Project no. 269977

APARSEN
Alliance for Permanent Access to the Records of Science Network

Instrument: Network of Excellence

Thematic Priority: ICT 6-4.1 – Digital Libraries and Digital Preservation

D33.1A Report on Peer Review of Research Data in Scholarly Communication (Part A of D33.1)

Document identifier:	**APARSEN-REP-D33_1A-01-1_1**
Due Date:	2012-02-29
Submission Date:	2012-04-30
Work package:	WP33
Partners:	AFPUM, AIRBUS, C.I.N.I., DNB, DPC, IKI-RAS, KB, PCL, SBA, STFC, STM
WP Lead Partner:	AFPUM
Document status	FINAL
URN	urn:nbn:de:101-20140516241

Abstract:

Quality assurance of scientific information is a precondition and integral part of digital long-term archiving. To operate successful digital long-term archiving, organizations from the fields of science, culture and business cooperate within the EU project, APARSEN. The objective of this project is to set up a "long-lived Virtual Centre of Digital Preservation Excellence". Securing permanent access of quality assured research data on reliable repositories is a central concern of APARSEN. This report documents ideas, developments and discussion concerning the quality assurance of research data. Focus is placed on action taken by science, e-infrastructure and publishers on quality assurance of research data. Such action is documented and classified in this report. Future fields of research are then identified based on this work.

Delivery Type	REPORT
Author(s)	AFPUM, AIRBUS, C.I.N.I., DNB, DPC, IKI-RAS, KB, PCL, SBA, STFC, STM
Approval	David Giaretta/Simon Lambert
Summary	Report on peer review of research data
Keyword List	trust, peer review
Availability	☐ PUBLIC

Document Status Sheet

Issue	Date	Comment	Author
0.1	2012-01-13	First Version	Heinz Pampel, Hans Pfeiffenberger, Angela Schäfer, Eefke Smit
0.2	2012-02-08	Completion in Chapter 3	Stefan Pröll
0.3	2012-02-14	Revised and consolidated version	Christoph Bruch Heinz Pampel, Hans Pfeiffenberger
1.0	2012-04-28	Format corrections	David Giaretta
1.1	2012-08-16	Further minor format changes after acceptance by EC	Simon Lambert

Project information

Project acronym:	**APARSEN**
Project full title:	**Alliance for Permanent Access to the Records of Science Network**
Proposal/Contract no.:	**269977**

Project Co-ordinator: Simon Lambert/David Giaretta

Address:	STFC, Rutherford Appleton Laboratory Chilton, Didcot, Oxon OX11 0QX, UK
Phone:	+44 1235 446235
Fax:	+44 1235 446362
Mobile:	+44 (0) 7770326304
E-mail:	simon.lambert@stfc.ac.uk / david.giaretta@stfc.ac.uk

CONTENTS

1 INTRODUCTION

Scientific progress is based on high quality information. The term, quality, is defined in the Academic Press Dictionary of Science and Technology as follows: "[…] an essential or distinctive characteristic of property of a thing […]"[1]. The metaphor "standing on the shoulders of giants", which vividly describes the scientific cognitive process clearly shows that new findings are always based on statements already published.[2] Access to information of which the quality is assured is therefore a precondition for scientific excellence.

Growth in the digitization of science is opening up a wide range of opportunities for scientists. The exchange of scientific results independent of time and location, collaboration in virtual research environments or the inclusion of laymen in the scientific process of cognition within the scope of so-called "citizen science" are just some examples of the potential of digital science. New perspectives have also emerged for quality assurance of scientific information: comment and assessment functions as well as new processes for checking plagiarism are examples of the new opportunities which are being increasingly incorporated in daily scientific work.

In addition to the various opportunities provided, there is also a wide range of challenges. As a result of digitization, STM[3] disciplines in particular are faced with the task of organizing and permanently maintaining a fast growing volume of digital research data. To enable excellent science it is essential to ensure lasting access to these digital information items. The Alliance for Permanent Access (APA) [4] and its members are addressing this issue. The mission of the APA is "to develop a shared vision and framework for a sustainable organizational infrastructure for permanent access to scientific information". To operate digital long-term archiving successfully, institutions from the science, culture and business sectors cooperate in the Alliance. In addition, under the umbrella of the APA "a long-lived Virtual Centre of Digital Preservation Excellence"[5] is being set up within the EU project APARSEN - Alliance Permanent Access to the Records of Science in Europe Network, which addresses the challenges of digital long-term archiving.

Quality assurance of scientific information is an essential precondition and an integral component of digital long-term archiving. APARSEN addresses the following quality assurance issues:

- Quality assurance of scientific e-infrastructures such as e.g. repositories.
- Quality assurance of digital items stored on e-infrastructures such as e.g. research data.

These two topics are analyzed together within APARSEN, in the work package "Peer review and 3rd party certification of repositories". The results are shown in two independent, parallel reports. This report focuses on quality assurance of digital items. Quality assurance of repositories is handled in a separate report (D33.1B).

This report documents ideas, attitudes, developments and discussion concerning quality assurance of research data. The focus is on action taken by scientists, e-infrastructure providers and scientific journals. Their measures are documented and categorized. Future fields of research are to be described based on this work.

1 Morris, C. (Ed.). (1991). Academic Press Dictionary of Science and Technology. London: Academic Press.

2 Refer to Wikipedia article "Standing on the shoulders of giants" Retrieved from http://en.wikipedia.org/wiki/Standing_on_the_shoulders_of_giants

3 Science, Technology and Medicine

4 http://www.alliancepermanentaccess.org

5 http://www.aparsen.eu

2 THE DATA CHALLENGE

The advance in digitization of science enables new processes for handling scientific data. In 2003, leading science organizations described the potential of the internet for the scientific process of cognition in the "Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities". The demand for open access and unrestricted re-use of scientific information is formulated in the declaration as follows:

"Open access contributions include original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material."[6]

Science organizations worldwide are addressing the potential of openly accessible research data. The vision of the competitive European Research Area (ERA)[7] of EUROHORCs and the European Science Foundation (ESF) stated in 2009:

"The collection of research data is a huge investment. Permanent access to such data, if quality controlled and in interoperable formats, will allow better use to be made of this investment because it allows other researchers to (re)use them. Furthermore it allows re-analysis and could play a role in ensuring research integrity."[8]

Improved accessibility to research data is also demanded on a political level. In 2007 the Organization for Economic Co-operation and Development (OECD) passed the "Principles and Guidelines for Access to Research Data from Public Funding". This paper demands an increase in societal benefit by means of openly accessible research data:

"[…] access to research data increases the returns from public investment in this area; reinforces open scientific inquiry; encourages diversity of studies and opinion; promotes new areas of work and enables the exploration of topics not envisioned by the initial investigators."[9]

The OECD also emphasizes the importance of quality standards of research data:

"Data managers, and data collection organisations, should pay particular attention to ensuring compliance with explicit quality standards. Where such standards do not yet exist, institutions and research associations should engage with their research community on their development. Although all areas of research can benefit from improved data quality, some require much more stringent standards than others. For this reason alone, universal data quality standards are not practical."[10]

This demand for open access to research data has already been taken up on a national level in some countries, e.g. in Germany. In 2010, the Alliance of German Science Organizations published "Principles for the Handling of Research Data". Extract:

"In accordance with important international organisations involved in funding and performing research, the Alliance supports the long-term preservation of, and the principle of open access to, data from publicly funded research."[11]

Infrastructure facilities such as libraries also recognize the necessity of pursuing new paths in handling research data and are addressing this issue. In its strategic plan, the Association of European Research Libraries (LIBER), Partner in the APARSEN network, states:

"Identification of the role and responsibilities for European libraries in terms of collecting, describing, curating and preserving digital materials, especially but not limited to primary data."[12]

---

6 Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. (2003). Retrieved from http://oa.mpg.de/files/2010/04/berlin_declaration.pdf

7 http://ec.europa.eu/research/era

8 EUROHORCs & ESF. (2009). EUROHORCs and ESF Vision on a Globally Competitive ERA and their Road Map for actions. Retrieved from http://www.era.gv.at/attach/EUROHORCs-ESF_Vision_and_RoadMap.pdf

9 OECD. (2007). OECD Principles and Guidelines for Access to Research Data from Public Funding. Paris: OECD Publications. Retrieved from http://www.oecd.org/dataoecd/9/61/38500813.pdf

10 Ibid

11 Alliance of German Science Organisations. (2010). Principles for the Handling of Research Data. Retrieved from http://www.allianzinitiative.de/en/core_activities/research_data/principles/

Scientific publishers are also addressing the challenge of contemporary handling of research data. In the "Brussels Declaration" of 2007, a policy document of the International Association of STM Publishers, also an APARSEN member, it is stated:

"Raw research data should be made freely available to all researchers. Publishers encourage the public posting of the raw data outputs of research. Sets or sub-sets of data that are submitted with a paper to a journal should wherever possible be made freely accessible to other scholars. "[13]

Research funders are increasingly demanding open access to data emerging from sponsored projects, in so-called data policies. Some examples:

