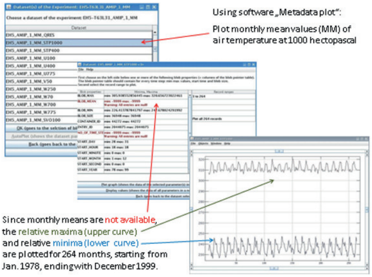
- Browse experiments
- Find datasets
- Find datasets by code
- Find entries by name / acronym
- Browse topic hierarchy
- Find topics by name / acronym
- Find DOI data publications

Utilities:

- Code lists
- Jblob - A command-line based tool for data download
- Date / record converter - For use with jblob
- Postprocessing - A collection of tools for data processing and visualisation

Figure F.21 The CERA WWW gateway[280]

and all 264 values for the runtime of the numerical simulation plotted (January 1978 through December 1999). Actually, since the monthly means of temperature had not been calculated and are therefore not contained in the database, the relative global maxima and minima were plotted instead (upper and lower curve, respectively).



Using software „Metadata plot":

Plot monthly mean values (MM) of air temperature at 1000 hectopascal

Since monthly means are not available, the relative maxima (upper curve) and relative minima (lower curve) are plotted for 264 months, starting from Jan. 1978, ending with December 1999.

Figure F.22 Plotting selected model output variables before data set download

Apart from general information about the project in the context of which this coupled climate model experiment was carried out, the following information is provided:

citation specification of the originator of the data, storage details of the downloadable output, the computing environment in which the data were produced, a list of related experiments, the model domain in three spatial dimensions, the temporal coverage, data formats and data set size, and whether/that download permission exists.

Access to CMIP5 data Much of section 2 has focused on the currently ongoing CMIP5 project that has been designed around Open Access principles. Ideally data will be released immediately after they have been quality-checked. The adoption of Open Access principles of the CMIP5 Federation is motivated by the highly risen interest in climate research results and data by a wide spectrum of users. These include members of the natural and social sciences that work on climate (change) impacts and options for adaptation who respond to increasing public pressure and demand for sustainable development.

According to PCMDI's website,[281] all model output in the CMIP5 archive is available for "non-commercial research and educational purposes", and a subset of the data has also been released for "unrestricted" use, i.e. users registering to access CMIP5 output will be granted access to some or all of the data, depending on which terms of use have been agreed on. For users seeking CMIP5 model output, a "getting started" tutorial is provided.[282]

According to information which we had received from PCMDI earlier this year and which was recompiled in Table F.1 (see section 2.3.2), five of the 21 modelling groups had announced that they would make their data available to all users, six groups would release their data for non-commercial use and ten groups had not (yet) specified by whom the data from their experiments may be obtained and used.

A check of the status of the CMIP5 archive[283] showed that data sets are now available from nine of the participating Centres, i.e. group numbers 1, 2, 4, 6, 12, 17, 18, 19 and 20 in Table F.1 (status on 31 August 2011). More details about available data sets can be seen in a CMIP5 wiki[284]

7 Challenges for Open Access e-Infrastructures in climate research

Concurring with *Science* staff writers, who point out in the introduction to *Science*'s Special Online Collection "Dealing with Data"[285] that "most scientific disciplines are finding the data deluge to be extremely challenging, and tremendous opportunities can be realised if we can better organise and access the data", the research that went into this climate science chapter has clearly shown this to be true. The organisational effort invested into planning the data management and assuring future re-usability in the CMIP projects was described in sections 2.3–2.7. Nonetheless, bottlenecks still exist in data

[281] http://cmip-pcmdi.llnl.gov/cmip5/terms.html.

[282] http://cmip-pcmdi.llnl.gov/cmip5/data_getting_started.html?submenuheader=3.

[283] http://cmip-pcmdi.llnl.gov/cmip5/availability.html?submenuheader=3.

[284] http://esgf.org/wiki/Cmip5Status/ArchiveView.

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

storage capacities, access and (re-)usage by users from a wider community, i.e. non-experts in climate research (or from the public), common metadata and securing sufficient funding to support archiving.

Scientists need to meet their responsibilities toward transparency, standardisation and data archiving because large integrated data sets can potentially provide a better understanding of nature and society, thereby allowing new approaches in research that address key societal problems like, for example, coping with climate change. The associated challenges are to:

1. equip data repositories with the necessary hard-, middle- and software to manage large data sets from climate and earth system research,
2. ensure blockage-free data flow through networks,
3. staff the data centres/repositories adequately to guarantee an effective re-usage of data.

The "wares" of the first challenge can be met by securing funding for appropriate shopping lists of corresponding products and technical support for them.

With regard to the second challenge, Kostas Glinos, head of the European Commission's GÉANT and e-Infrastructures Unit, highlighted some European e-Infrastructure mainstays[286] at a conference on the "Role of e-infrastructures for Climate Change Research"[287] in May 2011 at the Abdus Salam International Centre for Theoretical Physics (ICTP) in Trieste.[288] While access to and connectivity with a global community of data terminals via the internet is a given today, the switching of ever-increasing data streams across (academic) networks is a challenge that the GÉANT2 Joint Research Programme[289,290] has taken on (Figure F.23). The European Commission and the National Research and Education Networks (NRENs) are funding this first international hybrid research and education network.[291] While still active (until April 2010) the EU project EGEE[292] (Enabling Grids for E-sciencE) managed the GÉANT network. More information on GÉANT2 can be found at EUROPA, the gateway to the European Union[293]

[286] http://cdsagenda5.ictp.trieste.it/askArchive.php?subtalk=1&base=agenda&categ=a10141&id=a10141s2t9/slides.

[287] The main themes of the conference were climate change modelling and adaptation/mitigation policies and the role of e-Infrastructures in climate change studies. Pertinent challenges and ways forward both from the organisational and policy perspective and from the enabling technology vantage point were illustrated. Final programme version: http://users.ictp.it/~smr2238/Program_einfrastructures.pdf.

[288] http://users.ictp.it/~smr2238.

[289] http://www.geant2.net.

[290] http://www.geant2.net/upload/pdf/PUB-06-014_point_to_point_leaflet.pdf.

[291] http://www.geant2.net/upload/pdf/GN2_Topology_Feb_09.pdf.

[292] http://www.eu-egee.org.

[293] http://europa.eu/rapid/pressReleasesAction.do?reference=MEMO/08/133&format=HTML&aged=1&language=EN&guiLanguage=en.

The third challenge is an especially important one because the provision of tools for accessing, downloading, visualising and using data sets for new analyses requires communication and contact to data producers and/or data curators.

In an editorial in *Nature*, entitled "Data's shameful neglect",[294] the appeal is made to research funding agencies to "recognise that preservation of and access to digital data are central to their mission, and need to be supported accordingly". Data provenance, giving credit to "data contributors" also with respect to their career opportunities, and decisions on a "data library infrastructure" are critical issues that need to be taken care of and require a dedicated effort at the highest levels. To this end the editorial cites as a progressive example the establishment of the Joint Information Systems Committee by seven UK research councils in 1993, which made data-sharing a priority, and helped to establish a digital curation centre.

The commentary closes by recommending/demanding that "information management" should be a mandatory subject and that "data management should be woven into every course in science, as one of the foundations of knowledge".

Some optimistic correspondence[295] to this editorial appeal ensued, i.e. from members of the natural history research community, propagating the Mammal Networked Information System[296] (Guralnick, Constable, Wieczorek, Moritz and Peterson, 2009).

In Germany, the Alliance of German Science Organisations[297] has created the priority initiative "Digital Information"[298] in 2008 with the focal areas of:

- national licensing
- Open Access
- national hosting strategy
- primary research data
- virtual research environments
- legal frameworks.

[294] http://www.nature.com/nature/journal/v461/n7261/full/461145a.html.

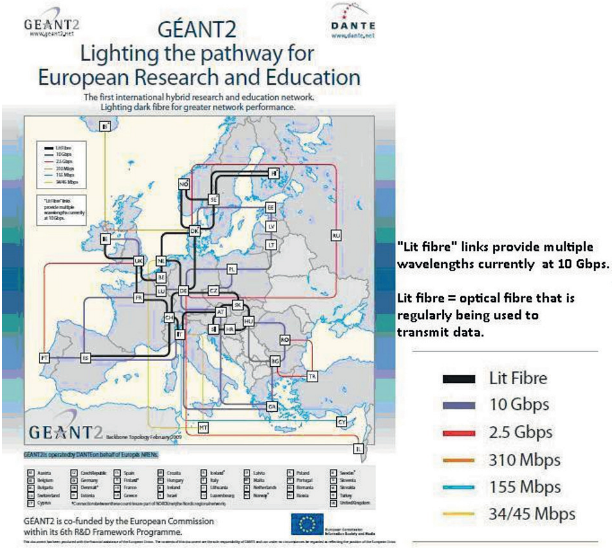
[295] http://www.nature.com/nature/journal/v462/n7269/full/462034a.html.

[296] http://manisnet.org.

[297] Members are: Alexander von Humboldt Foundation, German Academic Exchange Service, German Research Foundation, Fraunhofer Society, Helmholtz Association of German Research Centers, Association of Universities and other higher Education Institutions in Germany, Leibniz Association, Max Planck Society, Germany Science Council.

[298] http://www.wissenschaftsrat.de/download/archiv/Allianz-digitale%20Info_engl.pdf.

[299] http://www.geant2.net/upload/pdf/GN2_Topology_Feb_09.pdf.


GÉANT2
www.geant2.net
GÉANT2
Lighting the pathway for
European Research and Education
The first international hybrid research and education network.
Lighting dark fibre for greater network performance.
DANTE
www.dante.net
GÉANT2 Backbone Topology February 2009
GÉANT2 is co-funded by the European Commission within its 6th R&D Framework Programme.
Lit Fibre
10 Gbps
2.5 Gbps
310 Mbps
155 Mbps
34/45 Mbps

"Lit fibre" links provide multiple wavelengths currently at 10 Gbps.

Lit fibre = optical fibre that is regularly being used to transmit data.

Figure F.23 GÉANT2, the first international hybrid research and education network (modified from source)[299]

Kommission Zukunft der Informationsinfrastruktur The Leibniz Association was mandated in 2009 to develop a concept for the technical information infrastructure in Germany (delivered in April 2011). A commission "Future of information infrastructure" (*Zukunft der Informationsinfrastruktur KII*),[300] consisting of approximately 135 persons from 60 institutions contributed to the concept, which included the topics licensing, hosting and longterm archiving, non-textual material, retro-digitalisation/cultural heritage, virtual research environments, Open Access/electronic publishing, research data and information competency/education. They took the work of the above-mentioned priority initiative "Digital Information" into account.

Furthermore, the DFG on behalf of the Alliance of German Science Organisations also funded a special study on the establishment of a federated strategy on perpetual access and hosting electronic resources for Germany[301] At the European level the High level Expert Group on Scientific Data drew up their terms of reference in 2010.[302]

[300] http://www.wgl.de/?nid=kiikom&nidap=&print=0.

[301] http://www.allianzinitiative.de/fileadmin/hosting_studie_e.pdf.

[302] http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/tor.pdf.

8 Future directions and summary

The data that are being generated in climate or earth system science are already and will be re-used not only by scientists but should continue to be beneficially exploited by other societal groups. The scientific community, politicians and stakeholders concerned with climate change (and adaption to and mitigation of it) are to a varying degree already familiar with the existence of data sets useful for their work. There are mechanisms in place that let them tap successfully into the climate data resource.

In Hamburg, for example, the CSC,[303] specifically former members of the "Service Group Adaption" (SGA), assist participants of the project KLIMZUG[304] by providing them – in cooperation with the German Weather Service – with a common database of regional climate model and climate monitoring data and guidance concerning methods of climate data analysis.[305] Climate indices like precipitation days, snow, frost, ice, summer days, tropical nights, days with strong winds, length of the growing season, hot days, wet days, and diagrams and animations of simulated data are also available.[306] This information is not only valuable economically, but may also help to anticipate direct and indirect effects to human health.[307]

Before KLIMZUG, the BMBF-funded research about ways to deal with climate change in its programme "klimazwei".[308] Cooperative projects were carried out in which business and societal challenges and opportunities that exist due to climate change were addressed in transdisciplinary approaches. The need to manage both mitigation measures, e.g. reduction of greenhouse gas emissions, and adaptation strategies spawned a number of projects whose results were presented in a brochure[309] in 2009. Innovative ways to minimise CO2 and other greenhouse gas emissions in various industrial processing chains were derived and several possible solutions to the heat and water stresses brought about by climate change were suggested for several metropolitan areas. Particularly in the second category (adaptation) effective ways of communicating uncertainties and risks were central to some projects.

[303] http://www.climate-service-center.de.

[304] http://www.klimzug.de/en/211.php, KLIMZUG is presently funded by BMBF in the category "Managing climate change in the regions for the future".

[305] http://mud.dkrz.de/projects-at-md/sg-adaptation/sga/sga-introduction-english/index.html.

[306] http://www.klimazwei.de/Portals/0/SGA_flyer_engl_jul09.pdf.

[307] http://wiki.bildungsserver.de/klimawandel/index.php/Klimawandel_und_Gesundheit.

[308] SGA was created during BMBF programme "klimazwei – research for climate protection and protection from climate impacts", which started in 2006.

[309] http://www.klimazwei.de/Portals/0/klimazwei-Ergebnisbrosch%C3%BCre.pdf.

The CSC's priority is thus to provide decision takers from politics, business, administration and society with pertinent climate information that lets them do an improved service to their customers, which may be, for example, users from forestry, farmers, tourism managers and insurance companies. As a new instrument supporting the CSC's principal mission, the Climate Navigator[310] was launched in July 2011. The information that is accessible through this online platform is provided by more than 30 research organisations/units.

The CSC also contributes to the City of Hamburg's Educational Server (Hamburger Bildungsserver) with the maintenance of a climate wiki[311] which informs about general facts and processes concerning climate (change).

As the effects of climate change vary regionally, the Helmholtz Association has created four regional climate offices in Germany,[312] each focusing on a different region and having slightly different priorities: (1) the North German Climate Office,[313] at the Geesthacht Centre for Materials and Coastal Research, concentrates on changes in storms, storm surges, ocean waves, and coastal climate; (2) the Climate Office for Central Germany in Leipzig "offers information on adaptation strategies and on the impact of climate change on the environment, land use and society" (host: Centre for Environmental Research, UFZ);[314] (3) the South German Climate Office in Karlsruhe scores with its expertise on regional climate modelling and informs about extreme weather events such as heavy precipitation and flooding; and (4) the Climate Office for Polar Regions and Sea Level Rise, hosted by AWI in Bremerhaven, concentrates on relaying exactly this information to the public. All climate offices summarise their information in an online regional climate atlas.[315]

At the European level, ESA's Group on Earth Observations maintains the GeoPortal,[316] an entry point to access remote sensing, geospatial static and in-situ data, information and services. The Global Earth Observation System of Systems[317] (GEOSS) in particular promises to provide decision-support tools to a wide variety of users. Browsing for "climate"[318] shows the following services as being available: early warning, monitoring, analysing, mapping,

[310] http://www.hzg.de/science_and_industrie/klimaberatung/csc_web/012225/index_0012225.html.de, in German.

[311] http://www.klimawiki.org/, in German only.

[312] http://www.klimabuero.de/index_en.html.

[313] http://www.norddeutsches-klimabuero.de.

[314] http://www.mitteldeutsches-klimabuero.de/.

[315] http://www.regionaler-klimaatlas.de.

[316] http://www.geoportal.org/web/guest/geo_home?cache_control=0.

[317] http://www.earthobservations.org/geoss.shtml.

[318] http://www.geoportal.org/web/guest/geo_search_overview?p_p_id=srgPortlet_WAR_geoportal&p_p_lifecycle=0&p_p_state=normal&p_p_mode=view&p_p_col_id=column-1&p_p_col_count=4&_srgPortlet_WAR_geoportal_searchType=browse&_srgPortlet_WAR_geoportal_sbaId=4.

assessment, alerting, geospatial web service, data processing and data provision. The services are based on global and regional resources and are intended to help understand and evaluate key indicators of climate change.

A potential interest in climate data exists, however, also outside of the large earth system science network and in new communities recruiting themselves from as-yet-unknown areas of the general public. To raise awareness about the existence of and foster the re-use of high-quality earth system science data sets, the international and interdisciplinary journal *Earth System Science Data* (ESSD)[319] was established in 2008, and is published by Copernicus (Copernicus Publications). Another important objective of this endeavour is to reward "data authors" with the recognition of their peer-reviewed and appropriately described data set as an academic achievement. The chief editors of ESSD, David Carlson[320] and Hans Pfeiffenberger,[321] describe the peer-review process that they envision in their paper introducing this journal,[322] and distinguish between a-priori and a-posteriori quality assurance. The former may already exist in researcher communities that utilise established methods of documentation, validation and explicit quality control levels when generating or processing primary and secondary data. For the most part, this may be assumed for the data discussed in the previous sections of this chapter. Guidance for the review of and a-posteriori quality assessment of potentially valuable data resulting from less tested methods is provided by the journal.[323]

The evaluation of data sets based on a peer review like, for example, the one practiced by ESSD, represents an added value, as the recognition or impact of a data publication is a function of the quality control/assurance, since issuing a DOI for a data set does not render it a publication in a sense that is comparable to a usual (peer-reviewed) scientific publication.

Besides quality stamping data sets, data repositories as well may earn recognition for trustworthy data management and stewardship. The Data Seal of Approval (DSA)[324] may be awarded to repositories holding digital research data if they score highly in the DSA's 16 guidelines[325]. While not being one of the six recipients of this seal, the WDCC may claim adher-

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

[320] Science Communication Director for the non-profit geodesy consortium UNAVCO, Boulder, Colorado, USA, http://www.unavco.org.

[321] Head of IT infrastructure at AWI, and speaker of the Helmholtz Association's Open Access working group.

[322] http://www.dlib.org/dlib/january11/pfeiffenberger/01pfeiffenberger.html.

[323] http://www.earth-system-science-data.net/review/ms_evaluation_criteria.html.

[324] http://www.datasealofapproval.org.

[325] http://www.datasealofapproval.org/sites/default/files/DSA_informationfolder_web_0.pdf.

ence to all those guidelines, as it has the technical, institutional and cultural frameworks required to support such open data access.

The founding of the WDC cluster Earth System Research[326] in 2004 implies a commitment to long-term archiving facilities and data libraries with funding to meet the data management infrastructure, expertise and manpower needs. Members are WDC-MARE, WDC-Climate, WDC-RSAT and WDC-TERRA (WDC of the Lithosphere, GeoForschungsZentrum, Potsdam). The WDC cluster guides the peer-review process for scientific data, acts as a publication agent (compare with section 2.7), checks that metadata are complete, that the methods used have been validated and that data values are quality controlled with respect to their precision, sequence and ranges. After that, the independent data entities suitable for publication can be identified and data publication is initiated, i.e. the data entities with persistent identifiers (DOI) are then citable.

DKRZ is also a partner in several national and international projects which further develop a distributed data handling and processing infrastructure. At the conference "Grid Computing: a new tool for Science and Innovation" (VeliLošinj, Croatia, August 2009) Stephan Kindermann presented experiences from the German C3-Grid project and the prototype C3-Grid/EGEE integration and the current data-handling effort for the next Intergovernmental Panel of Climate Change (IPCC) assessment report. The abstract of his lecture "Distributed Data Handling Infrastructures in Climatology and the Grid" has been published on page 20 of the proceedings of this event[327] (compare also with sections 2.2–2.4).

More future developments in climate science e-Infrastructure were voiced at the ICTP conference mentioned in section 7. Kostas Glinos pointed to project METAFOR as an "integrated and coordinated approach to capture the comprehensive metadata requirements of the Climate Research community to create the Common Information Model (CIM) and the tools that will exploit it" and emphasised that the European Grid Initiative (EGI)[328] is the world's largest multiscience grid with over 350 sites including more than 200,000 CPUs and 100 petabytes of storage, which are being used to complete approximately 150,000 jobs per day.

At present, from July to November 2011, the 10th FP7-Infrastructures call is open for applications, and support is intended for the third implementation phase of Partnership for Advanced Computing (PRACE).[329] The roadmap,

[326] http://www.dkrz.de/daten-en/wdcc/wdc-cluster-earth-system-research?set_language=en.

[327] http://indico.cern.ch/getFile.py/access?resId=0&materialId=paper&confId=60021.

[328] http://www.egi.eu.

