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-to-date 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 require-

ments. 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.