- In 2003 the National Institute of Health (NIH) published a "NIH Data Sharing Policy".[14]
- In 2007 the Wellcome Trust issued a "Policy on Data Management and Sharing".[15]
- In 2011 the US National Science Foundation (NSF) issued a "Data Sharing Policy".[16]

The discussion of challenges of permanent access to research data is also conducted by leading scientific journals. This is reflected in the way the issue is dealt with in Nature and Science. Both journals regularly address the topic:

- 2008: Nature Special on Big Data[17]
- 2009: Nature Special on Data Sharing[18]
- 2011: Science Special on Dealing with Data[19]
- 2011: Science Special on Data Replication and Reproducibility[20]

Processes and methods of data sharing are distributed differently within the scientific disciplines. The practice of data exchange is especially distinctive in genetic research. A significant step was taken towards openly accessible research data in this field in 1996 with the passing of the "Bermuda Principles", within the scope of the Human Genome Project. In the "Bermuda Principles" it is stated that:

"All human genomic sequence data generated by centres funded for large-scale human sequencing should be freely available and in the public domain to encourage research and development and to maximize the benefit to society."[21]

In the "Bermuda Principles" a scientific community coordinated with sponsor organizations to create self-obligating rules for handling research data. This method is also supported by scientific journals in the field of biomedical science. In their editorial policies these journals call upon their authors to have the data on which a publication is based made accessible on a repository. For example, in the editorial policy of Nature Cell Biology it is stated:

---

12 Ligue des Bibliothèques Européennes de Recherche. (2009). Making the case for European research libraries. LIBER Strategic Plan 2009-2012. Retrieved from http://www.libereurope.eu/sites/default/files/d5/LIBER-Strategy-FINAL.pdf

13 International Association of STM Publishers. (2007). Brussels Declaration. Electronic Publishing. Retrieved from http://www.stm-assoc.org/brussels-declaration/

14 National Institutes of Health. (2003). Final NIH Statement on Sharing Research Data was. Retrieved from http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

15 Wellcome Trust. (2010). Policy on data management and sharing. Retrieved from http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm

16 National Science Foundation. (2011). Proposal and Award Policies and Procedures Guide. Chapter VI - Other Post Award Requirements and Considerations. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4

17 Nature Special on "Big Data". 2008. Online: http://www.nature.com/news/specials/bigdata/

18 Nature Special on "Data Sharing". 2009. Online: http://www.nature.com/news/specials/datasharing/

19 Nature Special on "Dealing with Data". 2011. Online: http://www.sciencemag.org/site/special/data/

20 Science Special on "Data Replication and Reproducibility". 2011. Online: http://www.sciencemag.org/site/special/data-rep/

21 Smith, D., & Carrano, A. (1996). International Large-Scale Sequencing Meeting. Human Genome News, 6(7). Retrieved from http://www.ornl.gov/sci/techresources/Human_Genome/publicat/hgn/v7n6/19intern.shtml

"An inherent principle of publication is that others should be able to replicate and build upon the authors' published claims. Therefore, a condition of publication in Nature Cell Biology is that authors are required to make materials, data and associated protocols available to readers on request." [22]

In addition, processes of data publication are described for individual data types. Example:

"Structures: Papers must state that atomic coordinates have been deposited in the Protein Data Bank (or Nucleic Acids Database, as appropriate), and must list the accession code(s). Accessibility must be designated 'for immediate release upon publication'."[23]

As a result of this method, the reviewer of an article submitted has the opportunity of considering the source data when assessing the work.[24]

[22] Nature Cell Biology. (n.d.). Editorial Policies. Retrieved from http://www.nature.com/ncb/about/ed_policies/index.html

[23] Ibid

[24] Processes of data publication and the interplay between science, libraries, data Centres and publishers was investigated by some APARSEN partners in the course of the EU project Opportunities for Data Exchange (ODE): http://ode-project.eu

3 DATA AND PUBLICATIONS

3.1 LINKING AND CITING

As part of project Opportunities for Data Exchange (ODE)[25] several APA partners investigated the ways in which data and publications are currently being integrated. To that purpose, the "Data Publications Pyramid" was developed (see illustration 1) to distinguish five different manifestation forms for data to appear inside or alongside publications.

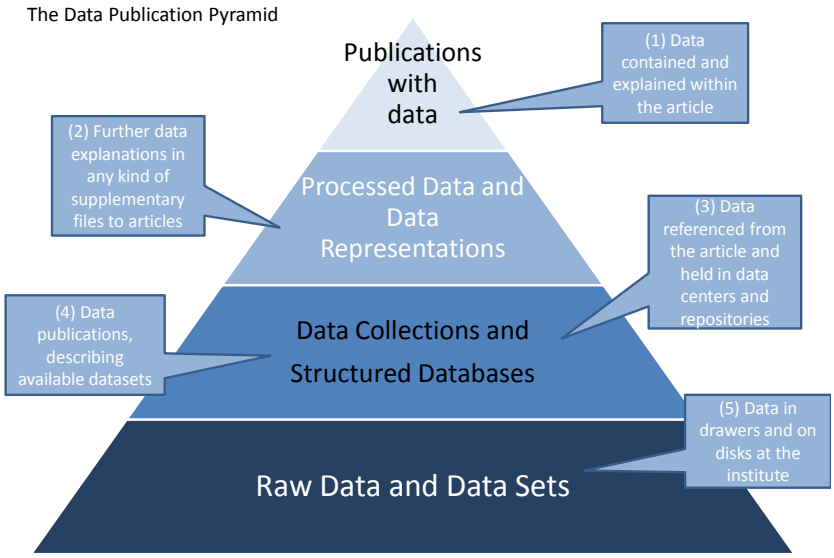

Figure 1 The "Data Publications Pyramid"

As drivers to promote further integration of data and publications, so that data as a first class research object ensures its preservation in the Record of Science, the following important opportunities were listed, together with the first examples setting such a course:[26]

- Require availability of underlying research material as an editorial policy (example: Nature, PLoS)
- More careful treatment of digital research data submitted to journals and ensure it is stored, curated and preserved in trustworthy places (several examples of collaboration with community endorsed repositories)
- Ensure (bi-directional) links and persistent identifiers (examples for listed public archives, DataCite, Dryad)
- Establish uniform citation practices (examples Elsevier-PANGAEA, ESSD, DataCite, Dryad, Thieme)
- Establish common practice for peer review of data (example ESSD)
- Develop data-publications and quality standards (example ESSD, GigaScience, IJRobotics Research)

In Chapter 5.3 of this report several of these drivers have been further investigated, by inviting the expert opinion from publishers and journal editors on the present status and their future ideas.

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

[26] Reilly, S., Schallier, W., Schrimpf, S., Smit, E., & Wilkinson, M. (2011). Report on Integration of Data and Publications. Retrieved from http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-ReportOnIntegrationOfDataAndPublications-1_1.pdf

3.2 INTERWEAVING DATA AND PUBLICATIONS

A fundamental requirement for reviewing scientific experiments and theories is the possibility of reproducing the claims and conclusions made by scientists in their publications. The validity of an experiment can only then be judged correctly, if it is possible to rerun a specific experimental setup under similar preconditions. This essential standard applies more than ever, when huge amounts of digital data are involved.

Current journal articles are mostly detached from the digital data they are based on. They hardly allow peer scientists to replicate the findings from data intensive experiments. Although research data are often added as a supplement to the static original article, it is too hard in many cases to assess the validity of the results published.

Several approaches try to integrate research data more closely into publications. Enhanced papers, rich Internet publications or executable papers are some of these developments that combine research data and articles and allow researchers in varying degrees to re-use, analyze and verify the data and the publication.[27] [28] [29] Enhanced papers refer to publications that are augmented with links to additional content. These links can point to technical documentation, comments, images and other sources available online and also to research data. Rich Internet applications feature multimedia content and interactive elements that support the visualization of research results, such as interactive maps or tools for data analysis. The last approach – executable papers – refers to publications that allow executing and therefore rerunning scientific workflows. All three concepts mentioned have in common that they are designed to advance the usability of research data in combination with scientific publications.

An example for current efforts into the direction of reproducibility, verifiability and re-usability of research data shall be given on the example of Elsevier. The publisher called for the Executable Paper Grand Challenge[30] in 2011 and investigated on the topic of combining traditional journal publications with live research data. The goal of this initiative was to promote the usage of research data directly in publications and to go beyond the simple supplement of data to traditional publications. Research data should be integrated directly into the interactive publications and allow consumers to use these to replicate the results. This should be achieved by working with actual data, algorithms and code of the research project and alter and change parameters. This would allow the re-running of experiments by using the exact same data and verify the results. It should also be possible to edit the data and methods in a convenient fashion. The effects of such changes in the parameters can be detected directly. By supporting this process with tools, the quality of peer reviews should be enhanced and the effort reduced simultaneously.