[329] http://www.prace-project.eu.

services and creation of a European High performance Computing (HPC) ecosystem were outlined by Sergio Bernardi,[330] representing CINECA,[331] the Italian “ConsorzioInteruniversitario” which is one of PRACE’s 21 members.[332]

The contribution of project METAFOR to creating a basic information infrastructure in support of climate science was illustrated in a comprehensive way (39 slides)[333] by Bryan Lawrence[334]

Steven Newhouse, another keynote speaker at the ICTP conference and project director of Integrated Sustainable Pan-European Infrastructure for Researchers in Europe (EGI-INSPIRE) described the EGI as indispensible for a sustainable production e-Infrastructure in the support of structured international research.[335]

Venkatramani Balaji[336] showed his and Dean N Williams’ presentation[337] on the Earth System Grid Center for Enabling Technologies (ESG-CET) and how data from multiple sources are being integrated (compare with sections 2.3 and 2.4). Taking CMIP5 as an example, Balaji summed up how a number of challenges in integration are being tackled. In a second presentation Balaji showed that in the project “Climate analytics on distributed exascale data archives” (ExArch),[338] investigations would be carried out on how to satisfy the need for policy-scale information from global-scale research and gave an example for removing uncertainty at regional scales (slides 25 and 26).[339]

Before listing the implications for OpenAIRE in the following section, we point to the various grid initiatives that are presently active in climate research such as WissGrid[340] and C3-Grid,[341] and support the call for an organised effort and leadership from funders, societies, journals, educators, individual scientists and the society at large to make measurable progress in

[330] http://cdsagenda5.ictp.trieste.it/askArchive.php?subtalk=1&base=agenda&categ=a10141&id=a10141s2t11/slides.

[331] http://www.cineca.it/en.

[332] http://www.prace-ri.eu/Members.

[333] http://cdsagenda5.ictp.trieste.it/askArchive.php?subtalk=1&base=agenda&categ=a10141&id=a10141s2t25/slides.

[334] Director of Environmental Data Curation at the UK Science and Technology Facilities Council, and head of the British Atmospheric Data Centre.

[335] http://cdsagenda5.ictp.trieste.it/askArchive.php?subtalk=1&base=agenda&categ=a10141&id=a10141s2t10/slides.

[336] Director of the Modeling Systems Group at NOAA/GFDL and Princeton University, USA.

[337] http://cdsagenda5.ictp.it//askArchive.php?categ=a10141&id=a10141s2t16&ifd=37961&down=1&type=slides.

[338] http://proj.badc.rl.ac.uk/exarch.

[339] http://cdsagenda5.ictp.trieste.it/askArchive.php?subtalk=1&base=agenda&categ=a10141&id=a10141s2t27/slides.

[340] http://www.wissgrid.de/index_en.html.

[341] http://www.c3grid.de/index.php?id=44&L=1.

handling and beneficial exploitation of the deluge of data that was made in the introduction to *Science*'s special issue on data handling[342]

9 Implications for OpenAIRE

In the foregoing sections, the infrastructure in climate research, climate data management and organisational and technical implications of some Open Access e-Infrastructures were described. From these and bearing in mind the four objectives stated in the OpenAIRE Annex I[343], the development of the following information services appears to be particularly important:

- climate science data need to be citeable
- the provenance of the growing amount of data needs to be trackable
- climate data sets and publications need to be assigned persistent identifiers
- the quality control procedure which data and publications have undergone needs to be documented and published together with the data
- it should be possible to publish data at a variety of publication levels: peer reviewed, quality checked, unchecked and at some intermediate stages
- support for harvesting the growing number of climate data archives needs to be established
- due to the international nature of climate science cooperation with the Open Access initiatives in other parts of the world needs to be fostered
- consistent e.g. OAIS-based[344] interfaces to long term climate data preservation repositories are needed to enable Open Access.

10 List of figures

[342] http://www.sciencemag.org/content/331/6018/692.short.

[343] Objective 1: Building Support Structures for Researchers in Depositing FP7 Research Publication (Networking); Objective 2: Establishment and Operation of the OpenAIRE e-Infrastructure for Peer-Reviewed Articles (Services); Objective 3: Exploration of and experimentation with Scientific Data Management Services (Research); Objective 4: Sustainability of the OpenAIRE e-Infrastructure and Supporting Structures, Exploitation and Promotion.

[344] Open Archival Information System, http://public.ccsds.org/publications/archive/650x0b1.PDF
http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683

11 List of tables

12 Bibliography

Brady, SR, Sinha, AK, & Gundersen, LC. (eds.) Proceedings from Geoinformatics 2008 – Data to Knowledge. Scientific Investigations Report 2008-5172. US Department of the Interior, US Geological Survey, 2008. http://pubs.usgs.gov/sir/2008/5172.

Curdt, C, & Bareth, G. (eds.) Proceedings from the Data Management Workshop, 9 30 Oct. 2009. Germany, University of Cologne, 2010.

Guralnick, R, Constable, H, Wieczorek, J, Moritz, C, & Peterson, AT. Sharing: lessons from natural history's success story. *Nature* 2009, 462, 34. doi:10.1038/462034a. http://www.nature.com/nature/journal/v462/n7269/full/462034a.html.

Hense, AV, Hense, AN, & Lautenschlager, M. Acquired, analysed, archived – climate data for our future. Presentation given at the DACH Conference in Bonn, 21 September 2010, 29 slides. 2010. http://umwelt.wikidora.com/wikidora/attach/Documents/PU_talk_DACH_20100924.pdf.

Hoeck, H. std-doi Publication of Climate Data at WDCC. Presentation given at the DataCite Summer Meeting in Hanover, 7–8 June 2010. http://datacite.org/slides/DataCite2010_Hoeck.ppt

Metz, B, Davidson, OR, Bosch, PR, Dave, R, & Meyer, LA. (eds.) Contribution of Working Group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. New York, Cambridge University Press, 2007. http://www.ipcc-wg3.de.

Parry, ML, Canziani, OF, Palutikof, JP, van der Linden, PJ, & Hanson, CE. (eds.) Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. New York, Cambridge University Press, 2007. http://www.ipcc-wg2.gov/publications/AR4/index.html.

Randall, DA, Wood, RA, Bony, S, Colman, R, Fichefet, T, Fyfe, J, Kattsov, V, Pitman, A, Shukla, J, Srinivasan, J, Stouffer, RJ, Sumi, A, & Taylor, KE. Climate models and their evaluation. In Solomon et al., 2007. pp. 589–662. http://www.ipcc.ch/pdf/assessment-report/ar4/wg1/ar4-wg1-chapter8.pdf.

Science Special online collection: Dealing with data. *Science* 2011, 331, 649.

Solomon, S, Qin, D, Manning, M, Chen, Z, Marquis, M, Averyt, KB, Tignor, M, & Miller, HL. (eds.) Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, New York, Cambridge University Press, 2007. http://www.ipcc.ch/publications_and_data/ar4/wg1/en/contents.html.

Spohr, D. Survey on the research infrastructure. 2010, Sample Questionary (see Appendix 1)

Taylor, KE, Stouffer, RJ, & Meehl, GA. A summary of the CMIP5 experiment design. 2009. http://cmip-pcmdi.llnl.gov/cmip5/docs/Taylor_CMIP5_design.pdf (updated 22 January 2011).

Taylor, KE, & Doutriaux, CM. CMIP5 Model Output Requirements: file contents and format, data structure and metadata. Presentation, 7 January 2010. http://cmip-pcmdi.llnl.gov/cmip5/docs/CMIP5_output_metadata_requirements.pdf.

Taylor, KE, Balaji, V, Hankin, S, Juckes, M, Lawrence, B, & Pascoe, S. CMIP5 Data Reference syntax (DRS) and Controlled Vocabularies, version 1.2. 9 March 2011. http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_syntax.pdf.

Toussaint, F, & Lautenschlager, M. The CERA-2 Meta Database and needs for a common information structure. Presentation given in Berlin, 14–15 October 2008. http://colab.mpdl.mpg.de/mediawiki/images/2/20/ESci08_Sem_3_CERA-2_Toussaint.pdf.

Williams, DN, Ananthakrishnan, R, Bernholdt, DE, Bharathi, S, Brown, D, Chen, M, Chervenak, AL, Cinquini, L, Drach, R, Foster, IT, Fox, P, Hankin, S, Henson, VE, Jones, P, Middleton, DE, Schwidder, J, Schweitzer, R, Schuler, R, Shoshani, A Siebenlist, F, Sim, A, Strand, WG, Wilhelmi, N, S& u, M. Data management and analysis for the Earth System Grid. *Journal of Physics: Conference Series* 2008, 125, 1, 01272.

Williams, DN, Doutriaux, CM, Drach, RS, & McCoy, RB. The flexible Climate Data Analysis Tools (CDAT) for multi-model climate simulation data. Proceedings of ICDM Workshops, 2009. pp. 254–261.

13 Appendices

Appendix 1: Survey on the research infrastructure at a specific institute, Sample Questionary

Survey on the research infrastructure at

NAME OF INSTITUT

This survey is part of a study which is being carried out within an EC-funded project. This study aims to derive discipline-specific requirements on research infrastructure.

The participation in this survey is voluntary. Your answers will be treated anonymously and will only be used for the purposes mentioned above. This research underlies the rules and regulations of the data protection act. Your answers are not linked with your person in any way.

After completing the questionnaire in your favourite PDF viewer, please print it as a PS or PDF file and send it to **NAME OF RESEARCHER (E-MAIL, address)**. In case you have any questions regarding the project, the present study or this survey, please contact **NAME OF RESEARCHER**.

We would like to thank you for your participation in this survey.

A. General information

A.1 Personal information

Title Ms/Mrs/Mr, Dr/PhD, Prof. Dr, Jun. Prof.

Last name, first name

Position

Role and duties

A.2 Information on the working group

Name of the group		
Primary research objects		
Size of the group		
	People	ca.
	PCs	ca.
	Servers	ca.
	Other research instruments	
Number of cooperations with other groups		
	inside **INSTITUTE**	ca.
	outside **INSTITUTE**	ca.

B. Data infrastructure

B.1 Data lifecycle

Please use the menus to sketch the different stages that research data typically pass through in your group, and explain them briefly wrt. the time frame, used instruments, metadata and possibly involved people. Please add further stages if necessary.

	Stage	Explanation (e.g. time frame, instruments, people)
1	Choose from: Data collection Processing Enrichment Archiving Reuse Distribution	
2	Choose from: See above	
3	Choose from: See above	
4	Choose from: See above	
5	Choose from: See above	

B.2 Data collection

Which types of primary data are typically collected in a series of experiments concerning a particular object of research, and what are the storage requirements in terms of size?

	Audio data	Choose from: kB range MB range GB range TB range
	Video data	Choose from: See above
	Textual data	Choose from: See above
	Other Data:	Choose from: See above
Do you develop software in order to collect primary data?		Y / N
To what extent do you rely on proprietary software in order to collect primary data?		Choose from: not at all hardly sometimes considerably (almost) entirely

B.3 Data processing and data formats

Which types of secondary data are typically derived on the basis of the primary data, and how are they represented?

Are the data further annotated with metadata?	Y / N
If so, which kinds of information are typically expressed?	
Are there established metadata standards in your field?	Y / N
If so, do you use them?	Choose from: not at all infrequently quite frequently (almost) always
Are there kinds of information for which you have to deviate from standard formats (e.g. because they cannot be represented)?	Y / N
If so, which are these kinds of information?	
Do you develop software to derive secondary data?	Y / N
To what extent do you rely on proprietary formats and software for data representation and processing? Formats Software	Choose from: not at all hardly sometimes considerably (almost) entirely Choose from: See above
Further explanations (e.g. information on proprietary formats and software):	

B.4 Data management

How are developed software as well as primary and secondary data stored and archived?

	Software	Primary	Secondary
In a repository within the group or institute (e.g. hard disk on a central file server of the group)			
In a repository shared with other groups or institutes (e.g. faculty-wide or university-wide)			
In an external repository			
Mainly on your own storage devices (e.g. on the hard disk of your office PC)			
Would you consider the situation in your group representative for your research discipline?	Y / N		
If not, what is common practice in this respect?			
Are there members of your group which deal with data management issues?	Y / N		
If not, would you consider it desirable or even necessary?	Choose from: yes, would be necessary desirable, but not necessary no, not desirable		
Why?			

B.5 Access to data

Which practices exist in your group in order to ensure the access to produced data?

Which tools support these practices?

B.6 Publication and exchange of research data

Would it be conceivable in your group to make research data available to others (in anonymised form where necessary)?	Software	Primary data	Secondary data
only among your close colleagues also to other research projects to the general public			
If not or only restricted, which disadvantages or boundaries do you see?			
Which incentives for data exchange are mentioned in your group, and are there technologies supporting these incentives?			
Would you consider the situation in your group representative for your research discipline?	Y / N		
If not, what is common practice in this respect?			

Do you make software available to other institutions?	Choose from: yes, as a matter of principle yes, sporadically no, but is conceivable no, and is not conceivable
If so, does this typically include source code?	Choose from: yes, as a matter of principle mostly frequently rarely never
Explanations and reasons	

Which rules exist in your group wrt. use, exclusiveness and time frames for exchanging data?

How is data exchange supported and which tools are used?

Which practices and tools enable data reuse, and which boundaries exist?

C. Literature

C.1 Organisation

Which tools do you use to organise internal and external publications?

Which forms of publication are preferred in your group?
(Check first column and choose from options given in third column)

	Publication as print medium (e.g. article or book)	Choose from: preferred and established preferred, but rather infrequent not preferred
	Electronic publication online or offline	Choose from: See above
	Combination (e.g. book/CD-ROM or proceedings/website)	Choose from: See above

Would you consider the situation in your group representative for your research discipline?	Y / N
If not, what is common practice in this respect?	

Which scientific journals, publishers etc. are preferred in your group/discipline?

C.2 Combination of literature and research data

Are there publishers which enable the exchange of data and/or literature?	Y / N
Is it currently possible to publish data and literature together?	Y / N
If not, would you consider this reasonable or desirable in your discipline?	Y / N
Which developments would be necessary in order to achieve this?	
Which (dis)advantages do you see?	

C.3 Open Access

Is Open Access established in your group?	Chose from: yes a bit no
Would you consider the situation in your group representative for your research discipline?	Y / N
If not, what is common practice in this respect?	
Which technologies and practices support Open Access publications?	

D. Outlook

Which future developments for research infrastructures do you see in your discipline? If possible, please comment on developments wrt. an Open Access infrastructure as well.

Data infrastructure

Literature infrastructure

Thank you very much for your participation!

DATE

Appendix 2: Terms of use for data from WDCC

Conditions for using the WDCC database[345]

1 Creative Commons Licence All data and pages available from WDCC are licensed under a Creative Commons Licence (http://creativecommons.org/licenses/by-nc-sa/2.0/de/deed.en) as far as those conditions are

[345] http://cera-www.dkrz.de/WDCC/ui/docs/TermsOfUse.html.

not in any way modified by the following conditions or by any conditions specific to data or pages.

2 Special data owner conditions You must agree with the special data owner conditions (CERA WWW-Gateway » Browse experiments » Show datasets » Click on dataset name » Distribution). If there is a conflict between the Creative Commons Licence and the special data owner conditions, the latter shall have precedence.

3 Always quote reference Articles, papers or written scientific works of any form, based in whole or in part on data supplied by WDCC, will contain an acknowledgment concerning the supplied data. Always quote reference of the experiment in the citation index when using WDCC data (CERA WWW-Gateway » Browse experiments » Experiment information » CERA UI Compact » Citation.

In addition to your comments and suggestions, we are VERY interested in how you use the data we distribute. We would be most appreciative if you would send us a line or two describing your application of these data to data@dkrz.de.

4 Used freely for research only Data from the projects available on the WDCC server may be used freely for research only.

Non-research use of WDCC data: If you do not feel able to accept the conditions for access to the WDCC research data service, you may still be able to acquire data via WDCC Data Services. Please contact data@dkrz.de.

5 Personal account The WDCC will provide you with a personal account and password. If you copy, distribute, display and perform the data, you may do it only under the conditions identical to this one. You are responsible for maintaining the confidentiality of any password(s) you are given to access the WDCC site and are fully responsible for all activities that occur under your password(s). You agree to notify WDCC immediately of any unauthorised use of your password(s).

6 Audit All downloads from WDC-Climate are audited.

7 Share alike Any person extracting data from this server will accept responsibility for informing all data users of these conditions.

8 Liability/warranty

(1) The data are delivered to the user without a warranty of any kind. The user is aware that the data were generated in keeping with the current state of science and technology.

(2) The WDCC assumes no obligations on the basis of this agreement towards third parties. There is no liability of the WDCC for any harm arising from the delivery and subsequent processing of the data products. The user exempts the WDCC from any liability towards third parties.

The disclaimer under 8 (1) and (2) does not apply if and in so far as the WDCC has acted with gross negligence or with intent.

9 Other provisions

- If the user is a merchant, a legal person under public law or a special asset under public law, Hamburg shall be the place of jurisdiction for all disputes arising from this use agreement.
- Should there be one or more provisions in this agreement that are void, in whole or in part, then this will not affect the validity of the other provisions. The invalid provisions will be replaced retroactively by an arrangement as similar as possible in content and coming closest to fulfilling the purpose of the intended provision.

G Health Sciences

Johanna McEntyre and Alma Swan

1 Introduction and methodology

1.1 Introduction

This chapter provides an overview of research data management in the health sciences, primarily focused upon the sort of data curated by the European Bioinformatics Institute and similar organisations. In this field, data management is well-advanced, with a sophisticated infrastructure created and maintained by the community for the benefit of all.

These advances have been brought about because the field has been data-intense for many years and has been driven by the challenges biology faces. Science in this area cannot be done on a small scale: it is effectively a collaborative effort where data must be shared for any advances to be made. This has long been acknowledged. The HUGO (Human Genome Project) set the standards, because the demands of that project were so great that only a concerted effort across the whole genome science community would enable the achievement of that goal. It established new norms of scientific behaviour in this discipline and has influenced cultural developments in the discipline ever since.

The human genome is now long-decoded, but today's scientific questions in health sciences are no less challenging. The infrastructure, practices, standards and norms established in the life sciences can be viewed as good practice markers for those who wish to learn from what has gone before. Not everything practised in the life sciences will read across to other fields and disciplines, but many basic principles of research data management practice have been established that will transfer readily elsewhere. Perhaps most importantly, the life sciences have now reached the stage where the issues of long term planning, organisation and sustainability are now being tackled. The answers to these things are only partially worked out as yet, but some fundamental principles are being elucidated and these will be useful in a more general sense.

1.2 Methodology

The material in this chapter was developed by the following means:

- Literature review and analysis
- Semi-structured interviews with experimental and theoretical scientists
- Observational studies of experimental biologists working at EBI, The Sanger Institute and in a number of universities in the UK

2 The European Bioinformatics Institute: an overview

EBI (the European Bioinformatics Institute) is part of the European Molecular Biology Laboratory (EMBL). It was established in Hinxton, UK, in 1994 to build on EMBL's pioneering work in providing publicly-available biological information in the form of databases to the scientific community.

Such information was beginning to accumulate rapidly as molecular biology technologies created increasing amounts of data. New skills and resources were required to collect, curate and store these data and present them to the research community through reliable, professionally managed channels.

From small beginnings, EBI has grown and now provides data resources across all molecular biology domains. It hosts a number of public databases, most through collaborative initiatives with partner organisations throughout the world, particularly in Europe, the US and Japan. Services include Ensembl (a genome database), ENA, the European Nucleotide Archive (containing DNA and RNA sequences), UniProt (containing protein sequences) and PDBe, the European arm of the Protein Data Bank. The expression data is captured by ArrayExpress Archive which is a database of functional genomics information and the Gene Expression Atlas, which contains expression data from Array Express that has been re-annotated for particular purposes. More recently, the EBI has developed CiteXplore, a database of biomedical abstracts from research articles and patents, and UKPMC (UK PubMed Central), a full-text article database. This source is used actively for gathering information from literature to create information-rich databases such as GOA (Gene Ontology Annotation) and IntAct, and for text mining information.

These and other services, such as ChEBI (Chemical Entities of Biological Interest), Reactome, and InterPro, offered mean that EBI is the primary provider of biological information in Europe, and one of the major global providers.

Integration and distribution of information from and to various sources requires standardisation of the data input, storage and distribution. EBI scientists have been active in developing or contributing to the community efforts towards the development of international standards for use in bioinformatics. Two examples involving EBI are the MIAME standard for microarray experiments (Minimal Information About a Microarray Experiment) and the Human Proteome Organisation's Proteomics Standards Initiative (PSI).

EBI plays a coordinating role in Europe with respect to bioinformatics work, such as the ELIXIR (European Life sciences Infrastructure for Biological Information) objective to fund and maintain a world-class infrastructure for life science information management in Europe; ENFIN (Experimental Network for Functional Integration), an initiative to bring together experimental and computational biologists to develop the next generation of informatics resources for systems biology; IMPACT (Experimental Network for Functional Integration), developing infrastructure to improve protein annotation; and the SLING Integrating Activity (Serving Life Science Information for the Next Generation) which aims to bring together a wide range of information sources and services and help them to keep pace with scientific developments.

EBI also has a substantial programme of research. Research groups collaborate in experimental areas such as genomics, developmental biology, protein structure, evolutionary studies and cellular interactions, amongst others. EBI also contributes to research in the areas of computational biology and the development of simulation and modelling systems and of mark-up standards for biological data.

A further area of operation for EBI is training in bioinformatics, providing very active international PhD and postdoctoral training programmes for young researchers aiming to become bioinformaticians.

Additionally, EBI runs training programmes for users of biomedical data to equip experimental biologists with the skills needed to best use the information resources that EBI provides. An e-learning programme is being developed to complement the face-to-face training programmes.

EBI also coordinates the Bioinformatics Roadshow, a mobile training programme run in collaboration with the Swiss Institute for Bioinformatics, the European Patent Office and the BRENDA project (BRaunschweig *ENzyme* Database, an initiative of the Technical University of Braunschweig).

Up-to-date information on the Institute's activities can always be found in its Annual Report[1].

3 The health sciences

3.1 The scope of research activities in health sciences

Health science is a very broad area. It spans some elements of environmental science at one end of the spectrum through biomedicine, clinical medicine and veterinary science to medical physics and mathematical biology. Health-related questions and issues are studied at multiple levels.

At the molecular level, researchers study biomolecules and their activities and interactions in fields such as genomics (the study of the genetic complement of organisms), transcriptiomics (the functional transcript of the genetic component), proteomics (the study of the structure and function of proteins), metabolomics (the study of small molecules, metabolites, that are generated by living systems), macromolecular structures and interactions, and bioinformatics (the application of computer methodologies and statistical analysis to the study of molecular systems). The above resources are used to study network biology and regulation of biological systems, to eventually give us an understanding of systems biology.

At the next level – the study of cellular processes and behaviour – research is aimed at elucidating the ways in which cells and tissues interact and influence others and how cellular systems are regulated. Scientists also work at understanding how these systems relate to known molecular pathways and events and what can lead to cell and organ dysfunction. Included in this area of research activity is the study of mechanisms that form the basis of disease. *In vitro* (experimental) model systems are developed for human and animal diseases and malfunctions in order to try to understand what processes are aberrant in disease conditions. In addition, *in vitro* or computer model systems are used to research potential therapeutic agents and to test for toxicity.

At whole-organism level research areas include infection and immunity (encompassing the areas of immunology, microbiology and pharmacology); the wide-ranging clinical medicine disciplines; therapeutics and translational research; transplantation and regeneration; toxicology and environmental health; public health; and aging and wellbeing research.

With such a broad-scope area as health sciences, research activities are necessarily hugely varied. Research can be, at one end of the scale, devel-

[1] http://www.ebi.ac.uk/Information/Brochures/pdf/Annual_Report_2010_low_res.pdf.

oping advanced instrumentation for clinical therapies to, at the other end, sequencing the mutated gene responsible for a very rare disease. At each of these steps, however, a consolidated source of information is very useful and this can be provided by various sources made available at EBI.

3.2 Types of research activity and the main experimental and theoretical methodologies used

For the purpose of this exercise we are not attempting to cover the whole gamut of health science research areas. Instead, we focus on the research that looks at biochemical and sub-cellular processes, along with any allied approaches this may entail.

The areas of focus therefore include: genomics, proteomics, and metabolomics; macromolecular structures and interactions; cellular structure, function and signalling; and bioinformatics.

A major component of the above approaches is the analysis and production of data on a large scale, which, in the biomedical sciences is a process facilitated by the availability of public databases for both data deposition and retrieval for analysis. The data resources at the EBI, for example, have grown dramatically in recent years (see Figure G.1 below), This has led in some cases to "data-driven science" (PMID: 14696046) in which analysis of large public datasets gives rise to new hypotheses.

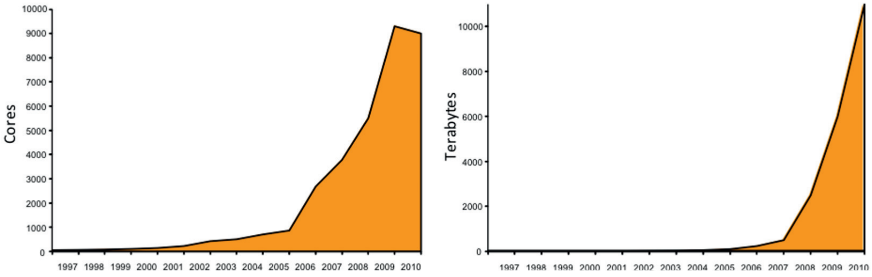


Figure G.1 Growth of compute and storage of data at EBI

Other experimental approaches that may be used in biomedical investigations include light and electron microscopy techniques (which produce images), scanning techniques (which also mainly produce images), and biochemical analytical techniques such as nuclear magnetic resonance and chromatography (which produce text or data files after computation of the machine analysis.

Theoretical approaches to research in this field include second-level informatics and modelling. These use computational techniques to further process data from experimental procedures. Models may be of a number of types, including mathematical models, computer simulations, computer models and 3-dimensional models. The development of integrative technologies is an area of considerable research focus: researchers are creating a wide range of data-integrative algorithms that enable combined analysis of diverse data sources.

These methodologies are described in more detail in section 4.1.

3.3 Types of research output and the way they are used

The research activities that are described here produce outputs of the following types:

(i) *"Big data"*. The deposition of data in public databases such as those provided by the EBI, are the end-point of some experiments but increasingly provide starting points for others. The deposition of data in public databases is increasingly a requirement of journals for publication. Storing data in centralised databases with uniform format and structure allows the development of computational tools for comparative data analysis (e.g. BLAST) and in-depth search and display mechanisms[2]. The figure below shows records available in some key public data resources maintained at the EBI:

(ii) *Research-lab generated datasets.* Experimental data are exploited first by their creators (researchers) and may:

  a. remain in the possession and care of those creators. In these cases, researchers may elect to share datasets with enquirers or with the research field at large, perhaps via a public website or service (see section 6).
  b. get deposited in the public databases (often a requirement of journals) (see section 6.2) or
  c. appended as supplemental data files attached to research articles published in peer reviewed journals (see section 6.3)

(iii) *Research articles in journals.* In health sciences, journal articles are the primary output type for research findings, contrasting with some other fields like computer science and engineering where peer-reviewed conference proceedings are the main dissemination channel. There are several thousand peer-reviewed journals covering biomedicine so finding an outlet for publication is not especially difficult.

Keeping up with the literature in this field is, however, challenging given the volume of papers published in each year. Probably half the total research

[2] See, for example: http://www.ebi.ac.uk/Tools/sss/

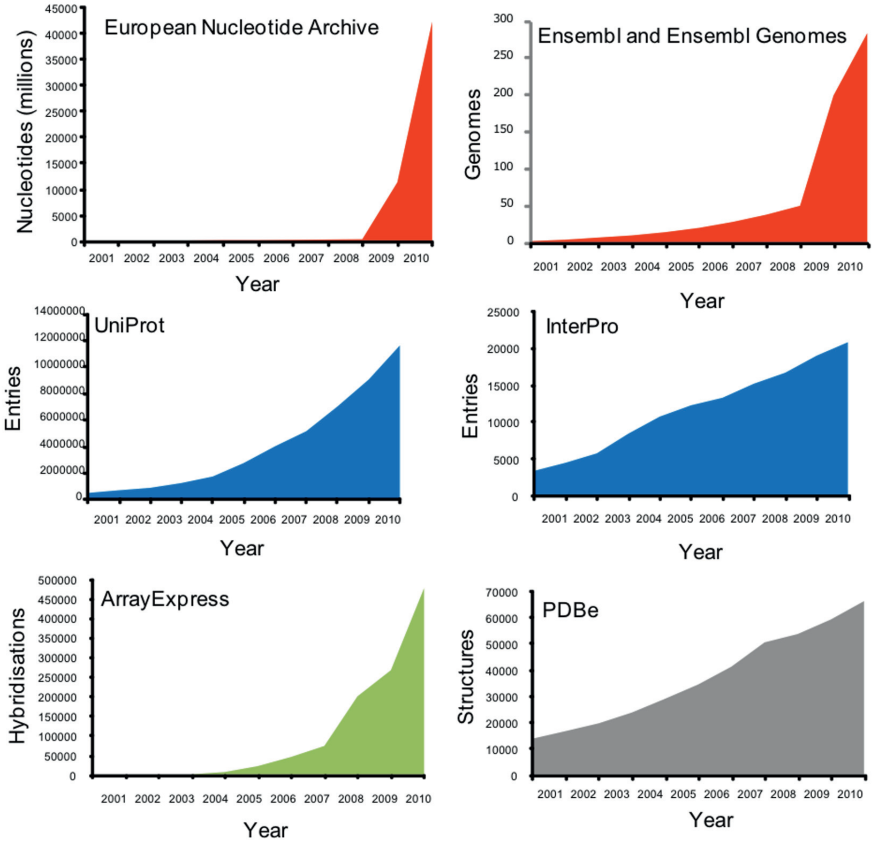


Figure G.2 Growth of key resources at EBI

literature is in health sciences, reflecting the priority that research in this area has for society and the levels of funding received from governments and other research funders.

Journals in the discipline are published by commercial publishers, medical charities, learned societies, medical institutions, and universities and research institutes.

(iv) *Conference papers.* Peer-reviewed conference proceedings are not common in health sciences, but they form an occasional outlet for research findings.

(v) *Outputs through more informal channels such as blogs, wikis, open notebooks and similar.* In recent years there has been growth in the use of online sites for the management of projects and the dissemination of materials from them. Especially in fields where the research cycle is short and progress is very rapid, such informal channels may be the best way to alert the community to new developments and findings.

Occasionally, laboratories do disseminate pre-publication results from analytical machine runs via wikis or blogs, for early communication through these routes. Access to data generated by other people's work may therefore be facilitated in this way and some bioinformaticians hunt for and harvest data from such websites for their own meta-analysis, sometimes by screen scraping. The data so obtained , however, do not have the parsable format that would make them more amenable to re-use and distribution.

One of the most successful and well-known examples of a community 'Web 2.0'-type facility is Open WetWare[3], a site that hosts individual laboratory websites and on which users share results, protocols, details about materials and so forth.

3.4 Workflows in life science research

3.4.1 Genomics

Gene sequencing experiments

- Experiment planning: Define experimental goals; identify source of sample(s); agree experimental conditions; plan and prepare for use of experimental machines; plan data handling procedures
- Experimental process: prepare samples and carry out machine run (on one or multiple samples)
- Data production: machine produces raw data (traces) and processed data (text-based outputs)
- Data storage and preservation: Optionally, store preliminary data from machine ('raw' pre-base called data) for future analysis or further processing if ever necessary: routinely, store the text-based base-called data
- Data quality checking: carry out manual quality check; discard datasets with obvious errors
- Data processing and enrichment: process data; annotate datasets where appropriate
- Data publishing and storage: complying with agreed standards, (e.g. MIARE, Minimum Information About an RNAi Experiment or MIAFGE, Minimum Information About a Functional Genomics Experiment) deposit representative dataset(s) in public databank (e.g. Genbank)

[3] http://openwetware.org/wiki/Main_Page

- Data analysis: capture relevant datasets from Genbank for computational analysis using preferred software
- Re-submit processed datasets if the data have been improved in some way
- Susceptibility or resistance to infection can be provided by SNP (single nucleotide polymorphism) or variation analysis or epigenetic analysis of the genome

Microarray experiments

- Experiment planning: Define experimental goals, identify source of sample(s); agree experimental conditions; plan and prepare for use of experimental machines; plan data handling procedures
- Experimental process: prepare samples and carry out machine run (on one or multiple samples)
- Data production: machine produces raw data (images)
- Data processing and enrichment: normalise raw data; statistically analyse data
- Data publishing and storage: complying with agreed standards (e.g. MIAME, Minimum Information About a Microarray Experiment), deposit representative dataset(s) in public databank (e.g. ArrayExpress)
- Data analysis: capture relevant datasets for computational analysis using preferred software
- Re-submit processed datasets if the data have been improved in some way

3.4.2 Proteomics

- Experiment planning: Define experimental goals, identify source of sample(s), agree experimental conditions, plan and prepare for use of experimental equipment; plan data handling procedures
- Experimental process: prepare samples and carry out experimental procedure (on one or multiple samples)
- Data production: collect results from experiment
- Data processing and enrichment: process data; annotate datasets where appropriate
- Data publishing and storage: complying with agreed standards, (e.g. MIAPEgelDB, Minimum Information About a Proteomics Experiment [gel electrophoresis]) deposit representative dataset(s) in public databank (e.g. PRIDE, PRoteomics IDEntifications database)
- Data analysis: capture relevant datasets for computational analysis using preferred software

- Re-submit processed datasets if the data have been improved in some way

3.4.3 Metabolomics

- Experiment planning: Define experimental goals, identify source of sample(s), agree experimental conditions, plan and prepare for use of experimental equipment; plan data handling procedures
- Experimental process: prepare samples and carry out experimental procedure (on one or multiple samples)
- Data production: collect data from experimental process
- Data processing and enrichment: process data; apply statistical analysis or visualisation techniques; annotate datasets where appropriate
- Data publishing and storage: complying with agreed standards
- Data analysis: capture relevant datasets for computational analysis using preferred software
- Re-submit processed datasets if the data have been improved in some way

3.4.4 Computational bioinformatics

- Experiment planning: Define experimental goals; identify source of data: agree experimental conditions; plan and prepare for use of experimental machines (if relevant); plan data handling procedures.
- Experimental process: prepare samples and carry out machine run (on one or multiple samples) if generating primary data: or, process previously-created data
- Data production: machine produces raw data (traces) and processed data (text-based outputs)
- Data storage and preservation: Store preliminary data from machine (‘raw’ pre-base called data) for future analysis or further processing if ever necessary
- Data quality checking: carry out manual quality check; discard datasets with obvious errors
- Data processing and enrichment: process data; annotate datasets where appropriate (for example, combine microarrays for analysis as a group, align gene sequence data to the genome, etc); convert data to appropriate commonly-used file formats (e.g. SAM/BAM for sequence alignment data)
- Data publishing and storage: complying with agreed standards, (e.g. MIARE, Minimum Information About an RNAi Experiment) or MIAFGE, Minimum Information About a Functional Genomics Experi-

ment) deposit representative dataset(s) in public databank (e.g. Genbank)
- Data analysis: capture relevant datasets from Genbank for computational analysis using preferred software
- Re-submit processed datasets if the data have been improved in some way. Options include adding data to the UCSC Genome Browser[4] for use by a larger audience

3.4.5 Microscopy

- Experiment planning: Define experimental goals, identify source of sample(s), agree experimental conditions, plan and prepare for use of experimental equipment; plan data handling procedures
- Experimental process: prepare samples and carry out experimental procedure
- Data production: collect data from experimental process (micrographs)
- Data processing and enrichment: manipulate and analyse image data by computational techniques
- Data publishing and storage I: store locally on hard drives or transportable media, or submit to public database (e.g. the Mouse Brain Library), depending on the type of project and collaborative nature of the work
- Data analysis: capture relevant datasets for computational analysis using preferred software
- Data publishing and storage II: store data derived from analytical/processing step locally on hard drives or transportable media, or submit to public database, depending on the type of project and collaborative nature of the work

3.5 Case studies: short description of typical use cases in health science research

Case study 1:
The biology of Trypanosoma brucei, the parasite that causes sleeping sickness

The research is aimed at obtaining a better understanding of the parasite's biology and virulence. It involves examining the genomic sequences that encode components of the flagellum (the parasite's organ of motility) and the molecular processes that drive the motor functions. The antigenic surface pro-

[4] http://genome.ucsc.edu/

teins on the parasite's coat and the behaviour of the parasite's chromosomes at cell division are also studied.

Methodologies used include comparative genomic analysis, protein analysis and cytological studies of chromosomes using the FISH technique (see section 3.1). Mass spectrometer data are generated from proteomic studies. Protein-protein and protein-small molecule data are obtained from co-purification studies and provide information about molecular interactions within the parasite and between host and parasite.

Curated databases like UniprotKB, Ensembl and others gather additional information and help the meta analysis. Data are also available via cross-referencing to the homologous genes from other organisms. These data may provide indicative evidence for the role of the proteins.

Meta-analyses in the form of curated databases are also produced when looking at potential homologies between genes. This work involves the use of data from other laboratories: these data are obtained from public databases or from datasets supporting journal articles.

Software for data analysis is either written in-house or an existing package is tailored to suit the research group's needs. Data are made available through Ensembl Genomes[5]. Annotated datasets are substituted each time the dataset is updated, so although original gene sequences are archived the annotated datasets are continually updated.

Published data are always processed and annotated to an extent, though effort is made to publish data in as useful a form as possible (this is not necessarily the norm throughout this research community). Software tools produced in the laboratory may also be made available on the research group's website if they have potential use outside that laboratory.

Case study 2:
neuroimaging in psychiatric diagnosis and therapy

This case is about research into the prevention of psychosis and the role of neuroimaging methods, and data curation and sharing, in this effort.

Isolated or transient symptoms of psychosis are common, but the development of a clinically-defined psychotic state only follows in a minority of patients. Patients with high risk of developing schizophrenia (risk based on genetic factors) undergo MRI (magnetic resonance imaging) scanning procedures for diagnostic evaluation, which is supported by genomic analysis. Many of the phenotypic characteristics are shared between many distinct genetic diseases. Also many of the diseases have multiple causes with similar manifestations. The MIM (Mendelian Inheritance in Man) database, a de-

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

scriptive database, is a major source of information on these various diseases. Two major developments are underway as part of the research, and these will provide publicly-available community resources for the future.

A collaborative effort to integrate image data from multiple sources is being undertaken involving clinical teams, imaging experts and e-scientists, to create a Grid-based network of neuroimaging centres and a neuroimaging toolkit. The aim is to share data, experience and expertise so as to facilitate the archiving, curation, retrieval and analysis of imaging data from multiple sites and enable large clinical studies. The process involves: collecting retrospective data to help develop ways of harmonising scans from different machines; integrating existing datasets (scans and other clinical information); and developing a generic ontology for a psychosis database that can be used in future studies and clinical management.

A second collaborative effort is a further health informatics initiative that aims to develop a functioning "e-community" and build a secure electronic database to hold anonymised clinical data about people presenting with first-episode psychosis. The focus is on working with the network of research centres to create a shared metadata model and ontology of terms for clinical and biological data relevant to psychosis. Decipher, a Database of Chromosomal Imbalance and Phenotype in Human diseases, uses Ensembl resources to help pinpoint chromosomal sources of imbalance in patients, though its data are not openly available.

For this project a formalised risk assessment process for digital repositories (DRAMBORA[6]) has been applied, along with the OAIS functional model for archival information systems, to consider recommended activities for a data archive for all these data.

Dissemination of findings from this case is through journal articles in basic and clinical neuroscience journals, and also in the form of image and numerical datasets that can be shared via the Grid-based system being created.

Case study 3: *the mechanics and dynamics of cell division*

The focus in this project is on the way in which the products of chromosome duplication are separated and moved equationally and simultaneously into the two daughter cells at cell division. If this process is faulty, the resultant daughter cells will be non-viable or malfunctioning.

The research questions are mainly of a molecular/mechanical nature and relate to the mass, speed, distance and timing of the main events of cell

[6] Digital Repository Audit Method based on Risk Assessment: http://www.repositoryaudit.eu/about/

division. Four components are involved: the chromosomes; a system of microtubules making up the mitotic spindle; the site to which the microtubules attach to the chromosome – the kinetochore; and the structure to which the microtubules are anchored at the poles of the mitotic spindle – the centrosome.

The research involves various kinds of approach: first, the use of advanced techniques of light microscopy (confocal microscopy, fluorescence microscopy) coupled to the use of immunofluorescent labels that attach specifically to defined components of the chromosome, the spindle, the kinetochore and the surrounding cell matrix; second, the use of mutant organisms in which cells lack specific proteins that are important for spindle function; third, the study of shifts in cell chemistry that coincide with microscopically observable events; and, fourth, the study of physical properties, such as visco-elasticity, and physical strain in relation to dimension and the energy required to move chromosomes quickly and directionally through the inside of a living cell.

Such studies have highlighted, amongst other things, the astonishing reliability of the system and the extensive fail-safe redundancies that have evolved to reduce the level of failure – quite important in a human, for example, where up to 100 million cell divisions are taking place at any one time.

Recent research in this area has involved a high level of interdisciplinary collaboration, with physicists, mathematicians, statisticians and engineers working alongside microscopists, geneticists, and molecular biologists. The experimental set-up, though microscopy-based, is heavily computerised, with software drivers for cameras, microscopes and analysing and manipulating the results. This analysis can be carried out during the experiment if required.

Dissemination of results is almost exclusively through journal articles summarising the work and which include the photomicrographs from microscopy. Image data are stored locally.

Case study 4:
quantitative models for simulating neuronal cell signalling

This research project develops computational models of cell signalling (interactions between cells that involve a signal-response effect) at multiple scales. The tools and technologies used are modelling environments and simulation software.

The group is also in charge of the world reference database of such models. As custodian of such things, one of the areas of work the group undertakes is the development of good practice procedures and standards. Data are readily shared with other scientists: standard XML formats are used which are richer and more easily re-usable and exploitable than text-based data or spreadsheet formats.

A large toolkit of standards, formats and ontologies has been created to describe, annotate and interface the models created by the community. As an example, one of these, Minimum Information Requested In the Annotation of biochemical Models (MIRIAM) is a set of rules defining minimum standards for encoding the models for the biochemical modelling community and was developed by the group in the mid-2000s.

Because such standards guide metadata/annotation practice and encourage the use of standardised terminology, searching for datasets of interest is facilitated, the community's confidence in found datasets is maximised and they can be re-used with precision. In all, the value is greatly increased as a result.