The challenge was supposed to result in a platform independent solution capable of executing files used in scientific environments and solve the problem of dealing with large files, which are common for research data sets. A crucial requirement is the capturing of provenance information that allows tracing all interactions with the system. The winner of this competition was the Collage Authoring Environment[31]. This eScience framework combines static textual information with interactive media. It provides a server infrastructure, which allows authors to collaboratively assemble executable papers and readers and reviewers to view these publications and use the embedded multimedia features in an interactive way. In the Collage terminology the executable code is called asset. Three different types exist so far: input forms for feeding data into the experiment, visualizations to render the output and code snippets that allow editing of the source code that is used for an experiment. The assets add the

[27] Woutersen-Windhouwer, S., Brandsma, R., Verhaar, P., Hogenaar, A., Hoogerwerf, M., Doorenbosch, P., Dürr, E., et al. (2009). Enhanced Publications. Linking Publications and Research Data in Digital Repositories. (M. Vernooy-Gerritsen, Ed.). Amsterdam: Amsterdam University Press. Retrieved from http://dare.uva.nl/aup/nl/record/316849

[28] Breure, L., Voorbij, H., & Hoogerwerf, M. (2011). Rich Internet Publications: "Show What You Tell." Journal of Digital Information, 12(1). Retrieved from http://journals.tdl.org/jodi/article/view/1606/1738

[29] Nowakowski, P., Ciepiela, E., Harężlak, D., Kocot, J., Kasztelnik, M., Bartyński, T., Meizner, J., et al. (2011). The Collage Authoring Environment. Procedia Computer Science, 4, 608-617. doi:10.1016/j.procs.2011.04.064

[30] http://www.executablepapers.com/

[31] http://www.sciencedirect.com/science/article/pii/S1877050911001220

necessary dynamics to the otherwise static publication and allow validation, reproduction and also the reuse of the underlying data by readers in general and by reviewers in particular.

Authors have to write the publication within the environment and provide the required data and the experimental setup. They can define interactive elements that allow readers to rerun and validate the results. The computations are carried out on the site of the publisher, which provides the required infrastructure for the executions. The framework is designed to run on different platforms, which provide the specific environments for the experiments. It follows a modular approach and allows communication across different systems. The user only requires a Web browser and does not need to install additional software. Interactive elements are rendered directly into the executable paper and have a similar layout and appearance as a classical publication. The enhancement with interactive features allows readers to verify the data in a straightforward and convenient fashion, which should also reduce the effort of reviewers.

Systems like the Collage Authoring Environment are first implementations of a new type of scientific publications. They provide reviewers with research results augmented with interactive media. Other approaches like Paper Mâché[32] or SHARE[33] make use of virtual machines that provide an environment for publishing executable papers. Such a virtual machine would include all required tools and the complete software setup, which is needed to reproduce and verify an experiment. The virtual machine may also contain data, the required scripts and embedded code snippets to generate updated revisions of a paper and allow reviewers to retrace the steps and verify results of the authors.

Promising approaches to interweave data and publications exist but many of them are still in an experimental stage. In particular, replicating experiments that require highly specialized hardware or high performance computing environments is still a challenge. Further on, executables pose the question whether these need to be preserved, for which time, by whom and - most challenging - how. Nevertheless the projects introduced are interesting approaches towards a new publishing paradigm.

[32] Brammer, G. R., Crosby, R. W., Matthews, S. J., & Williams, T. L. (2011). Paper Mâché: Creating Dynamic Reproducible Science. Procedia Computer Science, 4, 658-667. doi:10.1016/j.procs.2011.04.069

[33] Van Gorp, P., & Mazanek, S. (2011). SHARE: a web portal for creating and sharing executable research papers. Procedia Computer Science, 4, 589-597. doi:10.1016/j.procs.2011.04.062

4 PEER REVIEW OF RESEARCH DATA - CHARACTERISTICS AND SPECIFICS

In STM disciplines the quality of a scientific result is conventionally secured by way of a peer review process. On submitting an article for scientific publication, the article is checked by members of the respective discipline in accordance with predefined criteria. These criteria are defined by the editors of the respective scientific publications.

The peer review process was created during the 17th century. In one of the first scientific journals, the Philosophical Transactions founded in 1665, an article had to be reviewed by a member of the Council of the Royal Society before publication.[34] In 1752 the journal established a "Committee on Papers" for quality assurance:

"The new regulation stipulated that five members of the committee would constitute a quorum. It also provided that the committee could call on 'any other members of the Society who are knowing and well skilled in that particular branch of Science that shall happen to be the subject matter of any paper which shall be then to come under their deliberations.'"[35]

Various peer review processes have been developed since then. The three most central processes are the following, which can be characterized by the level of anonymization of participants:

- Single blind: The identities of the reviewers are not known to the authors. The identities of the authors are known to the reviewers.
- Double blind: The identities of the reviewers are not known to the authors. The identities of the authors are not known to the reviewers.
- Open peer review: Collective term for several processes in which the anonymity of participants is partially or wholly excluded.

The peer review processes have different functions depending on the respective participant: while the filter function is of priority with regard to the potential reader, the concern of a discipline is to improve the publication; the most important aspect for the author in the case of successful publication is his reputation.[36]

A useful categorization of quality assurance processes for research data can be found in the study, "To Share or not to Share" of the Research Information Networks (RIN). The study states:

"*The term 'quality' is conventionally associated with the notion of being 'fit for purpose'. With regard to creating, publishing and sharing datasets we identified three key purposes: first, the datasets must meet the purpose of fulfilling the goals of' the data creators' original work; second, they must provide an appropriate record of the work that has been undertaken, so that it can be checked and validated by other researchers; third, they should ideally be discoverable, accessible and re-usable by others. Fulfilling the first and second of these purposes implies a focus on scholarly method and content; the third implies an additional focus on the technical aspects of how data are created and curated."*[37]

4.1 QUALITY ASSURANCE PROCESSES CATEGORIES FOR RESEARCH DATA

Based on interviews with over 100 scientists, data managers and data experts, the RIN study identifies three categories of quality assurance process. Waaijers & Van der Graaf[38] took on this categorization in 2011 and drew up a description of the respective categories:

[34] Müller, U. T. (2008). Peer-Review-Verfahren zur Qualitätssicherung von Open-Access-Zeitschriften – Systematische Klassifikation und empirische Untersuchung. Berlin. Retrieved from http://nbn-resolving.de/urn:nbn:de:kobv:11-10096430

[35] Kronick, D. A. (1990). Peer Review in 18th-Century Scientific Journalism. JAMA: The Journal of the American Medical Association, 263(10), 1321-1322. doi:10.1001/jama.1990.03440100021002

[36] Regarding functions of peer review processes, refer to e.g. Müller, U. T. (2008). Peer-Review-Verfahren zur Qualitätssicherung von Open-Access-Zeitschriften – Systematische Klassifikation und empirische Untersuchung. Berlin. Retrieved from http://nbn-resolving.de/urn:nbn:de:kobv:11-10096430

[37] Research Information Network. (2008). To Share or not to Share: Publication and Quality Assurance of Research Data Outputs. Main report. Retrieved from http://www.rin.ac.uk/system/files/attachments/To-share-data-outputs-report.pdf

[38] Waaijers, L., & van der Graaf, M. (2011). Quality of Research Data, an Operational Approach. D-Lib Magazine, 17(1/2). doi:10.1045/january2011-waaijers

Quality assurance in the data creation process: In the first category, priority is given to method and data collection. The selection of method, work environment, tools used and calibration of instruments is of central importance.

Data management planning: The second category focuses on management of data. The objective of data management is to ensure permanent access to data. Reuse of such data is enabled as a result of exact description of the data and its process of formation.

Quality assessment of datasets: The third category addresses the "assessment of the scientific/scholarly quality of research data". Waaijers & Van der Graaf discuss the reviewing of data within the scope of peer review processes and refer to innovative publication strategies such as Data Publications. RIN summarizes the need for action in this sector as follows:

"Funders should work with interested researchers, data centres and other stakeholders to consider further what approaches to the formal assessment of datasets – in terms of their scholarly and technical qualities – are most appropriate, acceptable to researchers, and effective across the disciplinary spectrum."[39]

While measures in the first category vary in discipline and form of data, generic measures of data management can be identified in the second category.

4.2 DATA MANAGEMENT

A prime example here is the work of the Science and Technology Facilities Council (STFC), an APARSEN partner, which issued a Scientific Data Policy in 2011.[40] The guideline of the science organization is based on the Common Principles on Data Policy of Research Councils UK (RCUK). The following is stated in the principles of the RCUK:

"Institutional and project specific data management policies and plans should be in accordance with relevant standards and community best practice. Data with acknowledged long-term value should be preserved and remain accessible and usable for future research."[41]

To support the principles described, the STFC emphasizes the necessity for data management plans in its policy, in which the handling of data emerging within the course of STFC projects has to be described:

"Data management plans should exist for all data within the scope of the policy. These should be prepared in consultation with relevant stakeholders and should aim to streamline activities utilising existing skills and capabilities, in particular for smaller projects."[42]

Reference is made to the work of the Digital Curation Centre (DCC) as an example of such data management plans. This institution supports scientific institutions in curating digital research data. With DMP Online, the DCC provides "[a] flexible web-based tool to assist users to create personalized plans according to their context or research funder"[43]. This tool helps scientists to prepare a data management plan.