Data and models from the research group are published in journal articles and on the group's websites as well as in relevant web-based public databases. Wikis are used to perform the work in the laboratory but not so far to exchange or annotate datasets.

Case study 5:
databases for mouse embryonic development

The research project is to develop a publicly-available resource, a detailed model of the mouse embryo through development from conception to birth. The database provides information about the morphology (shape, size and structure) and histology (tissue structures at cellular level) of the embryo at different stages of development. It also provides the framework for adding information from genetic studies about gene function and expression in the embryo.

The database differs from other, tabular databases by enabling different types of information to be mapped onto the 3-dimensional organism. An extensive anatomical ontology has been developed to aid in the understanding of the relationships between the embryo's anatomy and other spatial, temporal and genetic information. The ontology of anatomical names was mapped to successive developmental stages of the mouse embryo.

A sister database of gene expression data has also been produced by the same project team. Data for this database were sourced from published reports, datasets in public databanks, original data from laboratories, and datasets from large-scale projects. Together, these two databases – of embryo structure data and gene expression data – provide an overall resource rich in detail and functionality that can be used in research and teaching[7].

[7] http://www.emouseatlas.org/emap/home.html

The research team consists of both biologists and computer scientists and has a full-time curator for the databases. Scientists in the community are invited to submit data for inclusion in the databases. Datasets that are accepted for inclusion are curated by the project team.

The databases are made publicly-available through a website. Users can view histological sections through an embryo in different planes, 3-D models and reconstructions, and videos of a rotating embryo prepared in different ways including by differential staining, to show its complete external morphology. The genetic data are analysable by text-based and spatial-based methods.

4 Current status of research infrastructure, workflows and life cycles

The research infrastructure should enable scientists to:

- access the physical resources, materials, and services necessary for their research, at the point at which they are needed
- access the information resources and the networks that transmit them, at the point at which they are needed
- have the means to access these resources at the time of need and have the skills required to use them to best advantage
- have confidence in the quality and integrity of these resources
- access the technologies they need for advanced or collaborative work
- have the means and skills to exploit these technologies to maximum effect
- have the analysis tools to include their data and analyse them with respect to the data already contained within the database

The following section describes the workflows and research systems and processes that operate in the areas of focus in biomedical research.

4.1 The experimental infrastructure: approaches and protocols

Experimental approaches vary between the fields that are the focus of this report.

4.1.1 Genomics

Genomics is the study of the genetic make-up of living organisms. Genes are composed of long runs of nucleotides (bases) that form the backbone of DNA. Their sequence along the DNA determines the function of the gene, since this sequence is transcribed into functional or messenger RNA and thence to proteins in the cell. The expression of genes at spatial, temporal and intensity levels is also studied.

(a) Genome sequencing

The sequencing of DNA is therefore at the core of genomic investigations, and since the human genome was sequenced in 2003, the technology for DNA sequencing has been dramatically advanced. The most recently developed massively parallel processes, referred to as "next generation sequencing" can output hundreds of millions of short DNA fragments of around 50 bases long in a matter of days. The bottleneck of making use of these data has switched from data generation to data analysis, with huge computational power now required to assemble these short strings of bases into complete genome sequences[8]

DNA sequencing, while still carried out in many research laboratories around the world in pursuit of specific research questions, is becoming an industrialized process, with many companies now offer next-generation sequencing services or related products. These technology changes are ensuring that sequence submissions to public databases continue to increase at exponential rates (see Figure G.3 below).

The availability of these powerful sequencing technologies are allowing genomic scientists to undertake comparative genomic experiments on a mass scale, giving rise to efforts such as the '1000 genomes project'[9], that have huge potential benefits for human health and well-being.

In addition, relatively cheap genome sequencing methodologies will have the power to revolutionise taxonomy and ecogenomics studies. Taxonomical relationships between organisms will be much easier and quicker to elucidate, and the application of molecular sequencing techniques on a genome-wide scale will have massive benefits for research that is aimed at better understanding ecological and evolutionary processes.

Figure G.4 shows a record of a short-read sequence from a nucleotide database.

(b) Gene expression

The other most commonly used methodology for looking at genetic activity is microarray technology. Microarray technology is a way of studying gene

[8] See for example: http://www.nature.com/nmeth/journal/v7/n7/fig_tab/nmeth0710-495_T1.html

[9] http://www.1000genomes.org

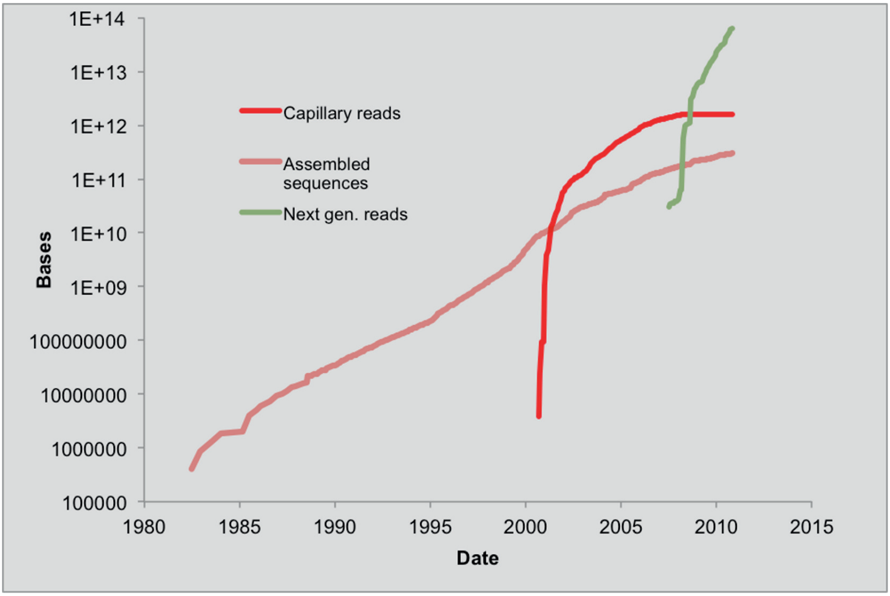
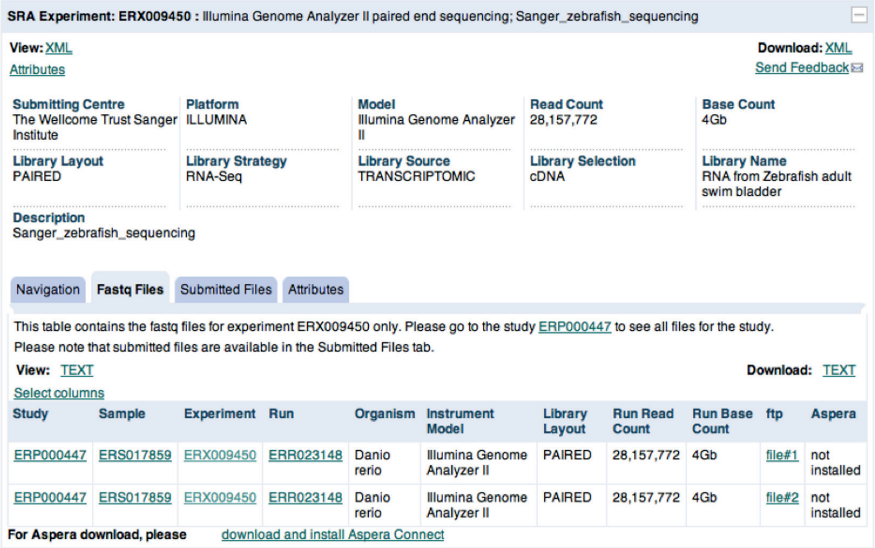


Figure G.3 public domain nucleic acid sequence data (kindly supplied by Guy Cochrane, EBI)

expression. In microarray work, thousands (sometimes millions) of genes or gene fragments can be assayed at once. This makes the work of looking at the expression of genes much quicker than it used to be, but the volume of data generated in the experimental process is very large.

The products of gene expression are messenger RNAs (mRNAs). In this process, thousands of genes or parts of genes are bound to a substrate on microarray plates. The mRNAs of interest (usually in the form of cDNA) are added to the plates and hybridise with (attach to) their complementary DNA sequence. The mRNAs are labelled, usually with a fluorescent dye, so that where they attach to a gene or gene fragment they can be visualised. The microarray plates are then scanned and the luminous dots of the fluorescent probes, which indicate where an mRNA has bound to a particular gene fragment, are recorded (Figure G.5).

Image analysis software can be used to process these findings, but very large datasets are produced as a result, sometimes hundreds of gigabytes in size. Downstream processing of such datasets requires considerable computing power. Microarray data also can be produced as text files that consist of rows and columns, sometimes in vast numbers (millions).



SRA Experiment: ERX009450 : Illumina Genome Analyzer II paired end sequencing; Sanger_zebrafish_sequencing

View: XML
Attributes

Download: XML
Send Feedback

Submitting Centre
The Wellcome Trust Sanger Institute

Platform
ILLUMINA

Model
Illumina Genome Analyzer II

Read Count
28,157,772

Base Count
4Gb

Library Layout
PAIRED

Library Strategy
RNA-Seq

Library Source
TRANSCRIPTOMIC

Library Selection
cDNA

Library Name
RNA from Zebrafish adult swim bladder

Description
Sanger_zebrafish_sequencing

This table contains the fastq files for experiment ERX009450 only. Please go to the study ERP000447 to see all files for the study.
Please note that submitted files are available in the Submitted Files tab.

View: TEXT
Download: TEXT
Select columns

Study	Sample	Experiment	Run	Organism	Instrument Model	Library Layout	Run Read Count	Run Base Count	ftp	Aspera
ERP000447	ERS017859	ERX009450	ERR023148	Danio rerio	Illumina Genome Analyzer II	PAIRED	28,157,772	4Gb	file#1	not installed
ERP000447	ERS017859	ERX009450	ERR023148	Danio rerio	Illumina Genome Analyzer II	PAIRED	28,157,772	4Gb	file#2	not installed

For Aspera download, please download and install Aspera Connect

Figure G.4 a short-read sequence record in the European Nucleotide Archive at EBI

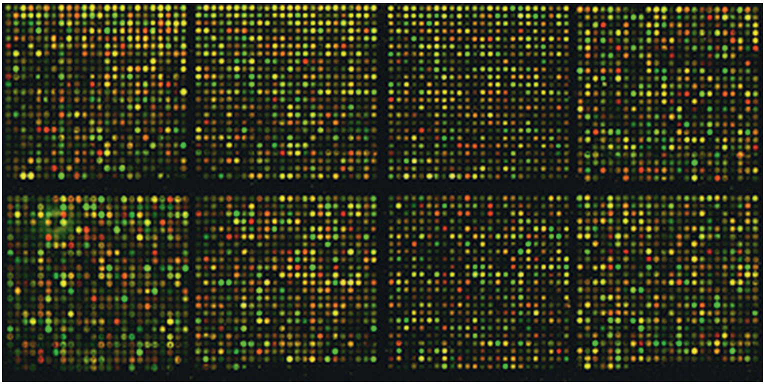


Figure G.5 Microarray plate showing gene expression differences between two muse tissues (red and green dots indicate which genes are turned on or off, yellow dots indicate that gene expression is unchanged. (Photo: Dr Jason Kang, National Cancer Institute, USA)

Genomics data are shared predominantly through a mature infrastructure of public databanks (see section 6). Researchers deposit datasets from machines as soon as is practicable, and this may be directly from the machine to the database. Large groups tend to deposit all their data while small groups are more likely to deposit a typical representative trace or other dataset because they do not have the resources to deposit the hundreds they might generate from each sample.

Summary findings are written up as articles and published in one or other of the many journals that cover health sciences. Functional genomics is a collaborative effort with application of standardised data integration, storage and dissemination policies and practices.

4.1.2 Proteomics

As well as studying the structure and function of genes, molecular biologists are interested in proteins, their molecular composition and how they function in regulating cellular processes. This field is nowadays called proteomics. Proteins are the final product of gene expression: they are composed of amino acids, the order of which in the molecule is determined by the sequence of nucleotides in the mRNA from which they were translated, and which in turn is determined by the sequence of nucleotides in the DNA of the original gene. To give a sense of the scale of the challenge, the 35,000 genes in the human genome can code for ten times as many proteins: in an extreme example, one gene can code for 1000 different proteins (in the case of genes expressed in the immune system).

Proteins can be sequenced (that is, the amino acids that compose them can be determined) by either of two methods. The most commonly used one is the Edman Degradation, a now-automated derivative of the original method developed by Edman in the 1950s. Proteins are degraded (broken down into constituent amino acids) and the individual amino acids released are assayed by high performance liquid chromatography (HPLC). The amino acids can be marked by compounds that produce a colour, enabling the presence and the amount of each amino acid to be determined.

The other method of sequencing proteins is mass spectrometry. This process is also automated. An electric current degrades the protein into its constituent amino acids or into small peptides (protein fragments) and these are identified by their individual mass. The spectra are expressed in terms of numeric data (i.e. peak intensities), as text-based data such as lists of protein IDs, or graphically. Considerable computational power is required for this process, but with improvements in this and in data storage, this technology is becoming common in protein studies. The process can be carried out very quickly – within seconds as opposed to many hours for earlier, manual methods.

Protein interaction data are becoming prominent as the functions of proteins are better analysed at the genomic or proteomic scale, and are becoming available as standardised databases. Pathways can be predicted, based on these data, to give a better picture of the combined functionality of the gene products in cells.

For most data types generated by proteomic studies, there existing public databases available for data deposition.

4.1.3 Metabolomics

Mass spectroscopy is also used for analysis of small molecules in metabolomics research, where metabolite levels in tissues or fluids are assayed. Techniques employed may be nuclear magnetic resonance spectroscopy and gas or liquid mass spectroscopy. All these tools are fully automated processes with sophisticated computational analysis at the other end of the process. For most data types generated by metabolomic studies, there are existing public databases available for data deposition.

4.1.4 Computational bioinformatics

Although the broad term ‘bioinformatics’ applies across all the technologies so far described, it can also be used in a more narrow way to describe the specific application of computer technologies to data integration, mining and other analytical practices. Such computational approaches to health research are usually termed bioinformatics, medical informatics or health informatics. Research in this area includes work on how to store, retrieve and use research data and findings, with a strong focus on manipulation of data, often collected from disparate sources, to derive conclusions or further data for further analysis.

Skills required are those of information science and software engineering as well as biomedical knowledge. Bioinformaticians may be biologists who train in computational technologies, or computer scientists or information scientists who gain knowledge and understanding of biological systems. The former pattern is more common. Either way, the field is interdisciplinary and is evolving rather fast.

4.1.5 Microscopy

Finally, there are visualisation methodologies. These include light microscopy (bright-field, phase-contrast, differential interference, fluorescence and confocal microscopy), electron microscopy (transmission and scanning types) and scanning technologies such as magnetic resonance imaging or computerised tomography. These technologies are used to study structure and function of tissues, cells or sub-cellular organelles, or to localise entities.

One example of the latter is the ability to use microscopy to map genes to specific locations on the whole chromosome set by the FISH (fluorescence *in-situ* hybridisation) technique. There are variations of this, but in essence it consists of attaching a fluorescent marker to the mRNA of interest and allowing the mRNA to hybridise to a chromosome set attached to a substrate. The mRNA shows up in fluorescence microscope images as luminous dots or bands. These may be computer-enhanced to improve the images or to enable easy discrimination between different genes where more than one mRNA has been used.

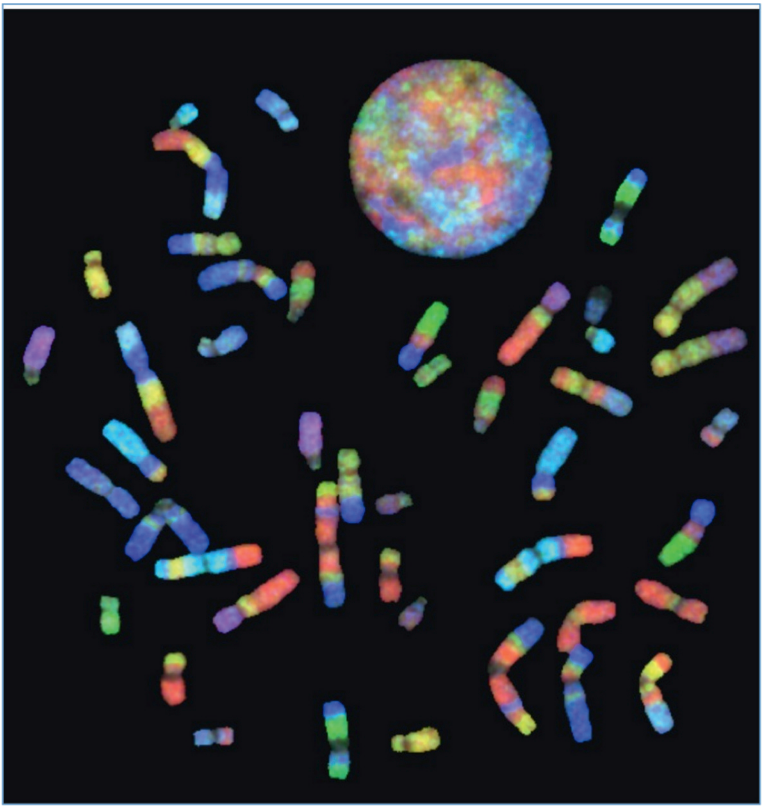


Figure G.6 In situ hybridisation of seven chromosome specific-paint probes derived from a gibbon to a set of human chromosomes (source: picture kindly provided by Dr. Fengtang Yang, The Sanger Centre, Cambridge UK)

If research is about a clinical condition, then additional techniques may be employed, such as MRI or CT (computerised tomography) scanning. Clinical imaging technologies produce large datasets delivering considerable storage and archiving challenges. Figure G.7 shows an example, a series of MRI images from a study of schizophrenia.

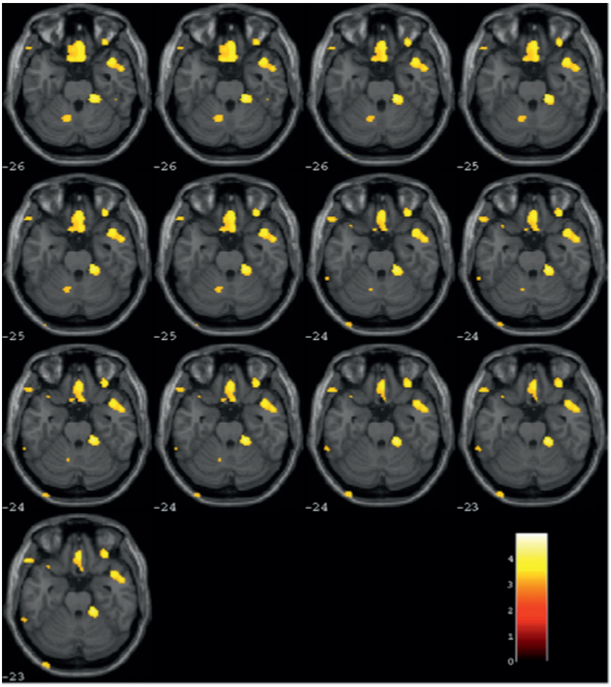


Figure G.7 group average difference map showing grey matter density in subjects in a schizophrenia study (courtesy of Professor Stephen Lawrie, University of Edinburgh)

4.2 The community infrastructure: collaborative research

The molecular biology community enjoys an extremely well-organised system for dissemination and curation of research results and through that system for connecting scientists and research groups with one another. The European Bioinformatics Institute is a focal point in the information infrastructure underlying research efforts in this discipline, providing the technologies and structural components required to collect, hold, curate and preserve research data outputs.

4.2.1 Interdisciplinary collaboration

Interdisciplinary collaborations are increasingly necessary in many fields. The growth in the application of computational technologies to bench-generated experimental data means that informatics experts from information science and software engineering are needed to complement the analytical skills of biologists.

Many larger groups now employ someone specifically dedicated to research data management, a role that requires a high level of skill and expertise: it encompasses not only data are properly stored and easily retrievable for further analysis but also of preparing data management plans when a new round of experimentation is planned, a didactic role in ensuring bench scientists understand and get optimal results from machines, tracking down data from other laboratories that may be needed for data-mining by the local research team, and writing scripts that enable best use of datasets from these machines or other research groups.

Data managers ensure best practice in the care, preservation and sharing of data, maximising confidence in biomedical data and encouraging data re-use and exploitation for new knowledge creation. This skill area is rapidly growing in importance as research becomes more data-intensive and as funders introduce formal data requirements to their funding process, and is becoming a career option for scientists with aptitude in this area.

Interdisciplinary approaches are also needed to tackle the challenges of experimental work in many areas. The increasing sophistication of light microscopy, scanners and other imaging techniques, for example, and the approaches needed to answer some of the questions at the cutting edge of cell biology and medicine, may draw on the skills of physicists, mathematicians, chemists and engineers as well as those with expertise in computational applications. This ‘systems biology’ paradigm, where scientific study in the life sciences is approached through synthetic, rather than reductionist, approaches is increasingly appropriate in responding to the scientific questions and challenges faced in the health science arena today.

4.2.2 Large-scale research and e-science

As well as interdisciplinary efforts, collaborations between research groups or laboratories are becoming more common in health sciences as major fields of research in this discipline become more data-intensive and e-science methodologies are applicable. Examples are the 1000 genomes, International Cancer Genome Consortium (ICGC) and the International Human Epigenome Consortium, in all of which the EBI are one of many global collaborators. It is useful to provide a short description of each of these to demonstrate the nature of the initiatives and to illustrate that collaborative approaches are necessary to deliver such ambitious and labour-intensive results.

The 1000 Genomes Project[10] is aimed at sequencing the entire genomes of 1000 people, in order to find most genetic variants that have frequencies of at least 1% in the populations that these individuals represent. The project has a steering committee of twenty-four scientists and a panel of several hundred scientists contributing to the laboratory work.

The goal of the International Cancer Genome Consortium[11] is to obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumour types and/or subtypes which are of clinical and societal importance across the globe. It coordinates a number of projects (35 at the time of writing) across the world with the aim of developing comprehensive catalogues of genomic abnormalities in these tumours and will make the data available to the entire research community as rapidly as possible, and with minimal restrictions, to accelerate research into the causes and control of cancer.

The International Human Epigenome Consortium[12] coordinates epigenetic mapping projects (projects that study the organisation of the genetic material in the cell and how this organisation affects gene expression and the control of cellular functions) worldwide. The aim is to prevent redundancy and duplication of effort and to implement high data quality standards, to coordinate data storage, management and analysis and to provide free access to the epigenomes produced.

In some areas of life science research, grid technologies are now warranted and used. This is particularly the case in fields where imaging technologies are intensively used. The neuro-imaging case study described in section 3.4, for example, has become a collaborative effort involving six or seven laboratories across the UK. The project has amassed clinical and demographic data on a terabyte scale, in the form of millions of individual files. Data management on this scale is a major undertaking.

Collaborations may also arise where large amounts of funding are needed for one piece of work, where individual expertise or technologies need to be pooled to answer a research question, or where large volumes of data need to be gathered for meta-analysis. In such cases, collaborative efforts may be transient, lasting only for the length of time needed to achieve that immediate goal, or they may persist for many years with repeat funding being attracted for further collaborative work. An example is the Bloodomics project[13].

[10] http://www.1000genomes.org/home

[11] http://www.icgc.org/content/icgc-home

[12] http://www.ihec-epigenomes.org/

[13] http://www.bloodomics.org/

4.3 The temporal infrastructure: research life cycles

While in some areas of health science (for example, epidemiology) the research life cycle is a long one, life cycles in many of the fields of focus here are relatively short. A DNA sequence run can be completed and the results deposited in a public databank for others to use within a few hours. The other biochemical analytical techniques described can also be carried out in a matter of hours or days.

Cytological experiments may require some days of specimen preparation and microscopy, plus more for computer manipulation of the resulting images to maximise the usefulness and clarity of the results. Large-scale clinical studies necessarily take much more time, though, sometimes requiring years to collect sufficient data for analysis. And computational (informatics) research is variable depending upon the complexity of the questions to be answered and the data to be manipulated.

It should be noted that for all these data-intensive research activities considerable curation, technical or computational/algorithmic expertise and effort are also required to ensure that datasets are usable by the research group that produced them and by others, and that these datasets are accessible and re-usable in time to come.

The research life cycle model related to knowledge creation and dissemination developed by Charles Humphrey[14] is useful here (Figure G.8).

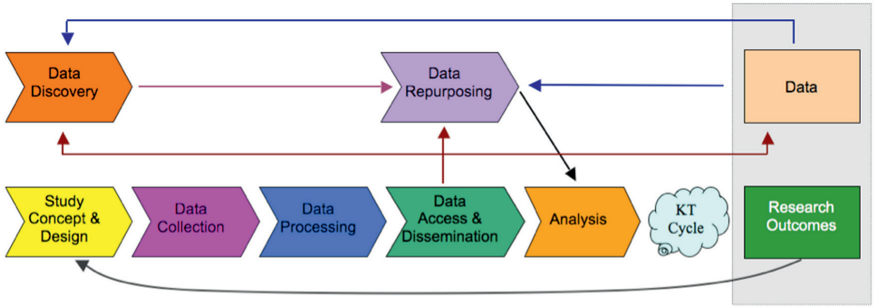


Figure G.8 The life cycle model of research knowledge creation (Humphrey, 2008) [‘KT Cycle’ is the Knowledge Transfer Cycle]

Experimental research is largely represented by the bottom set of activities in the diagram. The top set represents informatics/e-science approaches,

[14] Humphrey, C (2008) e-Science and the life cycle of research. http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc

where experimental data are re-purposed and analysed to create new data (which may themselves be re-purposed and analysed).

The key issue with respect to data services is that the community both contributes to and interacts with them. The community creates the databases, and EBI is the custodian of those resources, curating the data to some degree as part of that custodial role. Some databases get only light curation (such as the sequence databases, where a fairly simple metadata check suffices), but others are heavily-curated (such as UniProt, where EBI curators search for articles about a gene, find evidence on it and add that to the UniProt database). It is worth saying at this point that data curation at this level has become an established career option, emphasised by the fact that there is a now professional society of curators.

As well as curators, the rest of the community accesses and uses the databases in specific ways. Users of different types interact with the established databases differently. Essentially, there are:

- Power users: those who routinely download a database in bulk to carry out, for example, whole-genome analysis. These users work using FTP and web services to access the volumes of data they need
- Biologists: those who need to access and use relevant, small-volume data for their work; for example, someone who is working on a particular disease and needs to check on sequencing data for genes that are implicated in the disease
- Occasional users: for example, teachers, students, editors, citizen scientists and so forth, who may on occasions wish to view or analyse a dataset

In summary, the life science databases are growing, developing entities. They form a hub for the community's activities, but that hub is dynamic. Users not only take and use the data, they provide feedback on the service, their requirements change, research develops in new directions, and the services evolve accordingly.

4.4 The skills and training infrastructure

Basic and clinical research practices and technologies are acquired in the usual way through postgraduate and postdoctoral training. Technologies may be complex to master and require significant intellectual effort as well as practical skills. Some cytological techniques require exceptional dexterity.

The importance of informatics expertise in many areas of health science research imposes a new requirement on researchers who have been trained in conventional biological methodologies and approaches.

Many biologists learn 'on the job' and pick up coding skills where needed. The past decade has, however, seen rapid growth in masters level training in informatics (bioinformatics and cheminformatics, especially) and many young health scientists are entering the field with these qualifications. In addition, some library and information science schools are offering new modules or courses in data management, something that is expected to become a career option for librarians in the future.

Data management roles are also becoming more common within research groups in response to the increasing data-intensity of research and the requirements of funders for curation and preservation of datasets. Where these posts exist, they may be occupied by senior researchers who have shown interest in and aptitude for the role. These people are normally called ‘data scientist’ or similar, and may be distinguished from data managers by differences in their role and position. The terminology is extremely fuzzy at the moment, but we distinguish between data scientists and data managers simply on the basis of the set of tasks that they carry out and the overall objective of their job:

- Data science: the conceptualisation, creation, use and appraisal of data, the selection of data for re-use, and the application of tools to re-use and exploit data
- Data management: a specialist area of computational science – database technology – which focuses on ensuring that data produced and needed by the researchers are properly stored, curated and preserved. Included here (but not exclusively) is the work carried out in data centres or professional databanks. Often computer scientists with biomedical research background not required

Where there is not a discrete data scientist role within a research group or laboratory, various members of the research group may undertake data-related tasks. They may or may not be formally trained in the skills required, but there is growing attention to this within the community and summer schools and short training programmes covering specific areas of health science data manipulation or informatics are becoming more common. Almost all biologists working in informatics-based fields are able to write scripts that enable them to use datasets from other laboratories or to mash together datasets from different machines. Computer scientists may be employed to bring their software or databasing skills to biomedical research teams. In this case they must assimilate the biological domain knowledge that they need ‘on the job’.

5 Current status of Open Access to the research literature

5.1 The policy foundation for Open Access to the biomedical literature

The first Open Access policy that covered any biomedical/health science literature was the institutional mandatory policy at Queensland University of Technology, Brisbane, Australia, in 2004. This was followed the same year by a second institutional mandate at the University of Minho in Portugal.

These institutions began a trend that has continued. At the time of writing there are 117 institutional policies on Open Access to journal articles and 30 sub-institutional ones (departments or schools within universities or research institutes), including Harvard Medical School.

Several research funders have also adopted Open Access policies and mandates, lead by the NIH in the USA and the London-based Wellcome Trust, and it is this that has affected significantly the proportion of the health sciences literature that is now openly available. Of the 47 current mandatory policies from research funders, 22 are from funders of health-related research. The list includes national research councils funding health research in Australia, Canada, Ireland, UK and USA, and around a dozen medical charities in these countries plus Italy. Researchers supported by these funders are required to deposit their articles in institutional or subject-specific repositories (such as PubMed Central and its growing national/regional variants).

In addition, the European Research Council has a mandatory policy requiring Open Access to outputs from research that it funds and some of this falls under the health sciences banner. The European Commission has a 'pilot' mandatory policy covering 20% of the current FP7 research programme, and the health research programme is included in this 20%. This will expand to cover 100% of EU-funded research.[15]

The National Institutes of Health (NIH) in the US, the largest funder of research in the world, alone covers some 90,000 journal articles published each year from research that it funds. The NIH policy originally began as a voluntary one that requested grant-holders to deposit copies of their journal articles in the PubMed Central repository. After two years under this policy, only 5% of relevant articles were deposited by their authors voluntarily. Publisher deposits raised the total to 19% but this was still disappointing.

In 2008, the US Congress instructed the NIH to make the policy mandatory, the result of which is that the percentage of articles being deposited climbed to over 70% in 2009. This was the strongest possible evidence that a policy

[15] http://www.youtube.com/watch?v=GIU14-3hYto

must be mandatory to work effectively: since that time new policies from health funders have been mandatory and existing policies based on voluntary action by authors have been revised to make them mandatory.

There is considerable evidence from other quarters that also supports the necessity for the mandatory nature of a policy on Open Access. The earliest study to produce data on this was by Sale (2006), who looked at the repositories of three Australian universities. Sale showed that the accumulation of Open Access articles at the University of Tasmania, where there was no policy but there was some active advocacy on OA, was extremely slow. At the University of Queensland, where there was a policy encouraging authors to deposit their work, supported by active advocacy from the library, the accumulation rate was higher. A fast rate of accumulation of content was seen, however, at Queensland University of Technology, where there was both active advocacy and practical support from the library *and* a mandatory policy.

This evidence that only mandatory policies work effectively was compounded by a recent study by Gargouri *et al* (2010) that compared, amongst other things, the level of OA articles in university repositories that have mandatory policies with the general level of accumulation of articles in non-mandated repositories. This 'control' level of accumulation is around 15% of total outputs from the institution, whereas mandated repositories are collecting 60% of total outputs.

Most policies accommodate a short embargo period, usually 6 months but in some cases 12 months, to enable publishers to continue to operate their subscription sales model. The argument is ongoing about whether even a short embargo is detrimental to research progress: certainly the speed at which research moves in health sciences means that a delay in access by medical researchers and medical practitioners outside of research institutions, patients, therapists and research-based small companies that need this research to innovate may all be disadvantaged by having to wait for access to new research findings.

Further policies from other funders are expected as the benefits of opening up the biomedical literature become more apparent. Additionally, there is continued growth in institutional policies: these help to boost the Open Access corpus in biomedicine incrementally.

5.2 Open Access to the research literature

5.2.1 Open Access repositories

Health sciences is one of the few disciplinary areas of research where extensive, subject-based repositories of Open Access material exist. The fact that there is such centralised infrastructure reflects available funding, the criticality of research in this discipline and the ability of the discipline to organise around coordinating bodies.

PubMed Central (PMC) was established in the US in the year 2000, with the contents of just two journals in the repository. Within two years it covered 55 journals and numbers have been growing ever since. The database currently has around 2 million full-text journal articles and receives the full contents of 600 journals as well as manuscripts deposited by authors. All are free to access and read, but only about 11% fall under the strictest definition of Open Access by being distributed under a Creative Commons (or Creative Commons-type) licence that permits more liberal re-use.

The NCBI (National Center for Biotechnology Information), which manages PMC, has added many features over the decade, including a good search function, linking between articles, and between articles and other types of content such as commentaries and books. More features are planned. Such features enhance the user experience and utility of the database.

In 2007, the first international PMC (PMCi) was established in the UK by the Wellcome Trust and a consortium of other research funders. This repository, UKPMC, collects articles in biomedicine from UK scientists and shares content with PMC itself. This is the first of what may be many PMCis: already a Canadian site has been announced, with discussion of additional sites in other regions, including the possibility of transforming the UK site into a European PMC.

UKPMC launched with the intention of providing cutting-edge services to researchers and has already broken new ground by creating XML-marked-up texts that are amenable to data-mining and text-mining. UKPMC is already being used by the UK's National Text-Mining Centre (NaCTeM) and the European Bioinformatics Institute (EBI) to extract facts, concepts and relationships from the literature within UKPMC and combine them to create new knowledge. UKPMC has also developed collections of other types of content, such as clinical guidelines and project grant details. It also enables users to cross-search PMC alongside CiteXplore, an indexing service that covers a number of other large research databases.

This informatics work is laying the foundations for a future where interoperability is truly achievable between Open Access research collections. Indeed, UKPMC is also developing ways to support UK research institutions in their quest to fill their own institutional repositories. The policy from many of the UKPMC funders is to require deposit of research outputs directly into UKPMC itself, thus conflicting with policy requirements of many UK universities that require deposit into the local institutional repository. To obviate the need for researchers to deposit in both services, UKPMC will serve articles to the institutions from which they originate for population of the local repository.

A well as these centralised collections of Open Access content, health science literature is accumulating in institutional repositories. Mandatory policies on these, of course, cover all research carried out in those institutions and are thus essential in 'sweeping up' outputs from unfunded research and from research not covered by funder mandates.

5.2.2 Open Access to journal articles

Health sciences are also well-represented in Open Access journals.

The largest open Access publisher, BioMed Central (now part of the Springer science publishing organisation), specialises in biomedical research, as is obvious from its name, though it does also now cover some chemistry and mathematics too. It publishes some 210 journals, most of which are in biomedicine. BioMed Central deposits all its journal articles in PMC at the time of publication as well as hosting them on its own website.

The Public Library of Science, another leading Open Access publisher, has not only developed some very high quality journals in biology and medicine (*PLoS Biology* and *PLoS Medicine*, plus others) but has changed the shape of publishing through *PLoS ONE*. This is a journal that covers all the natural sciences. It introduced a new system of quality control, still based up on peer review, where referees are asked to judge an article purely on the basis of whether the work has been carried out in a sound scientific manner. Judgments about its relevance, significance and impact are made through community response post-publication. The model has proved very successful and has recently been emulated by the Nature Publishing Group with the launch of *Nature Scientific Reports*[16].

The Scielo (Scientific Electronic Library Online), a collection of peer-reviewed Open Access journals published mainly from South American countries in Spanish or Portuguese, covers over 800 journals. Of these 45 are in biological sciences and 261 in health sciences, representing a large part of the Latin American biomedical literature.

Bioline International, a service that provides a free electronic publishing platform for small publishers wishing to publish Open Access journals in the biosciences, has over 50 journals in its collection, all from developing and emerging countries, covering biomedicine and agriculture.

The Directory of Open Access Journals[17], a listing of Open Access journals from around the world, currently details 709 journals in its 'health sciences'

[16] http://www.nature.com/srep/marketing/index.html

[17] www.doaj.org

category (covering medicine, dentistry etc) and a further 217 in its 'biology' category. This list overlaps with the specific services mentioned above.

In addition to the fully Open Access journals, many publishers have now offered a so-called 'hybrid' Open Access option, whereby authors can pay a publication fee and have their article made Open Access within an otherwise subscription journal. Take-up on these options is not high, largely because of the level of fee, and it should be noted that many journals offering this option do not make the articles available under a liberal licence, meaning they are free to access and read but often not to re-use in other ways, including computing upon them.

5.2.3 The proportion of Open Access literature in health sciences

Some attempts have been made to measure how much research in total is available in Open Access, and to break this down by discipline in some cases.

Björk and co-workers estimated that in 2008 using a sample of almost 1850 articles, that 20.4% of the total literature was available in some form of Open Access (in OA journals, in repositories or on author websites) (Björk *et al*, 2010). This compares to a previous study by the same authors of the situation in 2006 (Björk *et al*, 2009) that found a total of 19.4% of the literature to be Open Access. The difference is within confidence limits.

Hajjem *et al* (2005), using a sample of 1.3 million journal articles, found that the proportion of Open Access articles varied between disciplines from 5% to 16%. A later study by [Gargouri *et al*, 2010] from the same group found the OA share overall to be 20%, with biology scoring 21% and clinical medicine 3%. Note that this study used only 'green' OA (articles self-archived by their authors in repositories, including PubMed Central, not those published in 'gold' OA journals). The latest estimate by this group of the percentage of research openly available through repositories is 20-22% and, if Björk's estimate of the percentage available through journals ('gold' Open Access) is added, the total is currently about 30%[18].

Matsubayashi *et al* (2009) studied the discipline of biomedicine specifically and found the OA availability of articles to be 26%.

The findings from studies so far are summarised in Table G.1.

The more recent of the two studies by Björk *et al* showed that for the fields of medicine, biochemistry/genetics/molecular biology, and 'other areas related to medicine', the proportion of OA articles in Open Access journals was higher than that in repositories. This position is reversed for all other fields in this study, presumably reflecting the domination of the 'Gold' Open

[18] Stevan Harnad and Yassine Gargouri, personal communication (to be published shortly)

Table G.1 Open Access availability of journal articles

Study	**Overall % OA**	**Specific fields % OA**
Björk *et al* (2009)	19.4	n/s
Björk *et al* (2010)	20.4	Medicine 21.7% (13.9% OA journals, 7.8% OA repositories) Biochemistry, genetics, molecular biology 19.9% (13.7% OA journals, 6.2% OA repositories)
Matsubayashi *et al* (2009)	n/s	Biomedicine 20%
Hajjem *et al* (2005)	n/s	Biology 15% Health 6% Psychology 7%
Gargouri (2009)	20.0	Biology 21% Clinical medicine 3% Health 18% Biomedicine 11% Psychology 25%

(ns = not studied)

Access journal-publishing arena by biomedical journals. The data from Björk *et al* (2009) on this point are shown in Figure G.9.

5.3 New developments in dissemination in health sciences

The vision of enhancing the traditional scientific article (or book) has been developing over the past few years. The Web provides the opportunity to link an article written and presented in the traditional format with supporting data, commentaries, similar articles, datasets in public databanks and so on. Semantic technologies now hold the promise of creating a scientific Web that is linked by meaning and context, with all research outputs fully and meaningfully linked to one another.

Traditional forms of peer review, sequential publishing, the way rights and ownership of knowledge are managed, and the emphasis on disseminating scientific findings in the form of reports 'crystallised' in time are likely to metamorphose into a system that makes maximum use of the opportunities offered by the Web and optimises research communication.

An FP7-funded project, Liquid Publications[19], has been investigating options and developing 'liquid journal' and 'liquid conference' use cases. A position paper defines some of the conditions and the benefits of liquid publishing:

[19] http://liquidpub.org/

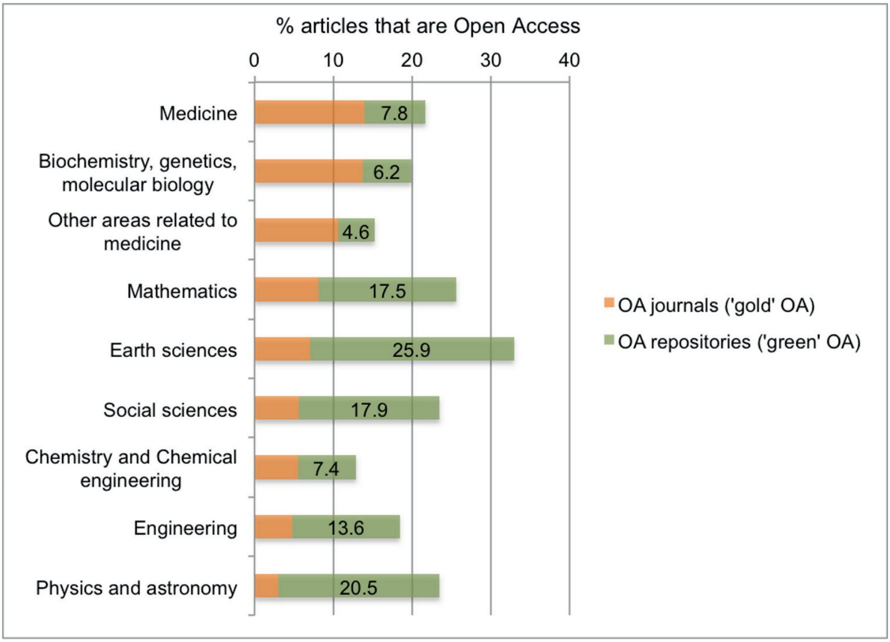


Figure G.9 Percentage of Open Access articles by discipline and mode of dissemination (data from Björk et al, 2010)

real-time dissemination of findings, encouragement of early sharing of results and ideas with reward systems in place to benefit researchers who maximise this behaviour, the concomitant increase in collaboration that will arise from early and widespread dissemination, and lightweight and real-time assessment and evaluation of findings.