Such data management plans are not only issued on an institutional basis. The inclusion of data management plans is increasingly becoming the standard also for larger scientific projects. One such example is the project, TERENO[44], of the APARSEN partner Helmholtz Association. In this project, a

[39] Research Information Network. (2008). To Share or not to Share: Publication and Quality Assurance of Research Data Outputs. Main report. Retrieved from http://www.rin.ac.uk/system/files/attachments/To-share-data-outputs-report.pdf

[40] Science and Technology Facilities Council. (2011). STFC scientific data policy. Retrieved from http://www.stfc.ac.uk/Resources/pdf/STFC_Scientific_Data_Policy.pdf

[41] Research Councils UK. (2011). RCUK Common Principles on Data Policy (p. 2011). Retrieved from http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx

[42] Science and Technology Facilities Council. (2011). STFC scientific data policy. Retrieved from http://www.stfc.ac.uk/Resources/pdf/STFC_Scientific_Data_Policy.pdf

[43] Digital Curation Centre. (n.d.). Data Management Plans. Retrieved from http://www.dcc.ac.uk/resources/data-management-plans

[44] Extract from the project description: "TERENO is embarking on new paths with an interdisciplinary and long-term research programme involving six Helmholtz Association Centres. TERENO spans an Earth observation network across

project-specific Data Policy[45] describes the basic conditions for handling research data. This guideline is enhanced by a Data Management Plan.

4.3 QUALITY ASSESSMENT OF DATASETS

The discussion involving review processes for research data has gained importance in recent years. The demand for open access and potential re-use of data brings with it the question of how the quality of research data can be ensured and which contribution peer review processes can make to securing quality. In this context, the aforementioned RIN study states:

"Peer review may involve checking supporting data in a more or less detailed way. In some disciplines, reviewers check data extremely thoroughly and are capable of unearthing flaws or inconsistencies at this point. In other cases, checking is less than thorough, partly because reviewers may not be able to judge the data satisfactorily, partly because datasets may be too large to review in their entirety, and partly because the data may be too complex to be judged in this way. Reviewers may check that the data are present and in the format and of the type that the work warrants, and leave it at that. Overall the approach is uneven. There is a concern also that even if peers have the skills to review the scholarly content, they may not be able to judge the technical aspects of a dataset that facilitate usability."

In her very readable paper, "Scholarship in the Digital Age", issued in 2007 Borgmann describes the challenges of reviewing research data:

"For publications that report data, the data are implicitly certified as part of the peer-review process. Reviewing data in the context of a publication, however, is much different than assessing their accuracy and veracity for reuse. Reviewers are expected to assess the face validity of the data, but only in certain fields are they expected to recompute analyses, verify mathematical proofs, or inspect original sources. Only a few scientific journals require authors to provide the full data set."[46]

In Great Britain, the Science and Technology Committee published a comprehensive survey in 2011 on peer review of scientific publications. The survey includes the issue of „the need to review data":[47] in a consultation process under the key word of "Replication" the committee deals with the question of the extent to which the data on which a submitted publication is based can be reviewed. The committee states:

"[..] that reproducibility should be the gold standard that all peer reviewers and editors aim for when assessing whether a manuscript has supplied sufficient information, about the underlying data and other materials, to allow others to repeat and build on the experiments."

However, a precondition for potential replication of data is its accessibility. The report goes on:

"If reviewers and editors are to assess whether authors of manuscripts are providing sufficient accompanying data, it is essential that they are given confidential access to relevant data associated with the work during the peer-review process. This can be problematical in the case of the large and complex datasets which are becoming increasingly common."

Lawrence et al. place the following demands on the review of research data:

Germany that extends from the North German lowlands to the Bavarian Alps. This unique large-scale project aims to catalogue the longterm ecological, social and economic impact of global change at regional level. Scientists and researchers want to use their findings to show how humankind can best respond to these changes."

[45] TERENO. (2011). TERENO Data Policy. Retrieved from http://teodoor.icg.kfa-juelich.de/overview/downloads/TERENO Data policy.pdf

[46] Borgman, C. L. (2007). Scholarship in the Digital Age. Information, Infrastructure, and the Internet. Cambridge, Massachusetts: MIT Press.

[47] House of Commons. (2011). Peer review in scientific publications. Report, together with formal minutes, oral and written evidence. London. Retrieved from http://www.publications.parliament.uk/pa/cm201012/cmselect/cmsctech/856/856.pdf

"The data peer review procedure must ensure that all metadata is as complete as possible, but it must also address other qualities expected of [p]ublication class material, such as the data's internal self-consistency, the merit of the algorithms used, the data importance, and its potential impact."[48]

In addition, Lawrence et al. have developed a "Generic Data Review Checklist". The checklist is divided into three categories: "data quality", "metadata quality" and "general". Some questions are proposed for each of the categories, with the aid of which a dataset can be assessed. The focus is on completeness and correctness of the metadata.

In summary it can be said that the reviewing of research data results in challenges which affect scientific disciplines, their e-infrastructures (such as e.g. research data repositories) and publishers (as issuers of scientific publications).

[48] Lawrence, B., Jones, C., Matthews, B., Pepler, S., & Callaghan, S. (2011). Citation and Peer Review of Data: Moving Towards Formal Data Publication. International Journal of Digital Curation, 6(2). doi:10.2218/ijdc.v6i2.205

5 LOOKING INTO CURRENT PRACTICE

The following section examines and documents activities and positions of science, publishers and information infrastructures (e.g. research data repositories) in relation to the quality assurance of research data.

5.1 THE SCIENTIST'S PERSPECTIVE

In their function as author, reviewer and sometimes editor, scientists are confronted with recommended and obligatory guidelines in the publishing of scientific results which make comments on the handling of the data on which a publication is based. For example, the Committee on Publication Ethics (COPE) recommends the following to reviewers:

"Reviewers should be asked to address ethical aspects of the submission such as: […] Is there any indication that the data has been fabricated or inappropriately manipulated?"[49]

Practical implementation of quality inspection of data during the peer review process varies according to discipline. An inspection of data in post-mortem examination is probably pretty rare. The following is stated in the aforementioned RIN study, in which the methods of quality assurance of research data were examined and assessed in eight research sections in Great Britain:[50]

"There is no consistent approach to the peer review of either the content of datasets, or the technical aspects that facilitate usability."[51]

The attitude of scientists to peer review has been examined in several studies.[52] The most central studies are those by Mark Ware Consulting from 2008 and Sense about Science from 2009. Both studies provide a comprehensive picture on the attitude of scientists to the peer review processes.

The "Peer Review Survey 2009" by Sense about Science interviewed more than 4000 authors and reviewers concerning this issue. The survey investigated the view of scientists on the reviewing of research data which is the basis for a paper.

It is stated that reviewers require access to data in order to expose incorrect scientific behaviour:

"It is widely believed that peer review should act as a filter and select only the best manuscripts for publication. Many believe it should be able to detect fraud (79%) and plagiarised work (81%), but few have expectation that it is able to do this. Comments from researchers suggest this is because reviewers are not in a position to detect fraud, this would require access to the raw data or re-doing the experiment."[53]

The study cites, for an example the comment of a medical scientist who describes the challenge of accessibility:

"Similarly it would be very difficult for reviewers to detect fraud since they do not have access to primary data. If reviewers were expected to sift through primary data to detect fraud, this would take so much time that the entire process would grind to a halt and probably people would simply start declining requests for review."[54]

According to the study, reviewers and authors consider reviewing of the data to be impractical. A stagnation of the reviewing system is feared:

---

49 Committee on Publication Ethics. (2008). Guidance for Editors: Research, Audit and Service Evaluations.

50 Research Information Network. (2008). To Share or not to Share: Publication and Quality Assurance of Research Data Outputs. Annex: detailed findings for the eight research areas. Retrieved from http://www.rin.ac.uk/system/files/attachments/To-share-data-outputs-annex.pdf

51 Research Information Network. (2008). To Share or not to Share: Publication and Quality Assurance of Research Data Outputs. Main report. Retrieved from http://www.rin.ac.uk/system/files/attachments/To-share-data-outputs-report.pdf

52 An overview is provided by: Ware, M. (2011). Peer Review: Recent Experience and Future Directions. New Review of Information Networking, 16(1), 23-53. doi:10.1080/13614576.2011.566812

53 Sense about Science. (2009). Peer Review Survey 2009: Full Report. Retrieved from http://www.senseaboutscience.org/data/files/Peer_Review/Peer_Review_Survey_Final_3.pdf

54 Ibid

"[…] researchers point out that examining all raw data would mean peer review grinds to a halt."[55]

A slightly more positive view of this issue is provided in the study published in 2008 by Mark Ware Consulting. 3000 scientists were interviewed about their position on the topic of peer review. With regard to the reviewing of research data it is stated:

"A majority of reviewers (63%) and editors (68%) say that it is desirable in principle to review authors' data. Perhaps surprisingly, a majority of reviewers (albeit a small one, 51%) said that they would be prepared to review authors' data themselves, compared to only 19% who disagreed. This was despite 40% of reviewers (and 45% of editors) saying that it was unrealistic to expect peer reviewers to review authors' data. Given that many reviewers also reported being overloaded, we wonder, however, whether they would still be as willing when it actually came to examine the data."[56]

Both studies come to the conclusion that the potential of peer reviewing of data is recognized, but that considerable doubt exists with regard to practical execution on account of the work involved with such reviewing. This conclusion is verified by the aforementioned RIN study which surveyed more than 100 scientists in Great Britain:

"In summary, there is some sympathy with the concept of expert assessments of the quality of datasets, but researchers don't see how it might work in practice and, given that they are not unhappy with the present situation, there is no grass-roots pressure to introduce a formal assessment process."[57]

Waaijers & Van der Graaf published a study in 2011 based on interviews with sixteen "Data Professionals", which was enhanced by a broad questionnaire to more than 2800 university professors and lecturers. The implementation of peer review processes concerning research data was also highlighted in the interviews. The paper states the following:

"In general, the interviewees had their doubts about the feasibility of peer review in advance because of the demand it would make on the peer reviewer's time. It was also pointed out that such a system would lead to an unnecessary loss of time before the dataset could be made available. Some respondents thought that it was theoretically impossible to assess the 'scholarly merit' of a dataset in isolation; the dataset exists, after all, in the context of a research question."[58]

This evaluation is in line with the studies already mentioned. It is interesting that Waaijers & Van der Graaf observe a positive attitude among the "data professionals" with regard to the new publications strategies being established for scientific data:

"Finally, it was suggested that, rather than setting up a separate quality assessment system for data, one could create a citation system for datasets, which would then form the basis for citation indices. The thinking behind this was that citation scores are a generally accepted yardstick for quality."[59]

The results of the survey of university professors and lecturers confirm the assessment of the sixteen "data professionals" consulted. The questionnaires reveal reservation concerning peer review of research data:

"It is striking that the high score in all disciplines for extending the peer review of an article to the replication data published along with it is largely negated by the objections. The reason given in the explanations is the excessive burden on peer reviewers. It would seem that it is here that the peer review system comes up against the limits of what is possible."

The potential of accessibility of data and the opportunity of innovative publication formats for research data is also emphasized here:

---

55 Ibid

56 Mark Ware Consulting. (2008). Peer review in scholarly journals: Perspective of the scholarly community – an international study. Retrieved from http://www.publishingresearch.net/documents/PeerReviewFullPRCReport-final.pdf

57 Research Information Network. (2008). To Share or not to Share: Publication and Quality Assurance of Research Data Outputs. Main report. Retrieved from http://www.rin.ac.uk/system/files/attachments/To-share-data-outputs-report.pdf

58 Waaijers, L., & van der Graaf, M. (2011). Quality of Research Data, an Operational Approach. D-Lib Magazine, 17(1/2). doi:10.1045/january2011-waaijers

59 Ibid

"Scientists and scholars in all disciplines would welcome greater clarity regarding the re-use of their data, both through citations and through comments by re users. Setting up special journals for data publications is also popular in all disciplines."

Waaijers & Van der Graaf also ascertain a negative attitude among the scientists questioned to obligatory measures of data management:

"The view regarding a mandatory section on data management in research proposals is also unanimous, but negative. The decisive factor here is a fear of bureaucracy."

Summary:

The studies specified show a uniform picture of the perspective of scientists to the peer review of scientific data:

- Scientists recognize that accessibility of data is a precondition for peer review of it.
- In principle, reviewers and editors find it preferable for data to be peer reviewed but many reservations exist about its feasibility; peer review may grind to a halt.
- Scientists fear that reviewing data in the course of the peer review process is not practical due to the amount of work and time involved.
- Scientists have a positive attitude towards innovative publication strategies of research data and welcome greater clarity regarding the re-use of their data.
- Scientists are sceptical about obligatory measures of data management, since they fear bureaucracy.

5.2 THE DATA REPOSITORY'S PERSPECTIVE

To support scientists in handling the "data deluge"[60], scientific infrastructure facilities such as data centres and libraries are required to provide reliable e-infrastructures (such as e.g. research data repositories), on which data can be made permanently accessible. The High Level Expert Group on Scientific Data of the European Commission outlines the following vision of handling research data in 2030 in its strategy paper, "Riding the Wave", published in 2010:

"Producers of data benefit from opening it to broad access, and prefer to deposit their data with confidence in reliable repositories. A framework of repositories is guided by international standards, to ensure they are trustworthy."[61]

To create reliable data repositories designed in accordance with disciplinary requirements, infrastructure facilities aim to support and develop the certification and audit of repositories.[62] This concern is also pursued within the APARSEN project.

The relevance of infrastructure facilities and their services for the quality assurance of scientific data is also emphasized in the "GRDI2020 Roadmap Report" published in 2011. This stresses data management as a precondition for high quality data:

"If research data are well organized, documented, preserved and accessible, and their accuracy and validity is controlled all times, the result is high quality data, efficient research, findings based on solid evidence and the saving of time and resources."[63]

The e-IRG Report on Data Management published in 2009 is more specific. According to experts, e-infrastructures are "the main advocates of quality assurance for research data".[64] The expert group specifies the following measures of repositories for quality assurance of stored data:

- checking the format of the data files
- checking whether a complete code book is available for coded data
- checking the anonymity of personal data; data are de-identified by expunging names, addresses, etc.
- checking for missing values and overall completeness / data integrity
- checking for consistency

The contribution made by data repositories is also underlined by the Research Information Network (RIN) in a study published in 2011 concerning the status of data centres in Great Britain:

"The curatorial role of the centre thus affects two important elements of data quality: first, ensuring that individual datasets are academically 'good' (as much as it can) and second, ensuring that it creates and preserves collections which can be a useful starting point for new research."[65]

This evaluation clearly shows that data repositories support quality assurance of research data via two complementary measures:

- via selection of data during the recording process and
- via curatorial measures of data management.

[60] Hey, A. J. G., & Trefethen, A. E. (2003). The Data Deluge: An e-Science Perspective. In F. Berman, G. Fox, & A. J. G. Hey (Eds.), Grid Computing - Making the Global Infrastructure a Reality (pp. 809-824). Chichester: Wiley and Sons. Retrieved from http://eprints.ecs.soton.ac.uk/7648/

[61] High Level Expert Group on Scientific Data. (2010). Riding the wave. How Europe can gain from the rising tide of scientific data. Retrieved from http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

[62] Klump, J. (2011). Criteria for the Trustworthiness of Data Centres. D-Lib Magazine, 17(1/2). doi:10.1045/january2011-klump

[63] GRDI2020. (2011). Global Research Data Infrastructures: The GRDI2020 Vision. Retrieved from http://www.grdi2020.eu/Repository/FileScaricati/6bdc07fb-b21d-4b90-81d4-d909fdb96b87.pdf

[64] e-Infrastructure Reflection Group, & European Strategy Forum on Research Infrastructures. (2009). e-IRG report on Data Management. Retrieved from http://www.e-irg.eu/images/stories/e-irg_dmtf_report_final.pdf

[65] Research Information Network. (2011). Data centres. Their use, value and impact. Retrieved from http://www.jisc.ac.uk/news/stories/2011/09/~/media/Data%20Centres-Updated.ashx

Up to now there have been only few interdisciplinary studies made on data repositories. The contribution of repositories to quality assurance, in particular, has found little interest up to now. The studies of the Research Information Network (RIN) are particularly useful. In its study, "To Share or not to Share" published in2008, RIN states:

"Data centres apply rigorous procedures to ensure that the datasets they hold meet quality standards in relation to the structure and format of the data themselves, and of the associated metadata. But many researchers lack the skills to meet those standards without substantial help from specialists."[66]

This evaluation is supported by the study on the role of data centres in Great Britain published in 2011. In the course of this study, research sponsors and users of five UK data centres were questioned on the work of these infrastructure facilities; the contribution made by data centres to quality assurance was also examined:

"There were high levels of agreement across all data centres with most of the statements about research benefits. Benefits to do with research efficiency were the most widely supported, with researchers mentioning ways in which the centres had saved them time, money and effort. Benefits to do with research quality related both to the quality of their own work, and the quality of the data that they access from the centre in order to undertake such work. In both cases, the data centres are perceived to add quality. Researcher training was more important in some centres than others."[67]

An internal survey was made in the course of the APARSEN work package „Annotation, Reputation and Data Quality". This survey included i.a. the examination of measures of data repositories within the APARSEN network with regard to quality assurance of stored research dat. 20 partners took part in the survey. The following measures of quality assurance were specified in a free text response:

- Business process documentation
- Completeness / Consistency checks
- Data curators technical review (methods, parameters, unit checks, consistency)
- Data management and sharing training
- File format validation
- Metadata checks
- Risk management
- Storage integrity verification
- Tools for annotating quality information

The following examples document the contributions of three data repositories to quality assurance of research data:

Example 1: The World Data Center for Marine Environmental Sciences (WDC-MARE), which is operated by the Alfred Wegener Institute for Polar and Marine Research (AWI) and the University of Bremen, secures the quality of stored data in an editorial process organized by the research data repository PANGAEA[68] and its staff:

"The PANGAEA data editorial ensures the integrity and authenticity of your data. Data might be submitted in the authors' format and will be converted to the final import and publication format. The PANGAEA editors will check the completeness and consistency of metadata and data. Our editors are scientists from the earth and life sciences. We may identify potential problems with your data (e.g. outliers). Nevertheless, we will only take full responsibility for the technical quality. You will be responsible for the scientific quality of your data (e.g. the validity of used methods). After data have been archived you will receive a DOI name and you are requested to proof-read before the final version is published. In case your data are supplementary to a journal article you might reference the data in the article. In addition our web services allow for embedding data references dynamically on