While such as system remains to be achieved, there have been steps taken towards developing an improved, linked scientific knowledge base. The far-seeing work of UKPMC in establishing a system that ensures material ingested into the repository is in XML and marked-up for semantic data-mining and text-mining tools is one important advance, and it is taking place in the health science discipline. That repository also provides the means for supporting datasets to be deposited and linked to articles, thus taking another step towards a properly-linked scientific corpus.

Publishers have also been active in this area. In 2007, the Public Library of Science launched PLoS ONE, a broad-scope Open Access journal covering the whole of science. PLoS ONE is interactive, providing the means for readers to post comments and discuss articles, and it also incorporates various additional features including a range of article-level metrics that inform the author about the usage and impact of their paper. Two years ago, Elsevier Science released prototypes of what it called 'the article of the future'[20], which was hyperlinked (mainly to other articles) and contained embedded video and audio files, and some integrated social media tools.

These are small steps and ones still bounded by the limitations of the traditional model of a scientific paper. The concept of true liquidity is a different level altogether. Nonetheless, the experimentation is commendable and the incremental advance is welcome. These moves signal something that will be of fundamental importance for the efficacy of the future scientific communication system.

In this scenario, biological databases are, and will be, also critical. Each database release is equivalent to 'publication' and most databases are highly fluid in the sense that sequences are modified as more alignments become available. These modifications change the data for all downstream databases and this realises/synchronisation process plays an important role in maintaining data consistency.

5.4 Open notebooks

Open notebooks – the open dissemination of the day-to-day experimental activities in the laboratory – have become fairly common in certain fields. Scientists record their experimental procedures and results and publish them on the Web, usually in blog form[21].

The idea of this form of communication is to speed up scientific endeavour in a field, to gather feedback from the relevant community, to engage the community in general discussion about the ongoing work, and to capitalise on the collective wisdom of the crowd.

The discipline where the use of open notebooks is furthest advanced is chemistry. This is partly because early-adopters of the concept were chemists and partly because chemistry is largely free from the issues of concern that health scientists may have about the practice.

There are two main areas of concern expressed by researchers in various fields of health science. First, there is concern about patient confidentiality and the imperative to safeguard patient anonymity. Second, there is the risk of releasing early data or information that subsequently turns out to be inaccurate but which, if used in the meantime, may have damaging consequences for people.

[20] http://www.elsevier.com/wps/find/authored_newsitem.cws_home/companynews05_01279

[21] For example, the chemist Cameron Neylon's open notebook:http://biolab.isis.rl.ac.uk/camerons_labblog

As a result, open notebook science may not be a concept that translates well to some areas of health science research. Nonetheless, there are areas, such as the molecular biology fields, where the two concerns have little relevance. In these cases, open notebook work or something akin to the concept, are used in practice (see section 5.3).

Implications for OpenAIRE

- Mandatory policies from the European Commission and from a growing number of health research funders and institutions will increase the amount of health science literature accumulating in Open Access repositories across Europe for harvesting by OpenAIRE
- Continuing development of OpenAIRE policy on content acquisition may wish to consider whether to enrich the resource by harvesting health science material from OA journals and repositories, or by linking to that content in its original locations
- OpenAIRE will need to consider whether to mark-up and enhance the content it harvests from institutional and other repositories so as to provide the functionality for the future that UKPMC (and the European PMCthat UKPMC is planned to be transformed into) is delivering.
- OpenAIRE should keep a watching brief on how the concept of open notebooks and similar initiatives develop in health sciences. There is an opportunity for enriching OpenAIRE content by linking to these things but the implications of that in management overheads could be significant

6 Current status of Open Access to research data

6.1 The policy foundation for Open Access to biomedical data

Policies on research data in biomedicine have been accumulating for some years now. The discipline is fairly well-advanced in this respect.

A requirement for the inclusion of a data management plan, including details of how data will be stored and managed for a future period after the cessation of funding, has been part of the policy of a number of large research funders for some time, although there are many funders with a mandatory policy on Open Access to the literature that do not have an accompanying one on data.

There are several permutations on funder position. They may have an OA policy on both literature and data, or just literature. The data policy may include the requirement for a data management plan (including how data may be shared and how they will be cared for in the longer terms as well as through the lifetime of the project) or not. The funder may provide guidelines or rules about data sharing and curation, or not. And the funder may specify detail, such as the period of time within which data must be deposited in an Open Access location, or not.

The European Research Council (ERC), for example, has a data archiving/sharing policy that requires grant-holders to make their data available for others and that this occurs within a period of 6 months after the completion of the project[22]. The data (such as 'nucleotide/protein sequences, macromolecular atomic coordinates and anonymised epidemiological data') must be placed in an appropriate public databank. Examples of appropriate databanks are given as GenBank and PDB (Protein DataBank).

In the UK, the Biotechnology & Biological Sciences Research Council (BBSRC) has a similar policy to this, as does the Fonds zur Förderung der wissenschaftlichen Forschung (Austrian Science Council). Other European biomedical funders – the Wellcome Trust, Cancer Research UK, the Medical Research Council (UK), and the Országos Tudományos Kutatási Alapprogramok (Hungarian Scientific Research Fund), along with funders outside Europe such as the NIH, National Science Foundation (NSF; US), Canadian Institutes of Health Research, Gordon & Betty Moore Foundation, Heart & Stroke Foundation of Canada, Michael Smith Foundation for Health Research and the Ontario Institute for Cancer Research, have data access policies without any stipulation of how much time must elapse before researchers make their data available.

To an extent, this is to accommodate a variation in needs between disciplines: full data exploitation by their creators takes much more time in some fields, such as epidemiology or certain clinical areas, than in genomics or proteomics, and funders wish to allow researchers sufficient time to carry out all the analyses they want before sharing their data openly. There is, however, an argument for reasonableness and most funders would not expect data to be withheld for a decade or more.

6.2 Formal infrastructure for sharing research data

In the life sciences, data sharing is mature in many areas of health sciences, notably those focused on in this chapter. This situation has evolved because an open approach is the only one that could enable the challenges of modern life science research to be tackled, based as it is on analytical and/or comparative approaches and intensely data-rich. Without the development of a

[22] http://erc.europa.eu/pdf/ScC_Guidelines_Open_Access_revised_Dec07_FINAL.pdf

formal infrastructure for data sharing, this research simply could not happen. Public databanks for various types of biomedical data are, as a result, long-standing and the organisational infrastructure to support these in the long term is established in many cases.

6.2.1 Large public databanks

Funding for these organisations and their ongoing work is from national and regional funders. The main players are as follows:

- NCBI (National Center for Biotechnology Information), established in 1988 as a division of the National Library of Medicine at the NIH, Bethesda, USA
- EBI (European Bioinformatics Institute), part of the European Molecular Biology Laboratory, based at Hinxton, UK
- Center for Information Biology and DNA Data Bank of Japan (CIB-DDBJ), established in 1987 at the National Institute of Genetics, Yata, Japan

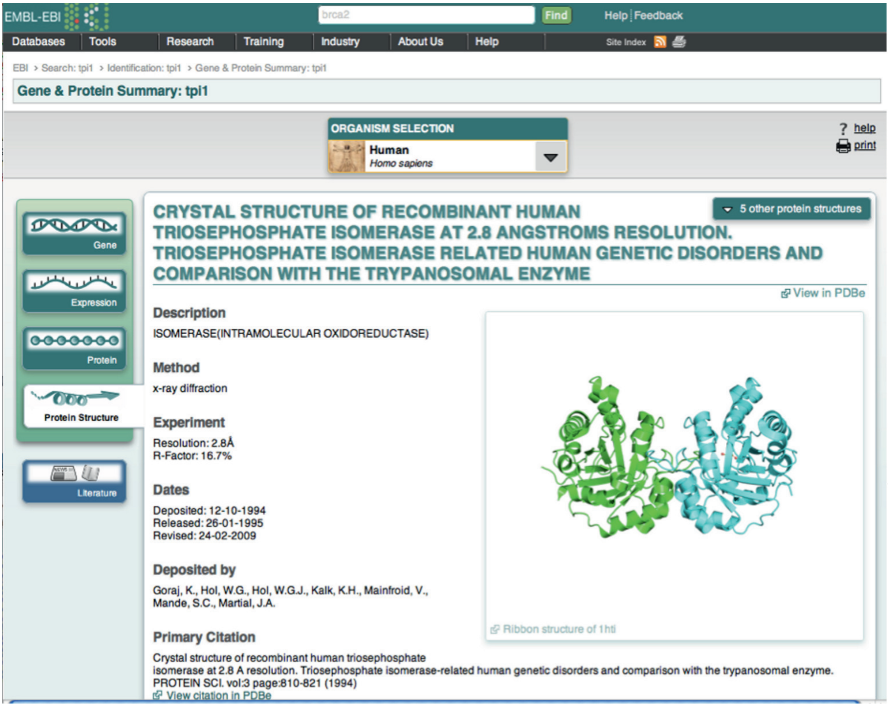
These three organisations curate and store biomedical data in a number of individual databases. They exchange and share data and researchers typically upload their data to, and download data from, the nearest site geographically. The original formal model for data exchange between sites was for nucleotide data resources[23], but this was such a successful example of formal data sharing on an international scale that it has now been followed by others.

Access is not enough in many cases, however. Tools for accessing data are also needed and the data custodians play a role here, too. They may carry out data assembly (put together multiple datasets to make a whole genome to save individual researchers having to do this), provide tools for searching and analysing datasets, and integrate data from other sources into the database (such as information on genes from journal articles). The outcome is a rich resource composed of standardised data elements, maximising the value to users. Figure G.10 shows an example of the multiple-view facility offered by EBI for a gene called Tpi1.

Curating life science data is not only about collecting and storing datasets and the operation of these large public databanks is sophisticated.

The metadata requirements are often demanding, ensuring that re-usability of the datasets is optimised. Researchers must follow strict rules on data structure and metadata entry when uploading datasets. The databanks employ a body of professional, highly-skilled developers and curators who check entries and will correspond with depositors if there are errors or inconsistencies in the metadata.

[23] For example http://insdc.org/



Gene & Protein Summary: tpi1

CRYSTAL STRUCTURE OF RECOMBINANT HUMAN TRIOSEPHOSPHATE ISOMERASE AT 2.8 ANGSTROMS RESOLUTION. TRIOSEPHOSPHATE ISOMERASE RELATED HUMAN GENETIC DISORDERS AND COMPARISON WITH THE TRYPANOSOMAL ENZYME

Description
ISOMERASE(INTRAMOLECULAR OXIDOREDUCTASE)

Method
x-ray diffraction

Experiment
Resolution: 2.8Å
R-Factor: 16.7%

Dates
Deposited: 12-10-1994
Released: 26-01-1995
Revised: 24-02-2009

Deposited by
Goraj, K., Hol, W.G., Hol, W.G.J., Kalk, K.H., Mainfroid, V., Mande, S.C., Martial, J.A.

Primary Citation
Crystal structure of recombinant human triosephosphate isomerase at 2.8 A resolution. Triosephosphate isomerase-related human genetic disorders and comparison with the trypanosomal enzyme.
PROTEIN SCI. vol:3 page:810-821 (1994)
View citation in PDBe

Figure G.10 Search results from the new EBI site search. Introductory information for genes (in this case, Tpi1) is shown in gene-, expression-, protein-, structure-, and literature-centric page views.

In some cases – such as microarray data – there is a further problem in that the data are meaningful only in the context of the particular individual sample used. Annotation therefore requires details of the experimental conditions and the gene name, but gene names are not yet fully standardised. Ambiguities in expression data and the possibility of many-to-many relationships between genes compound this problem. There are international efforts to establish ontologies and other standards that will resolve this. The MIAME standard (Minimal Information About a Microarray Experiment), developed at EBI and others, is a major advance here.

Moreover, primary data are often not sufficient for the type of work that needs to be done. The data curators therefore add value – considerable value in many cases, making data more accessible and more usable for the different constituencies that will use them: the curation process may include data integration, which is an area where much effort is expended and which still presents many challenges.

The process can be thought of as tiered, with some databases undergoing more curation than others, and curation can be by human or machine:

- Preliminary data sets: examples of these are genomic short-reads (short fragments of DNA whose sequence is ascertained)
- Primary data records: for example, a whole-gene sequence
- Computationally-annotated data records (for example, assembled-sequence records, where a computer has put together separate sequences to construct something much large, perhaps even a whole genome of an organism)
- Curated data records (for example UniProt or RefSeq databases where human curation in the form of the searching out and adding of further contextual information, perhaps from journal articles or other sources, has taken place)
- Curated 'views' (for example the Reactome database, where expert curators construct biological pathways using data from many sources)

6.2.2 Small public databases

As well as the large databanks described in the previous section, there are thousands of small, specialised databases that are made openly available. Most of these are hosted on the websites of research groups or specific research projects. Examples are databases containing sequences from the genome of a single organism or information about single genes or gene families.

The journal *Nucleic Acids Research* (an Open Access journal), publishes a list of databases in molecular biology each year: the latest one features over 1300 of them[24]. This listing only covers molecular biology: outside of this field there are many more public databases covering diseases, therapies, diagnostics and so on. The discipline overall is rich with information.

The problem that can arise is one of sustainability. Many of these small databases are supported by project funding and when the project comes to an end, the database may no longer be updated or curated and may even disappear.

6.2.3 Journals

A number of major journals, particularly in molecular biology, have policies that require authors to make their data freely available to others when a paper is published.

[24] http://nar.oxfordjournals.org/content/35/suppl_1#EDITORIAL

The usual way of ensuring that this is done for data that should be in a public database is for the journal to require the accession number of the dataset, proving that it has been deposited in a database and providing the direct link to the dataset.

This is not a foolproof system, however. One study looked at the level of compliance with journal policies by checking for datasets that should be in GenBank. It showed that 9% of articles did not cite accession numbers of the datasets, even though the datasets were in GenBank (Noor *et al*, 2006). A further 7% had not submitted the datasets to GenBank.

Nonetheless, these are small percentages. The norm in molecular biology fields, at least, is to make data available for sharing. In these fields data sharing is relatively non-contentious because the purpose of the experiment is often to determine a genetic sequence and thus the scientist has achieved his/her objective by doing this and publishing the result.

For other types of research data and outputs, such as software, the journal itself may host the outputs on its website. This may apply quite widely: data from PubMed Central show that 25% of articles published in 2009, for example, have supplemental datasets attached to them. In such cases, reviewers may reject a paper if the supporting material does not accompany it.

One reward that researchers may enjoy from sharing their data is increased impact for their research. Piwowar *et al* ((2007), examining citations to microarray clinical trials, demonstrated that articles where supporting datasets had been made publicly available enjoyed an average 69% increase in citations compared to articles with no available data.

In fields where data take considerable time to analyse and exploit, there tends to be more reluctance on the part of researchers to relinquish their data to the community within a short period. Funders understand that there are significant cultural differences between research communities on this issue and generally word policies to accommodate these differences.

6.3 Informal infrastructure for sharing research data

Informal data sharing also goes on in health sciences, most commonly in fields outside of molecular biology. Research groups will usually supply data if asked by another group (though the data may not always be in a usable format). Such negotiations may also lead to more fruitful engagement in terms of formal collaboration or joint publication.

In many instances, though, data remain stored locally and never shared. Or there may be an attempt at sharing, by including some data in published articles, though in practice this makes access and re-use by third parties extremely difficult. A table of transcriptomics data in the published PDF file of a journal article is effectively unusable without a great deal of work in manually transcribing the table contents into a software programme that can manipulate and analyse the data or integrate them with other datasets for further analysis.

What is clear is that large amounts of data that are unsuitable for the big public databank services (either because the datasets are too small or they are not an appropriate type) remain stored locally by their creators on hard drives or portable media. Discovery of these datasets is almost impossible when metadata are not made available on the Web and so they languish unused when they might be exploitable by others.

Institutions are rising to this challenge to a degree. There has been considerable work in the library community to scope and analyse the needs here, and some studies have gone some way towards providing a cost analysis and guidelines on practice for institutional efforts to preserve and curate research data[25]. There is, however, a clear need for this situation to be clarified and acted upon at a European level, and OpenAIRE may take a leading role here.

The diagram below shows the overall picture with respect to the literature, and data creation, manipulation and management, in the biomedical domain.

Implications for OpenAIRE

- Mandatory policies from the European Commission and from a growing number of health research funders and institutions will increase the amount of health science data accumulating in databases and repositories across Europe
- Continuing development of OpenAIRE policy on content acquisition may wish to consider whether to enrich the resource by harvesting health science data from institutional repositories, or by linking to that content in its original locations
- OpenAIRE might consider providing a storage and curation service for research datasets that are not suitable for the professional databanks but that should be made openly available. This service should be offered for all publicly-funded research, not only Framework Programme research

7 Challenges and opportunities

As mentioned in the Introduction to this chapter, research data management in the life sciences is comparatively advanced. Many basic principles of good practice and infrastructure development have already been established. The

[25] For example, one guideline for preserving and curating data is the Data Sea of Approval: http://www.datasealofapproval.org

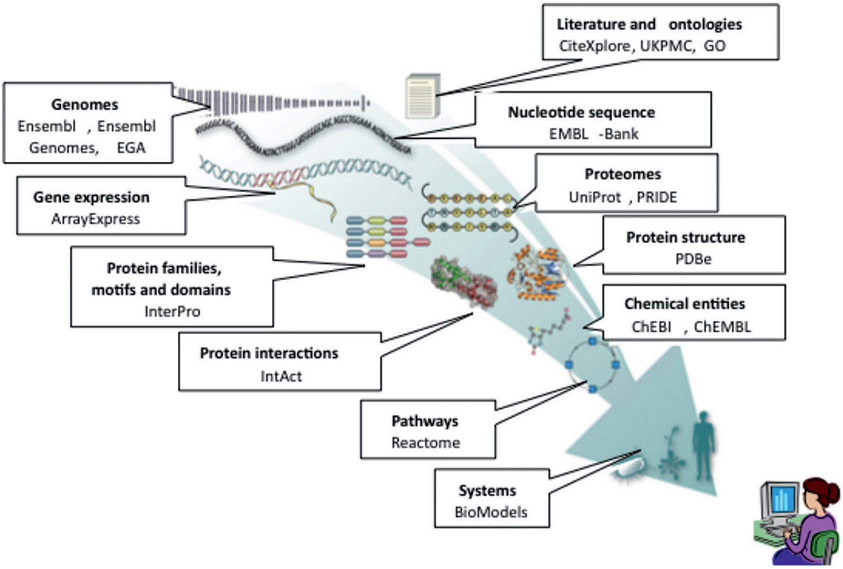


Figure G.11 The biomedical data arena

field is well-regarded by many as a worthy example of how to organise on a community level and put in place solutions that work for all. That does not mean that all problems have been solved: many remain, not least how to plan for and deal with the longer term challenges of data management.

The main information-related opportunities and challenges in health science research that are now receiving attention are as follows:

- Managing the increasing volumes of data generated. This introduces challenges in terms of providing access, storing and preserving the data
- Making cost / benefit assessments of data storage and preservation processes, so that decisions on what to keep and how are arrived at objectively
- What to do with 'small' datasets that are too small for the professional databases and will require manual curation
- Cost / benefit considerations for data curation: what do we want to keep / document?
- How to link data sets to core databases and the literature to create additional value

- How to resolve a number of generic issues around standardising metadata for discovery, data documentation and packaging of related files (including documentation)
- Persistence of datasets: there are initiatives, such as DataCite, that are attempting to address long term findability but the persistence of the actual datasets through the long term is still an area of concern, particularly for ‘small’ data
- Sustainability of data curation services in health sciences. The ELIXIR initiative, where European funding is provided for core data services in life sciences, is certainly part of the answer but the problem is bigger than this and growing
- Effecting behaviour change in areas where sharing is not the norm
- Attribution and intellectual property rights when datasets are ‘stacked’ (created from datasets that were in turn created by many others)
- Sharing of systems that include components that were provided by third parties. There are currently many barriers to sharing because third-party components are licensed in ways that prevent this
- Integration of data curation and data exchange facilities in the workflows of research groups, both technically and organizationally
- Problems of the use of data from secondary sources that could be better annotated
- Incentives for data curation and sharing by researchers: there is currently no career-advancement advantage in sharing data or putting effort into curating data to enhance their value to others and biologists do these things altruistically. A more formal reward system, akin to that traditionally offered for publishing research articles, would help here
- Text-mining the literature for material that enables data enrichment: this is a technical challenge, partly, but even more so a process challenge for database curators
- Privacy and data protection issues: this is an issue of major importance in health sciences, but will be particularly so with respect to ‘personal genomes’ (where an individual’s genome is sequenced: what are the implications in terms of privacy, insurance and so forth?)