[66] Research Information Network. (2008). To Share or not to Share: Publication and Quality Assurance of Research Data Outputs. Main report. Retrieved from http://www.rin.ac.uk/system/files/attachments/To-share-data-outputs-report.pdf

[67] Research Information Network. (2011). Data centres. Their use, value and impact. Retrieved from http://www.jisc.ac.uk/news/stories/2011/09/~/media/Data%20Centres-Updated.ashx

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

the article splash page [...]. In case there is a moratorium on your data you can ask for access constraints."[69]

Example 2: The World Data Centre for Climate (WDC Climate) at the German Climate Computing Centre (DKRZ) secures quality of research data in a two-stage process. A differentiation is made between a technical and a scientific review of data. In the course of the "Scientific Quality Assurance (SQA)", the quality of data is inspected within the scope of a documentation process. This process is supported by a "web-based software system". During this inspection of data, the following conditions i.a. are checked:

- number of data sets is correct and > 0
- size of every data set is > 0
- the data sets and corresponding metadata are accessible
- the data sizes are controlled and correct
- the spatial-temporal coverage description (metadata) is consistent to the data, time steps are correct and the time coordinate is continuous
- the format is correct
- variable description and data are consistent

The two quality assurance processes may vary depending on the form and format of data. After successful completion of the two consecutive processes, the dataset is addressed with a Digital Object Identifier (DOI).[70]

Example 3: The APARSEN partner, Data Archiving and Networked Services (DANS), has enabled commenting of datasets stored in the „online archiving system EASY since 2010 in accordance with pre-defined criteria. EASY enables access "to thousands of datasets in the humanities, the social sciences and other disciplines. EASY can also be used for the online depositing of research data"[71]. The assessment of a dataset becomes visible for the user, if two assessments have been submitted for a dataset.

69 Refer to: http://www.pangaea.de/submit/

70 Based on data at http://www.dkrz.de

71 https://easy.dans.knaw.nl/

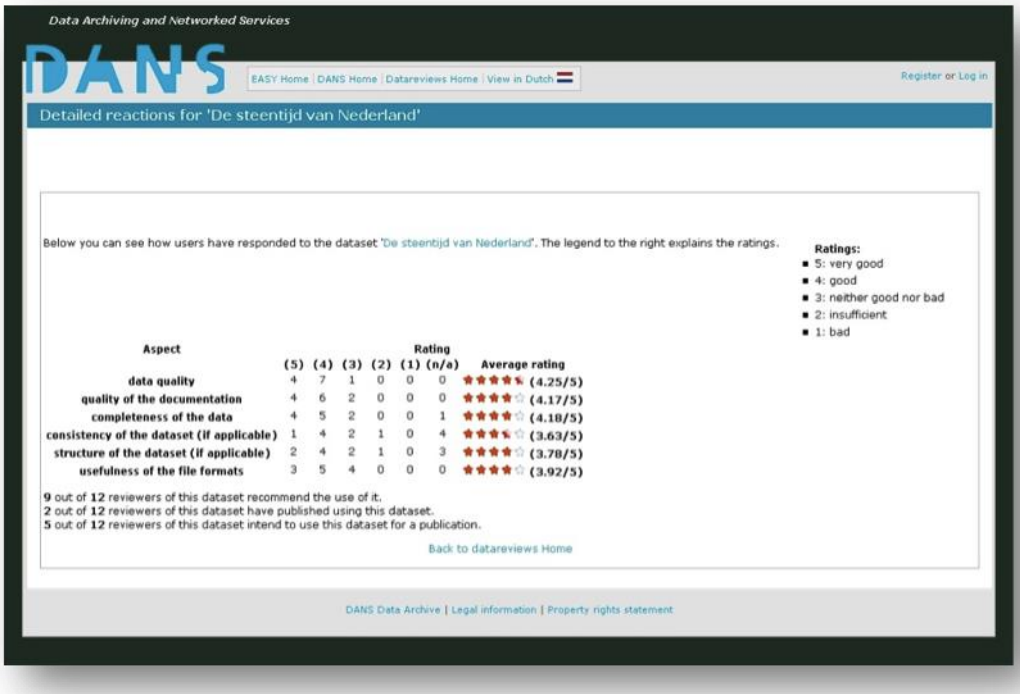

Figure 2 Assessment of the dataset "De steentijd van Nederland"

In 2011 DANS published an appraisal of 280 reviews. The study 2011 emphasized the significance of quality of data:

"The average scores for the aspects of the datasets surveyed are around 4 on a scale from 1 to 5, with 'quality of the data' ranking first (4.14). This is something to be pleased with. Among the researchers (57% of the respondents), most averages are even a fraction higher. As many as 91% of the respondents would recommend the dataset to others; this gives a strong impression of the quality of the datasets."[72]

Activities of the three data centres, WDC MARE, WDC Climate and DANS make it clear that e-infrastructures provide a contribution to quality assurance of data. In addition these institutions aim to secure and improve the quality of their services in the course of certification and audit.

Summary:

Up to now only few studies have been conducted on the activities of repositories in the field of quality assurance of scientific data. The studies specified and the examples of the three data repositories can be summarized as follows:

- Data repositories make a contribution to quality assurance of stored data.
- Data management is assessed as an essential contribution to quality assurance of data. The selection process and subsequent verification of data (via persistent addressing) is seen as very important.
- The measures contributed by repositories to quality assurance vary depending on the form, scope and discipline of data.
- Certification and audit secure the quality of data repositories and affect the quality assurance of data.

[72] Data Archiving and Networked Services. (2011). Data Reviews. Peer-reviewed research data. Retrieved from http://www.dans.knaw.nl/en/content/categorieen/publicaties/dans-studies-digital-archiving-5

5.3 THE JOURNAL'S PERSPECTIVE

Publishers and editors of scientific journals are increasingly looking for ways in which research data underlying the claims of a paper can be made permanently accessible. Robert Campbell and Cliff Morgan of John Wiley & Sons formulate the challenge of handling scientific data as follows:

"The real challenge is how to deal with the growth in research data that sits behind the journal article. Policies for data curation and sharing are emerging but there is no related peer review process or quality control."[73]

Editorial policies of scientific journals increasingly include statements concerning the handling of articles which form the basis of a publication. [74] For example, the editorial policy of the Nature journal family states:

"[...] condition of publication in a Nature journal is that authors are required to make materials, data and associated protocols promptly available to others without preconditions."

In addition, notes are provided concerning subject-specific features.[75]

An explicit "Policy on Referencing Data in and Archiving Data for AGU Publications" applies for publications of the American Geophysical Union (AGU). This describes concrete requirements of data repositories and the citation of research data.[76]

In the open access journal PLoS ONE of the Public Library of Science (PLoS) the following is stated in the section "Sharing of Materials, Methods, and Data" in the editorial policy:

"PLoS is committed to ensuring the availability of data and materials that underpin any articles published in PLoS journals."

In addition, references are made to appropriate research data repositories.[77] However, an article that appeared in PLoS in 2009 and in which the availability of such data was examined, found that only in one case out of 10 the underlying research data to a paper were indeed available.[78]

A similar commitment is already practice in some specialist fields of the life sciences. For example, the following sentence is to be found in the policy of the journal, Cell:

"One of the terms and conditions of publishing in Cell is that authors be willing to distribute any materials and protocols used in the published experiments to qualified researchers for their own use."

At Cell, for example, the nucleotide and protein sequences on which a paper is based must be accessible in appropriate repositories, such as e.g. the Worldwide Protein Data Bank (wwPDB), without restriction as of the time of publication of a paper and must be identifiable by way of specification of an "accession number".[79]

The data policies specified in the aforementioned examples illustrate the importance of interplay between journal and research data repositories. This cooperation takes a different form depending on the respective publication model of the research data. Categorizations of publication models can be

[73] House of Commons. (2011). Peer review in scientific publications. Report, together with formal minutes, oral and written evidence. London. Retrieved from http://www.publications.parliament.uk/pa/cm201012/cmselect/cmsctech/856/856.pdf

[74] Refer to: Pampel, H. & Bertelmann, R. (2011) „Data Policies" im Spannungsfeld zwischen Empfehlung und Verpflichtung. In S. Büttner, H.-C. Hobohm, & L. Müller (Eds.), Handbuch Forschungsdatenmanagement (pp. 49-61). Bad Honnef: Bock + Herchen. Retrieved from http://opus.kobv.de/fhpotsdam/volltexte/2011/228/

[75] Nature. (2009). Guide to Publication Policies of the Nature Journals. Retrieved from http://www.nature.com/authors/gta.pdf

[76] American Geophysical Union. (1996). Policy on Referencing Data in and Archiving Data for AGU Publications. Retrieved from: http://www.agu.org/pubs/authors/policies/data_policy.shtml

[77] PLoS ONE. (n.d.). PLoS ONE Editorial and Publishing Policies. Sharing of Materials, Methods, and Data. Retrieved from http://www.plosone.org/static/policies.action#sharing

[78] Savage, C. J., & Vickers, A. J. (2009). Empirical Study of Data Sharing by Authors Publishing in PLoS Journals. PLoS ONE, 4(9), e7078. doi:10.1371/journal.pone.0007078

[79] Cell. (2011). Information for Authors. Retrieved from http://www.cell.com/authors

found in papers by Dallmeier-Tiessen[80] and Lawrence et al.[81] Three publication models are focused on in Dallmeier-Tiessen:

- Publication of research data as an independent item in a repository.
- Publication of research data with textual documentation, as a so-called data publication.
- Publication of research data as an enrichment of an interpretative text publication.