8 List of figures

9 List of tables

10 Bibliography

Björk B-C, Roos A & Lauri M (2009). "Scientific journal publishing: yearly volume and open access availability" *Information Research*, **14**(1) paper 391.

http://InformationR.net/ir/14-1/paper391.html

Björk B-C, Welling P, Laakso M, Majlender P, Hedlund T, et al. (2010) Open Access to the scientific journal literature: Situation 2009. PLoS ONE 5(6): e11273. doi:10.1371/journal.pone.0011273 http://www.plosone.org/article/info:doi/10.1371/journal.pone.0011273

Sale, AHJ (2006) Comparison of IR content policies in Australia. *First Monday*, **11 (4).** http://eprints.utas.edu.au/264/

Gargouri Y, Hajjem C, Lariviere V, Gingras Y, Brody T, Carr L and Harnad S (2010) Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research. *PLOS ONE*, 5 (10). e13636 http://eprints.ecs.soton.ac.uk/18493/

Hajjem, C., Harnad, S. and Gingras, Y. (2005) Ten-Year Cross-Disciplinary Comparison of the Growth of Open
Access and How it Increases Research Citation Impact. *IEEE Data Engineering Bulletin*, 28 (4). pp. 39–47. http://eprints.ecs.soton.ac.uk/12906/

Matsubayashi M, Kurata K, Sakai Y, Morioka T, Kato S, et al. (2009) Status of open access in the biomedical field in 2005. Journal of the Medical Library Association 97: 4–11. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2605039/pdf/mlab-97-01-4.pdf

Noor MAF, Zimmerman KJ, Teeter KC (2006) Data Sharing: How Much Doesn't Get Submitted to GenBank? PLoS Biol 4(7): e228. doi:10.1371/journal.pbio.0040228 http://www.plosbiology.org/article/info:doi%2F10.1371%2Fjournal.pbio.0040228

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

H Subject-Specific Requirements for Open Access Infrastructure – Attempt at a Synthesis

Christian Meier zu Verl and Wolfram Horstmann

1 Introduction

This study addresses the question how to characterise subject-specific requirements for research infrastructure with a focus on the influences of Open Access (OA), in the general sense covering open access to literature, open data and open science. The introduction – which is assumed to have been read before this synthesis – specified the following.

We refer to the scope of OA in terms of the Berlin Declaration:

Establishing open access as a worthwhile procedure ideally requires the active commitment of each and every individual producer of scientific knowledge and holder of cultural heritage. Open access contributions include original scientific research results, raw data and meta data, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material.

We refer to the definition of Open Access in slightly modified terms of the Budapest Declaration:

By open access, we mean its immediate, free availability on the public internet, permitting any users to read, download, copy, distribute, [export], search or link to the [materials], crawl them for indexing, pass them as data to software or use them for any other lawful purpose.

By research infrastructure we mean the entirety of production and service, which includes instruments like large sensors, satellites, laboratories, and many more facilities, like digital services and virtual research environments. The research process within that refers to all facilitating processes: the researcher and his or her behaviour is not part of the infrastructure.

The chapters in this study present subject-specific views on OA infrastructure for research by analysing research workflows as well as researcher behaviours. They specifically take into account two aspects, namely (i) working with literature and (ii) working with data. Throughout the preceding chapters and throughout this chapter, the topic of OA infrastructure is centred on digital resources. Even though there are many transitions between physical and digital resources mentioned – for example, between the human researcher and the computer or a digital resource, and the physical, experimental as well computational facilities – these transitions will not be addressed explicitly in most of the cases for the sake of lingual simplicity.

The following sections will discuss commonalities of and differences between the different presented views on OA infrastructure and formulate recommendations for supporting the development of infrastructure (e.g. through funding initiatives) under specific consideration of the question how principles of "openness" or OA can be applied. In line with the qualitative approach of this whole study, the synthesis will be provided as an interpretative account.

When comparing the chapters, the most obvious observation can be summarised in one word: diversity. The archaeologist in a desert excavation has different requirements from a climate researcher crunching observational satellite data or an engineer building a biologically inspired robot hand. On the first view, this diversity may appear to be the natural enemy of infrastructure, since infrastructure is about commonalities in terms of global standards, joint facilities and shared resources rather than about differences between diverse subject-specific requirements. Simultaneously, it is obvious that research must be extremely diverse in terms of thematic and methodological specialisation in order to tackle the ever-more specific challenges of the world. **Thus, any roadmap for OA infrastructure must address this natural tension between diversity and infrastructure.** This study chose the approach of addressing this tension directly by providing an account of diversity and then reflect this diversity in specific aspects of OA infrastructure such as OA to literature and OA to data. It is not expected that the study will provide a complete picture and a detailed plan for the next decades: rather it is expected that the reader will gather impressions of diversity and develop a (maybe sometimes tacit) understanding of how diversity can be managed within research infrastructure development in a way that leaves research with sufficient degrees of freedom for self-organised developments while supporting the emergence of synergies between those developments through shared resources that apply principles of openness.

Attempting to provide a synthesis, the following sections will consequently analyse the commonalities and differences. This is done first on a high conceptual level and then on a detailed, systematic case-by-case basis. Thus, the resulting qualitative, rather than quantitative, account shall inform strategic decisions for future developments with respect to conceptual rather than procedural aspects. The specific measures, programmes or plans are assumed to be the result of these strategic decisions.

2 Methodological reflections

The approach taken in this study is unusual – or even extremely particular. Rather than analysing the principles of research practice through large-scale, representative questionnaire exercises, a small selection of partners provide individual and often descriptive accounts of specific subject areas. Rather than mapping the world of research with a broad account aiming at comprehensiveness, the analyses in the specific subject areas dive deep from the institutional and departmental level to the individual researcher and even research project.

2.1 Localising the study in the world of research

The world of research represents the most specialised activities in human behaviour. Being always on the verge of the unknown – things that never have been experienced and discovered before – researchers have to develop extremely resourceful, creative and swift capabilities in order to "squeeze" novel knowledge out of their minds and the world. Additionally, considering how much knowledge has already been generated through research in the last centuries and decades, the questions posed and methods used are becoming ever more capillary. At the same time, the phenomena analysed by research are becoming ever more complex and significant. Topics such as cancer, climate, consciousness or terrorism require many researchers of different subject areas to join forces.

The question of how to characterise research in a comprehensive sense is the subject of specialised research (e.g. philosophy of science or science studies) and goes far beyond the scope of this study. This study rather shall provide an explorative account of a very specific aspect of research practice, namely OA infrastructure. Thus, this study deliberately did not attempt to provide a representative account of research. Instead a pragmatic approach was taken: six partners (institutions, organisations) were chosen to provide their subjective view on OA infrastructure. The selection of partners originally referred to funding areas of the European Commission (EC), which were chosen as pilot areas for implementing the OA policy of the EC. These partners are considered as exemplars of infrastructure institutions in a given subject area (Table H.1): they not only perform research in a given subject area but also provide some sort of infrastructure for their subject area. Thereby, the analysis of both aspects – subject-specific requirements and infrastructure – should be made possible.

Table H.1 Pairings of partners and subject areas

Partner	Subject area (corresponding to EC funding)
CGIAR	Environment (health)
CITEC	ICT – cognitive interaction and robotics
CNR/NKUA	ICT/capacities – e-Infrastructure
DANS	Science and society
EBI	Health
WDCC/DKRZ	Environment

CGIAR, Italian Consultative Group on International Agricultural Research and Bioversity International; CITEC, Cognitive Interaction Technology – Center of Excellence; CNR/NKUA, Consiglio Nazionale delle Ricerche – Istituto di Scienza e Tecnologie dell'Informazione and the Department of Informatics and Telecommunications of the National Kapodistrian University of Athens; DANS, Data Archiving and Networked Services; EBI, European Molecular Biology Laboratory/European Bioinformatics Institute; WDCC/DKRZ, World Data Center for Climate/Deutsches Klima Rechenzentrum.

The six subject areas and corresponding institutions definitely do not represent the complete world of research. However, they are spread across different domains of research, such as (natural) science, the social sciences and the humanities. Many of the partners are themselves highly interdisciplinarily organised, sometimes bridging between sciences and humanities (e.g. CITEC and CGIAR) and often showing overlaps in their constituent disciplines with other partners participating in this study, For example, both CGIAR and WDCC include environmental science, and both CITEC and EBI include computer science. In a sense, the approach taken in this study tries to provide vertical "drilling cores" into the world of research infrastructure rather than represent research infrastructure with a horizontal coverage.

2.2 The process of writing the chapters

Each exemplar partner appointed one or more chapter authors. In addition to all participants receiving written briefings from the editors, all chapter authors physically met three times to discuss concepts and progress. It was a deliberate decision not to provide too strict methodologies and structures for the chapter authors in the writing process. The reason for this liberal methodological approach was to provide a degree of freedom that could elucidate obvious but also subtle differences between the subject areas and inform future studies about possible approaches. The authors are experts in their fields and it was assumed that they know best themselves how to characterise their subject area. The reader should be provided with an expert subjective view of the given subject area with a taste of the sometimes implicit principles of thinking and working in that subject area rather than a normalised account constrained by too many pre-fabricated assumptions. In this sense, the free choice of chapter authors in how to characterise their subject areas is part of the design of this study, since the individual methodology chosen by authors to characterise their subject area also informs the reader about the self-perception within that subject area. Further, comparing the different methodologies applied in the chapters provides in itself an account of diversity between the subject areas.

2.3 Observations during the writing process

Why is the understanding of diversity so important for the future development of research infrastructure? Why did this study not try to focus on uniformity? It became immediately clear in the discussions at the meetings and the written correspondence that chapter authors were highlighting differences rather than commonalities. Everybody pointed very much to the “special character of their case” and that it is “not comparable to other cases”. Since it can be assumed that this attitude would become prevalent in any measure that is aimed at implementing infrastructure on a broader scale, it was decided to address this diversity directly. Thus, the diversity of requirements has to be studied, understood and respected with the greatest possible care.

An obvious observation with respect to diversity is that almost every partner emphasised that it is impossible to provide a single typical research workflow, even within the work scope of a given institution. Thus, a generic model of research workflows applying to all subject areas was not feasible. Accordingly, almost each partner subdivided the corresponding chapter into several sections, describing different typical researchers, research groups, disciplines or a number of different research workflows, which defined their very specific requirements for the infrastructure services. The methodologies used to characterise these typical research workflows varied from descriptive, observational accounts to semi-structured expert interviews and systematic questionnaires. This variation in methodology shows that the authors found different methods appropriate to characterising their subject area and supports, again, the presence of strong differences within and between subject areas.

However, the constituent disciplines in one subject area show overlaps between partners in a non-systematic fashion, with almost no overlaps in the subject-specific infrastructure services described, even in cases where the same discipline is involved in two subject areas. In other words, the infrastructure services for one subject area, say biology, provided by two different institutions, say EMBL-EBI and CGIAR, serve very different functions in the scientific community, even though they may be used by the very same researcher. But it has to mentioned that the descriptions of OA to literature show much more homogeneity with respect to infrastructure services than the descriptions of OA to data.

In sum, these unsystematic overlaps between research practices and infrastructure services show that not even a partial Cartesian "map of research" can be produced by analysing and comparing the different chapters. Rather than a traditional disciplinary division, say infrastructure for biology vs. infrastructure for geology, **specific research problems and their corresponding research projects performed by collaborative interdisciplinary organised groups can be identified as the drivers of research infrastructure**. Thus, a multidimensional organisation of research infrastructure – a network model – appears to be the appropriate model for describing research. both a layer cake model, in which a subject research layer is based on a data layer, in turn based on an ICT layer, or a hierarchical matrix model in which layers are pervaded by subject-specific "columns", seem too simple to catch the subtleties of research infrastructure.

2.4 Initial observations summarised

The analyses provided by the exemplars are "drilling cores" that characterise research infrastructure in a given subject area. Initial observations about these drilling cores can be summarised in a first coarse approximation as follows:

i. Each institution or organisation provides research infrastructure specific to the subject area in terms of multiple and focused requirement satisfactions, defined by the constituent subject-specific research processes.

ii. Institutions or organisations, although considering themselves subject-specific, do not have the self-perception for serving a single subject area. Rather, they serve a multitude of disciplines, with the tendency of becoming even broader in their constituent disciplines.

iii. Infrastructure provided by the institutions or organisations is designed to support the sharing of resources and collaborative research with a multitude of different services, such as databases, repositories, analytic software or communication tools. Those tools seem to be modular to serve the diverse needs of the researchers involved rather than providing an integrated virtual research environment for one subject area.

iv. OA to literature is described as a relevant phenomenon in each different subject area. The degree to which OA to data is established in a given subject area varies.

v. OA to data is characterised as much less established than OA to literature and often accompanied by enumerations of obstacles that prevent OA to data.

3 General assumptions throughout the chapters: the benefits and obstacles of OA infrastructure

Even more prominent than the question of characterising research in the different subject areas, the current state and perspectives of OA infrastructure was addressed by this study. Before providing a more detailed account in the later sections of this summary, a general interpretation of assumptions regarding OA infrastructure is given here first.

3.1 Benefits of OA infrastructure

OA infrastructure is a complex concept determined by multiple aspects, most obviously by the two aspects infrastructure and OA. These two aspects will now be characterised separately in terms of their benefits, as can be concluded from the chapters.

3.1.1 Benefits of infrastructure

Cost considerations As the predominant benefit, cost considerations can be easily identified as a benefit of providing infrastructure. It is generally assumed to be more efficient when a given service, say a database, is provided once to a research community rather than providing the service twice or multiple times in different locations. Today's digital services easily allow remote access to a single shared service from different locations for different users, wherever they are and whatever their particular research interest.

Enabling research Another benefit is providing researchers with access to resources that would otherwise be not accessible, to enable research processes that would otherwise be not possible. Examples are access to licensed literature for which the individual researcher has no access rights and access to research results (e.g. personalised surveys) that are only accessible on special conditions (e.g. highly secured workstations) or expensive experimental facilities.

Transparency and comparability Good research practice, for example in terms of reproducibility of research results, dictates the comparability of research results in order to verify and falsify them. When researchers use the same infrastructure, say again a database, the research processes are more likely to be comparable than when differing infrastructures are used: file formats, metadata standards and statistical methods tend to be similar in an integrated infrastructure. The emergence and the application of standards is thus a very important implication with respect to transparency and comparability.

Synergies Providing infrastructure is a way of sharing resources among researchers. Synergies emerge through sharing when the research process can develop a novel quality that would not be possible without sharing. A prominent example is the Human Genome Project, in which joint infrastructure and standards were used to collaboratively build a resource of research results that could practically only be achieved in a global and collaborative manner.

3.1.2 Benefits of OA

Cost considerations Any barrier to resources for research causes costs. In a simple case, licensing access to electronic literature requires researchers or institutions to work with registration or accreditation obstacles (e.g. logins or IP-checks, digital rights management) and payments (e.g. invoice processing, bank transfers). In a more complex case, missing access to primary research data can force research funders to finance the same experimental projects several times. In general, the innovation capacity and creativity of research is limited wherever research resources are kept behind barriers. Thus, anything that is OA can help to reduce the effort and costs incurred when dealing with barriers.

Enabling research OA can even play a more crucial role when a given research project is simply not possible without OA, i.e. situations in which a researcher is endowed with access to resources specifically because of their open character. This is seen, for example, when a researcher grounds a project on data that have to be open in order to be re-used, say for an application that performs runtime public transport monitoring.

Transparency and comparability It almost goes without saying that OA enhances the possibilities for researchers to use, analyse, assess and check the work of their peers. The recent trend of data publishing as a supplement to research literature corroborates this observation.

Synergies Intensified peer communication and collaboration through OA resources is instrumental to effective division of labour and complementary, rather than redundant, research projects. OA can enhance the information flow between otherwise isolated research activities and is therefore crucial for performing collaborative, interdisciplinary research projects.

Summary The benefits of infrastructure and OA considered separately reveals a strong relation between these two main aspects addressed in this study. Even though it might seem trivial, it should be noted at this point that OA and infrastructure are two completely different phenomena: OA is a mode of communication while infrastructure refers to facilities. However, the benefits of both can be characterised referring to the same aspects of research: cost considerations, enabling research, transparency, comparability and synergies. The most obvious reason is that **both infrastructure and OA imply a notion of sharing**

This study shows that there are general assumptions underlying the analyses of OA infrastructure. In summary, infrastructure is an essential prerequisite of research that:

- reduces costs by providing shared resources instead of building multiple local solutions,
- enables research that is otherwise not possible,
- enhances comparability by providing joint standards and methodological frameworks,
- creates synergies between researchers, groups or disciplines by sharing the same resources.

If infrastructure is operated according to OA principles, all benefits of infrastructures are boosted because the degree to which the sharing of resources can be exploited is maximised.

3.2 Obstacles to OA infrastructure

The obstacles mentioned in the chapters are so manifold that such an analysis would justify a dedicated study on these obstacles. Consider only one example within one specific chapter, namely data collected at a archaeological excavation site: the necessity of barriers to excavation data is explained by the protection of the data against the possible abuse by treasure hunters and the possible abuse by political activists. Treasure hunters or political activism are rather surprising in the context of research resources! It would be interesting to collect all such examples throughout the chapters, but that would not emphasise the obvious observations with respect to the obstacles for OA, namely:

i. There may be good reasons against a completely open research infrastructure, particularly when they are grounded in the research processes themselves, for example needing time to exploit the results before someone else (competition among colleagues or industries), protection of privacy (medical records, surveys) and risk of abuse (dangerous technology).

ii. These good reasons apply to a much lesser degree to the aspect of OA to literature than to the aspect of OA to data, since literature is localised at the end of the research process, where many processing and refinement steps on the results to prepare them for publication have already been performed.

iii. Obstacles to OA infrastructure vary so dramatically across subjects that they cannot be foreseen in a general OA policy. Therefore, a procedure for allowing exceptions from a general OA policy is required. Exceptions can be justified and assessed on a case-by-case basis for each research project, particularly with respect to the question of OA to data.

4 Comparative analysis

The following sections provide a comparative account of the main aspects addressed in the chapters, namely the characteristics of the research lifecycles as a whole and its constituent aspects of literature management and data management.

4.1 Research lifecycles

Each subject area is organised in different ways as a result of differing research lifecycles. Even individual fields in one subject area can be organised differently. Therefore, it is necessary to compare parts of research practices of locally situated units belonging to an entire research field. Depending on the subject of research itself, it is possible to find concordances of data management between research fields (e.g. between climate research and ICT). Thus, the purpose of this comparison is to find commonalities and differences in research workflows and to emphasise the research steps described as at the core of research activity. All considerations below are grounded on rather abstract and minimal descriptions of observable workflows, which were described in each subject chapter. We will discuss each research workflow by pointing at essential steps of research practices. In the end of this section we will highlight common steps and main research activities of each presented case by comparing all research lifecycles.

CGIAR/Bioversity International observed four projects to define requirements on agricultural research by focusing on typical groups within this field. The research field of agriculture is highly interdisciplinary and includes economics, geography, geology and climate research as well as biology. Most observed work groups are collaborating internationally to study agricultural developments in different parts of the world. Therefore, these projects need simple and stable instruments to measure, for example, developments of plants or behaviour of farmers. Generic steps of workflows within the field of agricultural research are (i) data collection, (ii) cleaning, (iii) archiving, (iv) use and (v) dissemination. It is not possible to locate the steps of the workflow to which researchers pay more attention, but the effort in collecting data is huge, which suggests that data collection and use are the steps with most activity.

CITEC categorised four different research areas within the institute. The categorisation runs along the following four sections:(natural) science, social science and the humanities, computer science, and robotics and engineering. Groups within one section behave similarly for data and literature management and also conduct research in similar ways. But there are major differences between these sections on research objectives, methods used and infrastructures that influence the entire way of conducting research. CITEC performs highly interdisciplinary research on ICT and each working group is well engineered. The common steps of research are: for (natural) science (i) data collection, (ii) processing, (iii) enrichment and (iv) re-use; for social science and the humanities (i) data collecting, (ii) processing, (iii) archiving and (iv) enrichment; for computer science (i) data collection or re-use, (ii) processing, (iii) archiving; and for robotics and engineering (i) data collection, (ii) enrichment, (iii) processing, (iv) archiving and (v) re-use. Beyond this interdisciplinary cooperation of groups, CITEC cooperates with international researchers and companies all over the world. It is not possible to locate main research activities but all groups basically have three steps in common: data collection, data processing and data archiving. So these steps are typical and most important for CITEC as an exemplar of ICT research.

CNR/NKUA describe six research workflows within the field of e-Infrastructure. The scope includes public and commercial research institutes and three of the cases have the following workflow: (i) requirement analysis, (ii) design, (iii) development, (iv) documentation and (v) testing and deployment. All six research lifecycles have standard phases in common, such as : (i) requirement analysis, (ii) designing and (iii) implementation. Depending on the research objective itself, e-Infrastructure research is heavily engineered and uses a vast amount of computing hard- and software. International collaborations are common to all research groups.