These categorizations, especially for the second model, have been further refined in the Data Publications pyramid as developed within project-ODE and as mentioned in chapter 3 of this paper:

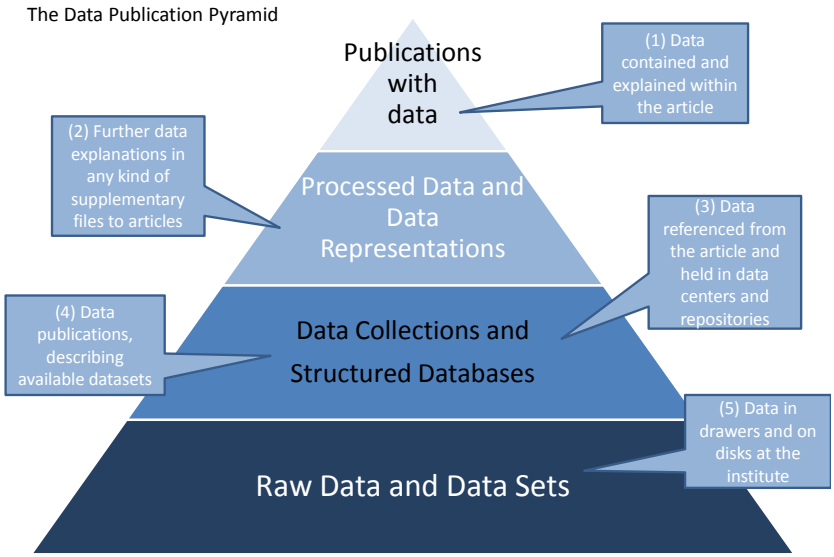

Figure 3 The "Data Publications Pyramid"

The following section deals with peer review for any of the categories 1 (data explained within an article), 2 (further data in supplements to a journal article), 3 (data referenced from an article and held in a repository) and 4 (data publications, describing available data sets). Experts have been asked for their opinion on the current and desirable implementation of peer review of data in these categories.

Data publications have a long tradition, especially in geological science. For example, the American Geophysical Union (AGU) and the Ecological Society of America (ESA)[82] have published data papers in their journals for a long time. The dataset described are checked in the course of a peer review process. In this context, the AGU states the following:

"Data sets that are the basis of data papers are subject to review. A sample of these data sufficient for the review process must be supplied with the submission of the paper. The reviewer is expected to comment on the data as if they were an integral part of the paper and on their usability."[83]

The concern of data publications is to document research data in a quality assured form in order to enable re-use. Chavan & Penev define this type of publication under the term, Data Paper:

[80] Dallmeier-Tiessen, S. (2011). Strategien bei der Veröffentlichung von Forschungsdaten. Retrieved from http://www.ratswd.de/download/RatSWD_WP_2011/RatSWD_WP_173.pd

[81] Lawrence, B., Jones, C., Matthews, B., Pepler, S., & Callaghan, S. (2011). Citation and Peer Review of Data: Moving Towards Formal Data Publication. International Journal of Digital Curation, 6(2). doi:10.2218/ijdc.v6i2.205

[82] Ecological Society of America. (n.d.). Data papers, supplements, and digital appendices for ESA journals. Retrieved from http://www.esapubs.org/archive/default.htm

[83] American Geophysical Union. (1996). Policy on Referencing Data in and Archiving Data for AGU Publications. Retrieved from http://www.agu.org/pubs/ authors/policies/data_policy.shtml

"We define a data paper as a scholarly publication of a searchable metadata document describing a particular online accessible dataset, or a group of datasets, published in accordance to the standard academic practices."[84]

Chavan & Penev go on to attribute three characteristics to this publication type:

"[...] to provide a citable journal publication that brings scholarly credit to data publishers; to describe the data in a structured human-readable form; and to bring the existence of the data to the attention of the scholarly community."

A new phenomenon that has appeared in the publishing landscape is that some publishers have recently started to set up dedicated data publications in the form of independent data journals. These include i.a. BioMed Central (GigaScience, Open Network Biology) and Copernicus Publication (Earth System Science Data). Elsevier launched two data journals: Nuclear Data Sheets[85] and Atomic Data and Nuclear Data Tables[86].

Other publishers announced the founding of data journals at the end of 2011: e.g. Faculty of 1000 (F1000 Research)[87], Pensoft Publishers (Biodiversity Data Journal)[88], Ubiquity Press (Open Archaeology Data)[89].

Many of these journals are still at an early stage of development and there are only a few in comparison to the vast numbers of thousands and thousands of traditional journals. But their emergence is something worthwhile to watch as many believe that the peer review of data(sets) will take place with much more rigour for these journals than for traditional research journals.

So-called data journals explicitly support the quality assurance of data. In the following the journal Earth System Science Data (ESSD) is used as an example to document the reviewing process of a data journal. This journal is one of the pioneers in its sector.

Earth System Science Data (ESSD) is a geo-scientific open access journal published by Copernicus Publication.[90] The journal publishes textual descriptions of datasets which have to be published on an appropriate repository[91]. The publishers formulate the focus of the journal as follows:

"The articles in this journal should enable the reviewer and the reader to review and use the data, respectively, with the least amount of effort. To this end, all necessary information should be presented through the article text and references in a concise manner and each article should publish as much data as possible. The aim is to minimize the overall workload of reviewers, e.g., by reviewing one instead of many articles, and to maximize the impact of each article."[92]

Articles submitted are published in the course of an "innovative two-stage publication process". After a brief check by an editor, the article is published on the website of the journal as a working paper. In this status, specialists can submit comments on the article. In addition to the comments of the community, two reviewers check the article and the research data made accessible on a data repository in accordance with a list of criteria. In this process, the following questions are focused on:[93]

---

84 Chavan, V. & Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. BMC Bioinformatics, 12(15), S2. doi: 10.1186/1471-2105-12-S15-S2

85 http://www.sciencedirect.com/science/journal/00903752

86 http://www.sciencedirect.com/science/journal/0092640X

87 Chan, A. (2011). F1000 Research. Retrieved from http://blog.f1000.com/2011/11/24/ismpp-2011-–-what's-next-for-f1000/

88 Chavan, V. & Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. BMC Bioinformatics, 12(15), S2. doi: 10.1186/1471-2105-12-S15-S2

89 http://www.openarchaeologydata.com

90 A detailed description can be found at: Pfeiffenberger, H. & Carlson, D . (2011). "Earth System Science Data" (ESSD) - A Peer Reviewed Journal for Publication of Data. D-Lib Magazine, 17(1/2). doi: 10.1045/january2011-pfeiffenberger

91 Refer to: Earth System Science Data. (n. d.). Repository Criteria. Retrieved from http://www.earth-system-science-data.net/general_information/repository_criteria.html

92 Refer to: Earth System Science Data. (n. d.). About this Journal. Retrieved from http://www.earth-system-science-data.net/general_information/about_this_journal.html

93 Earth System Science Data. (n. d.). Review Guidelines. Retrieved from http://www.earth-system-science-data.net/review/ms_evaluation_criteria.html

1. Read the manuscript: Is the article itself appropriate to support the publication of a dataset?
2. Check the data quality: Is the dataset significant – unique, useful and complete?
3. Consider article and dataset: Is the dataset itself of high quality?
4. Check the presentation quality: Is the dataset publication, as submitted, of high quality?
5. Finally: By reading the article and downloading the dataset would you be able to understand and (re-)use the dataset in the future?

The reports of the reviewers and comments contributed by specialists are published together with the working paper. The authors can respond to the reports and comments. After consideration of all reports and comments, the papers are published in ESSD.