DANS describes a research workflow with five steps within the field of the humanities and social science, which is based on the workflow of the large-scale activity Digital Research Infrastructure for the Arts and Humanities (DARIAH). There are the following steps: (i) search/discovery, (ii) gather, (iii) analysis/experiment, (iv) publish/disseminate and (v) store/archive. Collaboration and sharing of current research results with the public or internal working groups is possible in all steps. It is not possible to identify steps in this research lifecycle that are more prominent than others.

EMBL-EBI describes five cases within the research field of health and life science. All cases have different objectives, such as examining genomic sequences, mechanics and dynamics of cell divisions, imaging brains, simulating neuronal cell signals and developing databases for mouse embryonic models. The common research lifecycle mentioned by EBI has seven steps: (i) data collection or re-use, (ii) processing, (iii) analysis, (iv) enrichment, (v) archiving, (vi) dissemination and (vii) publication of literature. All five cases use other methods to explore their subjects but it is impossible to make general statements about any kind of emphasis of a specific activity. Only research modelling is primarily related to data processing (for modelling) and archiving. The rest of research workflows are equally distributed in their activities throughout the complete research lifecycle.

WDCC/DKRZ describes five different cases of research institutes within the field of climate research. Most types of data are observational data such as images or sheets of numbers. A common research workflow includes four steps: (i) data collection or re-use, (ii) processing, (iii) enrichment and (iv) archiving and re-use. Climate research is very well engineered and uses a vast amount of computing hard- and software and technical equipments such as satellites, airplanes and observation stations. Researchers share their facilities internationally to constantly use these expensive instruments. Therefore, data sharing with colleagues and/or the public is commonly established at the climate research community. Some institutes are more specialised in data collection and archiving than others, and some researchers spend more time on data collection, archiving and disseminating than researchers who have no access to data or access to data-collecting facilities only for a limited period of time.

Summary Comparing these descriptions of research workflows from different subject areas and cases, it becomes obvious that some workflow steps are generic to conduct research beyond disciplinary and institutional boundaries. Even if we sample different research fields we can observe five steps that emerge in nearly every workflow. These five steps are (i) data collection (as direct or indirect collection by searching through databases), (ii) processing, (iii) enriching, (iv) archiving and (v) re-using. These steps are rather abstract and you can discover more differences by focusing in-depth on a single step. For example, collecting data enforces other research practices and facilities in climate research as in the field of social science.

Another aspect of the research workflow is the order of individual steps. Hence, it is instructive to have a look at the ordinal dimension of research workflows. Most descriptions of workflows start with the collection of data but some start with the re-use of data or requirement analysis, for example.

Also, the combination of steps is different, even in within one institute, and depends mostly on the research subjects, applied methods, technologies and collaborations. WDCC and CITEC have one arrangement of workflow in common. All other observed research workflows are different in combination. Most research workflows are workflows with four, and sometimes three, main steps. Sometimes, even the understanding of what can be count as an autonomous step of research workflow varies from case to case. But a common understanding of steps necessary to conduct research or to build an entire research workflow is observed. This is generic for all analysed data-driven research practices. As we mentioned, subjects, approaches and applied methods diverge at more complex levels; therefore, generic infrastructure has to be highly configurable in combination (such as modules to rebuild individual research rhythms) and suitable for different research environments by adapting the modules.

We conclude that there are five generic steps of workflow, even if these steps are always very specific on closer consideration. The arrangement of steps depends for the most part on the objectives, applied methods, technologies and collaborations. Therefore, any approach to generic infrastructure has to be highly configurable.

4.2 Literature management

A common final good of all research is literature. Almost all significant research knowledge is transformed into literature at a certain point and to a certain extent. The most obvious advantage of literature as container of knowledge is the way it supports understanding and dissemination of insights through time and space. Of course literature is indexical and written in different terminologies (with which one has to be familiar) but it is more durable and reaches more recipients than talk and more generic than data. Hence, management of literature is a generic task beyond disciplinary boundaries to reach large audiences. We differentiate three dimensions of literature management: (i) production, (ii) organising and (iii) dissemination. Nowadays, researchers explore new ways to present their literature. E-publishing, social media, OA and data publishing are only a few aspects depicting the current change of research publication. These upcoming developments influence all of the dimensions mentioned above and even restructure the principles of research. To serve new needs, it is necessary to analyse the management of literature in different research fields. How is literature managed through different disciplines? Which reasons can be observed for differing literature management? Where are the most progressive developments of literature management? How are these new developments organised? In the following section, we will look at each individual chapter one by one to finally compare all of the approaches and discuss commonalities and differences.

CGIAR activities are massively dominated by data management so that literature management is characterised concisely. There are several branches, which show established practices of OA publishing (gold) and CGIAR manages 14 OA repositories spread over all partner institutions. Since 2006 CGIAR provides a virtual library which gives access internal and external research literature on agriculture, hunger, poverty and the environment. This is a shared, integrated service that allows users to tap into leading agricultural information databases, including the online libraries of all 15 CGIAR Centers.

CITEC describes several ways to manage literature. Self-written literature can be presented through the central service PUB which is provided by the Bielefeld University Library. This service manages the bibliographic information as a generic service for all departments of Bielefeld University, which is locally configured to specific needs. Future developments by CITEC are semantic enrichment which allows formal representation of literature and the relations between them. Beyond this generic literature management, CITEC has four different groups which diverge because they are using different tools to write, manage and publish literature. The BehNatNeur group uses Endnote, Mendeley and Reference Manager to manage non-self-written literature. Data and literature can be published together. There are two forms of publishing which are preferred within the group: printed versions and electronic versions (accessible via the Internet) and 34% of published literature is OA (followed the green way). The SocHum group uses BiTeX, Citavi, Zotero and Mendeley. For collaborative writing they use Google Docs and Subversion. Data and literature are usually not published as a compound object and 57% of published literature is OA (followed the green way). The CompSci group uses BibTeX, Mendeley and Drupal (for metadata management of literature and for the literature itself, with modifications). Collaborative writing is managed by Subversion. It is not possible to publish data and literature together. The RobEng group uses Drupal (for metadata management of literature and for the literature itself, with modifications), Endnote, BibTeX and Subversion to manage literature. Both forms are established to publish as a printed version and as an electronic version (via the Internet), and 68% of published literature is OA (followed the green way).

CNR/NKUA stores most of the research literature locally on personal computers and manages and shares via e-mail or with software tools such as Google Docs and Dropbox. Literature writing is often realised with online tools like Google Docs to produce texts cooperatively, but each interviewed group behaves in a slightly different way. The D-Lib group searches to find literature via Google, Google Scholar, Wikipedia and DRIVER. If they write literature collaboratively, they use Google Docs or share their file via Dropbox or BSCW. The Agro-know group uses Google Scholar to find literature and to manage literature via Mendeley. They write collaboratively in many different ways, for example via e-mail, BSCW, Dropbox, Google Docs and Wiki (only if they collaborate with external researchers). Publishing data and literature together is not established. The group prefers to publish their literature OA. The researchers within the National Documentation Center search for literature via Google Scholar and Scopus. Literature is managed with CitUlike. They write documents collaboratively with SVN. They prefer to publish OA. The Greek Research & Technology Network uses Google to search for literature. The researchers manage literature with Mendeley and publish in journals and conference proceedings. A combination of literature and data publishing is not established. They do not prefer to publish OA. The MADGIK group searches for literature with Citeseer, Google Scholar and also with DRIVER. They collaboratively write documents via Google Docs but mainly they exchange documents via e-mail. In general, the common literature lifecycle is: (i) survey, (ii) analysis of literature, (iii) drafting and (iv) publishing. OA publishing is not desired by researchers within one mentioned organisation but by all the others.

DANS facilitates a self-archiving system called EASY (Electronical Archiving SYstem) which can archive both literature and data. There are four interviewed researchers, who come from different disciplines and manage their literature in different ways. (i) By using eDNA it is possible for archaeologists to conduct desk-based research with access to literature and data of other excavations. OA journals are not highly rated within the field of archaeology, so therefore they are not preferred. Some, but not every, researchers have an online list which shows his or her record of publications. Normally publications from excavations are published as reports under institutional copyright. These reports are necessary to understand the datasets in a better way. (ii) The historians described collaboratively written author literature for historical demography data. The specific role sharing depends on the difficulties in handling these demography data. Therefore, some historians prepare the datasets and the others interpret the datasets. (iii) The social scientist remarked that there is nowadays an inflation of publications. Literature writing focuses on articles, which are the major form of publication. There is no high-quality OA journal within the field of social science and publishing OA is not preferred. (iv) For linguists, literature is not only literature for reading but also data for research; therefore there is a clear tendency to OA with literature which can be used in both directions. The world of linguistic publications shifts towards enhanced forms, which make it possible to publish literature with data together. Some publishers embargo the literature for a period of time before it can be distributed OA (green).

EBI mentioned that journal articles are the primary output of life science. Most journals are published by commercial publishers, medical charities, learned societies, medical institutions, universities and research institutes. There are extensive, subject-based repositories of OA literature which are a well-established and integral part of the life science community. OA publication ratio varies between disciplines from 5% to 16%.

WDCC mentioned that online access is established in most subject-specific journals. One interviewed climate researcher reports that publishers support the publications of literature and data together. But only some formats of data are published, for example it is not possible to publish video data within one document. One climate points out that the German national license (covered by the DFG) provides access to the most relevant publication repositories for climate research. The interviewed researcher of the Climate Service Center mentioned that some publishers enable the exchange of data within literature. OA is established within the Climate Service Center. The climate researcher at the Karlsruhe Institute of Technology mentioned that they use Zotero to manage their collections, citations and sharing of literature. They prefer printed forms of literature. The interviewed researcher at the Alfred Wegener Institute for Polar and Marine Research mentioned that they prefer printed forms of literature and that OA is established in some extent.

Summary The landscape of research literature includes six fields of research which are currently similarly organised. If you compare these literature management descriptions, it is obvious that there are various tools to manage, write, publish and find literature. Some tools are common and you can find them all over research fields. These tools serve generic needs going beyond each disciplinary requirement. This is particularly the case for all literature management tools. Even though there are many tools like BibTeX, Citavi, CitUlike, Drupal (with modifications), Endnote, Mendeley and Zotero, they serve the same needs with slightly different modifications.

The infrastructure for literature is well established. The choice may depend on personal or institutional reasons. There are some tools which serve the form of writing via the Internet. Google Docs, BSCW, Dropbox, SVN and e-mail are the mentioned tools to write and share within the writing process documents. The common way is to write one document with many different versions which have to be merged by someone. Google Docs and Wikis serve the function to edit one document through different authors without document exchange. This can be done at the same time and the document will be stored at the server.

The use of OA publications is different between subject areas. Life science and climate research have well-established OA repositories. In the field of social science and the humanities, there are no high rated (golden) OA publication options. Most of the fields prefer to publish articles. Only two fields mentioned publishing books or using websites.

In sum, there is a common ground of literature management on which a generic infrastructure can build to manage the metadata and the literature itself. With respect to the publishing of data together with literature, it is obvious that there is no generic way to do this yet, but there are emerging techniques such as standardised forms of "enhanced publications".

We can conclude that there exists a generic and specific infrastructure which serves the needs of researcher at different research fields. On the one hand, management, discovery and writing literature is organised with the same tools and is not heavily dependent on subject-specific requirements. On the other hand, publication locations are organised in a subject-specific manner through different publishers, journals and OA repositories. OA is well established in life science and climate research and partly established in ICT and agricultural research, but to a lesser extent in social science and humanities.

4.3 Data management

In describing the data lifecycle and how data management is organised among research fields, we first describe the data lifecycle or parts of the cycle worth considering for comparisons. Then we compare these different subject-specific ways of managing data. It should be noted that research data management is but one part of the research lifecycle workflow and does not cover the complete lifecycle.

CGIAR research is data intensive, just as agricultural research is generally characterised. Therefore, the CGIAR chapter focuses on the openness of data sources and not on data management practices in general. Common steps are (i) data collection, (ii) cleaning, (iii) storage, (iv) use and (v) release. Data collection includes, for example, researchers installing portable laboratories in undeveloped landscapes to study agricultural processes. Therefore, these researchers need robust, simple and user-friendly instruments. Data down- and upload can be organised via a cell-phone Internet connection. For cleaning steps, software and manpower were used to describe the structuring process for collected data in order to decide which parts of the data have to be archived and which parts can be deleted. Afterwards, the cleaned data will be stored at a server. Storage also goes beyond the backup data in that these data will be reviewed according to formal standards for data archiving. After reviewing, the data are used and analysed for reports or publications. The data itself will be prepared for release after the publication of the research outputs. Describing metadata follows the standards in the field of agricultural research. CGIAR has several technical solutions for data management which depend on the research objectives. Dataverse is one example for a technical environment of data management and is used for water and agricultural research. Dataverse is a data repository run by Harvard University which provides metadata storage, file format conversion, collection management and customisation of display.

CITEC research is very data intensive. Between the three common steps of data management within the CITEC (data collection; processing, enrichment and analysis; and archiving), there can be additional steps, and especially the last step is not generalised. Archiving is performed by different groups in different ways. For example, there is no common server that archives everything. After archiving, the question of data exchange is important to all researchers. Currently, most researchers exchange their data by personal request and only a few data (e.g. Open Source software) are freely accessible without asking for permission. Open Source software is archived and distributed on a dedicated Open Source server and repository that manages software developments and data. This shall serve as an example for establishing data management on a broader scale. Currently, each group is managing their research data on their own based on a common internal infrastructure with local policies on group level.

CNR/NKUA describes several ways to manage data, from local storage up to Cloud or Grid storage. Different SVN and CMS solutions are used in e-Infrastructure research to manage and disseminate data. Research data are often stored locally; only software as a special kind of data are often stored and found at software sources on the web. These sources are well known within the e-Infrastructure community. Within e-Infrastructure, many kinds of data are produced, processed and archived, but there are no common standards for metadata to simplify data exchange: within one project or organisation, data exchange is well established but there are obstacles to exchanging data with the entire community or the public. This is true for nearly all kinds of data – software, again, being an exception. Reports, technical descriptions and system logs are shared with more access restrictions.

DANS is developing different data management solutions for different research fields. But there is one generic national data management system for the entire field of social science and the humanities which serves demands such as archiving data, curation and publication of data by DANS' staff. Data management is based on a research lifecycle model and supports archiving and exchange of data. In general, data management is integrating the following research steps: (i) discovery, (ii) collection, (iii) annotation and enrichment and (iv) publishing. First of all, data corpora have to be discoverable. Second, data are collected and generated with different kinds of tools. Most data are digital but sometimes digitising artefacts of archaeological excavations is time consuming. Third, annotation and enrichment of data is mostly necessary for all researchers to understand and interpret data correctly. Fourth, publishing data accompanied by literature is not well established within social science and the humanities. There are problems such as no standards for referencing data and less rewards for publishing data than literature.

EBI describes data management as different challenges for different subparts within the field of health and life science. All subjects within the field have databases that store and disseminate data to researchers and the public in general. Data publication is well established in the life sciences as long as the collected and published data do not touch personal rights. All other kinds of data are mostly archived in databases that are accessible for the scientific community. In many cases, there are standards, formats and ontologies that support data exchange.

WDCC describes data management as a major objective in climate research. Hence, there are international projects to organise data storage and dissemination to climate researchers and a broader public. Standards for data exchange and archiving are established within the field of climate research: most research facilities are expensive and therefore data are shared by big collaborative working groups distributed all over the world. The Coupled Model Intercomparison Project 3 and 5 are two large projects which are part of the current infrastructure of storage and exchange of data. Collection, quality control, annotation with metadata and publication of data is well established within the field of climate research.

All institutions and research fields analysed are managing large amounts of diverse data. There are many differences: some fields are more data driven than other fields. Looking at the technical basis of data, such as data types, formats, standards and metadata, it is obvious that data management is organised in many different ways but it can be observed that many fields use similar types of data. Building on similar data types, it could be possible to construct generic research infrastructure, for example managing image data across social science, the humanities, health, life science, and climate research. It also becomes obvious that data management becomes more restricted whenever privacy issues are involved.

Summary The comparative analysis provided descriptions of the main aspects addressed in the chapters, namely the characteristics of the research lifecycles as a whole and its constituent aspects of literature management and data management. Research lifecycles show common steps: (i) data collection, (ii) processing, (iii) enriching, (iv) archiving, and (v) re-using. However, the variance in the descriptions appears stronger than these rather abstract commonalities. Literature management shows strong commonalities in tooling but strong differences in publishing practices: data management shows a large variance in both tooling and data management practice. The step models for the different aspects provided in the chapters indicate the presence of systematic infrastructural services, but the variance of the step models indicates that each infrastructural service is built around a very specific research question or project. Corroborating the general observations in the beginning of this chapter, the comparative analysis shows that OA to literature is a growing or established practice in the subject areas studies but is not yet fully developed. OA to data is considered an important future activity.

5 Conclusions

The general observations on the writing process of all subject-specific chapters and the overall impressions as well as the comparative analysis point to one key challenge: developing research infrastructure that operates in an open mode and thereby supports the diversity of research practices. In a way, infrastructure is an opponent to diversity since infrastructure is not only an essential prerequisite but also a collection of rigid conditions or constraints: it is an inherent property and explicit objective of infrastructure to make research uniform. Openness, however, is a way to maximise the permeability of research resources (literature and data) within research infrastructure so that the collaborative, interdisciplinary and international research activities needed to tackle the next given challenge can emerge.

Measures to support infrastructure developments (e.g. funding programmes) should therefore take into account the following observations, interpreted on the basis of the subject-specific requirement descriptions throughout this volume.

i. Digital literature and data resources are an essential precondition of research. The provision of digital literature and data resources through infrastructural services are perceived as a matter of course (or implicitness) and are not questioned unless they are obviously missing. Thus, knowledge infrastructure, as the entirety of resources and processes related to digital literature and data resources used in research, is not conceived as an explicit facility but rather as an invisible capacity.

ii. OA is described as a *modus operandi* for working with digital literature and data resources, rather than as an end in itself or an ethical principle.

iii. OA to literature and OA to data refer to very different parts of the research process. While literature shows universally generic characteristics, data are much more related to subject-specific methodologies and facilities. Even though the benefits are the same for literature and data – namely cost considerations, enabling research otherwise not possible, transparency, comparability and synergy – the obstacles vary broadly and require that OA to literature and OA to data are treated separately in policy and infrastructure development.

iv. Due to the universally generic role of text-based resources in research, OA to literature can be regarded as a general prerequisite for efficient and effective as well as innovative research and should be mandated uniformly over all subject areas – even if the specific implementation of OA to literature is left at the discretion of the subject areas (e.g. through subject-specific repositories) – and should be arranged in the grant conditions. For non-subsidised research results, organisations should strive for access as open as possible.

v. OA to data has (yet) to be reflected in a fully subject-specific way in policy and research infrastructure development. The emerging practice of mandatory project-specific data management plans that address the question of OA to data could be sharpened by asking the question: "Are data open and if not, why not?" Also, OA in data management plans could be supported by providing a generic Open Data policy with subject-specific *addendi* to such a generic policy. A given subject-specific addendum to a generic Open Data policy may well be mandatory in a given subject area.

vi. The difference between OA to literature and OA to data may be transient as more and more systematic connections between literature and data are made. In many cases, the literature is the data: text-mining and text-annotation enrichment treat text as data and therefore contribute to provide a continuum of semantically connected knowledge resources on the long run. Explorations towards infrastructural linkage between literature and data (e.g. enhanced publications) should be intensified.

vii. The provision of research infrastructure services by institutions and organisations is requirement driven and depends on the research context – even within a smaller subject area – but supports collaboration among researchers from various disciplines. The development scheme in practice tends to be incremental and evolutionary and based on prototypes and working solutions rather than applying theoretical frameworks and capacious facilities.

viii. The layer cake model of research infrastructure does not reflect the complex organisation of research infrastructure. The distinction between horizontal developments, based on generic research processes and ICT standards, and vertical developments, based on subject-specific research questions, is helpful since it breaks up the layer cake model and suggests a hierarchical matrix model. However, a network model of research infrastructure, consisting of a multitude of subject-specific nodes that apply common local design principles (e.g. metadata standards, exchange protocols) in order to communicate with one another and share resources amongst other nodes, reflects best this study's descriptions of research infrastructure and is assumed to be the most promising approach for designing future research infrastructure developments.

Future research infrastructure developments should consider the following principles in order to reflect the diversity of research as the key challenge:

i. Support subject-specific developments that are research driven, incremental and evolutionary in order to match and adapt to the established situated practices.

ii. In a separate strand, support the development of generic infrastructural services and standards applicable in local subject-specific nodes. Services and standards should obviously be maintained by institutions and organisations with long-term commitment.

iii. Provide systematic cross-talk between the subject-specific and generic developments by:

   a. providing research and development programmes that explicitly address the question of how to link subject-specific and generic developments. Examples for activities are science and technology studies, networking events and focused infrastructure projects,

   b. installing advisory boards or oversight groups for projects and funding programmes that have representations of both subject specialists and infrastructure specialists, and

   c. enforcing the mutual participation of subject specialists and infrastructure specialists in assessments and reviews.

iv. Apply OA as a *modus operandi* in all activities. It should be mandatory for literature and is recommended for data. Appropriate exceptions for specific subjects can be considered.