5.3.1 EXPERT OPINIONS ON PEER REVIEW OF DATA

Several publishers and editors of journals were consulted on the topic of peer review of research data in the course of the APARSEN project. The aim of such consultation was to find out the views of the publishers and editors concerning this issue. The survey was made from November to December 2011 among 13 publishing experts. Three publishers and four editors supported the APARSEN project with their responses. The following issues were dealt with in the consultation:

a) Availability of underlying research data

b) Peer review of underlying research data

c) Checking quality and soundness of research data

d) Willingness of reviewers

e) Publishing data in data journals

Answers were anonymized. The answers can be summarized as follows, divided into the five topics:

a) Availability of underlying research data

The first group of questions involves the availability of and accessibility to research data. The question made a distinction between availability to the reviewers (underscribed by all respondents) and availability to readers (preferred by most but sometimes limited by privacy or confidentiality issues). All answers refer to the positive potential of open accessibility of data. However, the reservation of many authors is also mentioned:

"Unfortunately, we have still lots of authors with doubts in terms of sharing data, even if is "only" for the reviewers. But we are convinced that such an "open data model" would be of great benefit for science." (response of a publisher)

The importance of standards is specified as a precondition for open accessibility of data:

"Until scientific fields establish standards and formats to make it simple to process and interpret data from other laboratories there is little point in attempting to distribute those data." (response of an editor)

Reference is also made to the different publication strategies and to the importance of reliable and permanent infrastructures:

"If a scientific community or journal requires access for reviewers to underlying data then the publisher, as a service provider, should support these requirements. For relatively small datasets data can be provided as additional files but publishers are not best-placed to store large datasets in the long-term, as this might cause potential problems for mirror repositories of publishers' content." (response of a publisher)

In addition, reference is made to the central role of the editor and the Editorial Policy:

"It is the editor's task to ensure that the referees have all the information they need to provide a review, including data that are necessary to review the paper. If an author does not make data available that are required for a meaningful review of a paper (upon request) we do not consider the associated paper for publication." (response of an editor)

And editors are not reluctant to stick to the rules if authors are hesitant in making the data available:

"If an author did not comply with a reasonable request for data, I would not hesitate to reject the article". (response of an editor)

b) Peer review of underlying research data

The main issue addressed in this range of questions is the current practice applied in quality assurance of data. The answers make it clear that reviewing practices are very community-specific and vary depending on the respective discipline and form of data. The majority of responses indicate that reviewing of data is not a standard in reviewing practice.

"All think this should be done, but it is difficult to force reviewers to do that." (response of an editor)

The peer review of publications is focused first and foremost on checking the claims made in the paper, not on the available research data or its quality:

"I expect that more peer reviews focus on the content of the paper than on the underlying data. An expectation that every reviewer should re-analyze all data sets and supporting information would be unrealistic. When data are available as additional files data are at least available if the reviewer or editor wishes to dig deeper." (response of a publisher)

"It is unclear what exactly would entail "peer-review of underlying data". Underlying data are made available to reviewers; reviewers' focus is on assessing the robustness, novelty and importance of the main conclusions of the papers, in the light of the evidence (including data) presented."(response of an editor)

Those responding referred to the time and work involved in reviewing and emphasized the necessity for effective reviewing processes. Publishers and editors are of the opinion that assistance and support for the reviewers in quality control of data are an essential precondition for the reviewing of data. The following two responses are typical of several similar comments on this topic:

"The main challenges are to define the review criteria in a way that a non-paid reviewer is willing (not only is possible) to review the data and to reach the balance between the time to be spent to review data in depth on the one hand, but to keep the efforts for the review short on the other hand. Reviewing data in depth is a great challenge. We have to find criteria and methods to allow reviewers to do a good review on data with moderate efforts and time." (response of a publisher)

"Whether to ask for underlying data and consider reviewing them is at the discretion of reviewers. If it is done, it usually leads to longer review times and periods until publication. The main challenge to a general review of data is the additional burden it imposes on reviewers." (response of a publisher)

In addition, scepticism is expressed by those involved in journals that cover a wide range of science as to whether recommendations can be formulated on the reviewing of data for each discipline and each data type. The necessity for appropriate repositories where the data are stored and to which the publications can link bidirectionally is also mentioned to in the responses submitted on these questions.

c) Checking quality and soundness of research data

This part of the consultation deals mainly with practical implementation of the review. Questions included i.a. how intensively data can be reviewed. Depending on respective disciplines, the answers describe different processes in which, however, the reviewing of data is always merely a supplement to the checking of the actual article. None of the responses describe a detailed reviewing process of data. The following response is typical for similar comments:

"We would expect reviewers to check the plausibility of a data set, in the light of their knowledge, and to assess whether the authors' methods, as given in the paper, are state-of-the-art and robust. We would also expect reviewers to note if error bounds are unrealistically low, in the light of what's been done, or if there are just not enough data to support the conclusion." (response of an editor)

But in some disciplines, the responses are much more cautious, in one case even negative about the possibility to properly peer review the data for all its quality facets:

"The quality of data depends in large part on the design of the experiment design, which is already assessed by current peer review. The quality of the data also depends in large part on what was actually done in the lab as the data were collected (the practical skill of the scientists), and that will remain inaccessible without inspectors in every laboratory. Reviewing the digitized data will require a lot of effort to achieve a review that is more complete, but which remains incomplete." *(response of an editor)*

Reference is also made in this group of questions to the basic precondition of reviewing: accessibility of data in a reproducible form:

"For data to be understandable by an external reviewer then it needs to meet certain standards. Data formatting standards can be field type specific. But at a basic level authors would need to provide a clean (i.e. reviewed systematically for duplicates, errors, and missing data etc), well annotated (sufficient information/metadata is given about each variable to allow replication of the originally published results) dataset in a suitable format so that statistical analyses could be conducted." (response of a publisher)

"Data can only be reviewed properly when all underlying metadata, experiment conditions, etc. are fully shared with reviewers. This requires high standards on data sharing. To share data and to review them is certainly beneficial to science, at the same time it puts additional strain on researchers. This needs to be compensated with incentives (acknowledge the efforts for making data including appropriate metadata available; acknowledge the additional work in reviewing them)." (response of a publisher)

d) Willingness of reviewers

Further questions of the consultation deal with the willingness of reviewers and the requirements made of them. Reference is made to the importance of the selection of reviewers in the responses. The majority of those responding are of the opinion that special data reviewers are not required. However, the reviewer does need to have experience in handling the form and format of the respective data:

"I don't see why a different type of reviewer would be required specifically for research data, although referees with different backgrounds and expertise are selected to look at different aspects of a paper." (response of an editor)

"In the vast majority of cases peer review of data does not require a different type of reviewer. The majority of researchers are extremely capable in assessing research data within their own area of expertise and consequently can manage this alongside the review of the article in question. The majority of reviewers are willing to review data alongside the article as the normal process of peer review." (response of a publisher)

Other respondents question the willingness of reviewers to step into the daunting task of assessing the quality of the underlying research data:

"Some are. Many perhaps not yet." (response of an editor)

"As a scientist, I would never consent to the tedious task of trying to see whether I can spot problems in raw data from another lab. As an editor, I would never ask reviewers to take on this burden. We have enough trouble recruiting the best possible reviewers to evaluate experimental designs, analyses and interpretations."*(response of an editor)*

e) Publishing data in data journals

In this section questions involve the opinion of editors and publishers to the dynamically developing field of the data journals. Questions focus on the requirements of reviewing data publications.

The majority of editors and publishers surveyed welcome the setting up of data journals:

"Of course, a positive development, if they can make high-quality data available in a useful and accessible form to the community." (response of an editor)

"Perhaps if new journals are specifically set up and funded with the resources and time to pore over all research data, even high-throughput data sets, as well as doing the kind of peer review other journals are doing already, then that would be beneficial." (response of a publisher)

However, responses show that this type of publication is still at an early stage of development. Responses recognize the potential of such publications, but are reserved concerning the evaluation of future development of the data journals.

A similar picture emerges for the assessment of the review process of "data papers":

"Where publication of a dataset is the primary purpose of a scholarly article, such as in the case of a data note, then it would be reasonable to infer a greater expectation of peer review of the related data." (response of an editor)

"It's not entirely clear that reviewing a set of data without a paper is the same as reviewing a paper with claims/arguments built upon data." (response of an editor)

"Because in a data journal there are no main conclusions to be assessed, a different guideline would be needed for referees. I am neither sure what that would be (just check the plausibility of the data and their error bounds?), nor whether that would be more rigorous. My guess would be less rigorous, because the question asked is pretty general." (response of an editor)

These three responses show that the majority of those surveyed feel there are still many issues to be dealt with concerning quality assurance of data publications.

Summary:

On account of the early stage of development of data journals, there are no studies available on quality assurance for this type of publication. A view of people involved in traditional journals and their practice in reviewing underlying data to publications, as well as the results of consulting publishers and editors, can be summarized as follows:

- Several journals require in their editorial policies the availability and accessibility of data, especially in the life sciences.
- Peer review of underlying research data is not always included in the standard peer review process of journals. Practice can vary between science disciplines and can be community-specific.
- In the peer review of publications, the main focus is on checking the claims and conclusions of the article. Peer review of underlying data plays a supportive role in this if and when useful to the reviewer.
- In order to organize the reviewing of data effectively, clearly defined criteria are essential. A lack of technical standards and of common practice impedes reviewing of data in some disciplines.
- Respondents believe that for the review of research data no special type of reviewers would be necessary, but…
- Full review of research data may put an unrealistic burden on the task of current peer reviewers.
- Publishers and editors have positive expectations of the development of data publications. They also expect that more in-depth peer review of data will take place for so-called data journals.

6 UPCOMING RESEARCH AREAS

The following section future research issues are specified which have been identified in Chapter 4, grouped around four groups of stakeholders: scientists, data repositories, journals and providers of tools.

Scientists
Although some studies exist concerning the attitude of scientists to this topic, a number of questions remain open. Research is particularly required in the following issues:

- Methods: Which different methods of quality assurance exist and how are they implemented in the respective discipline?
- Reputation: How can the extensive work required in reviewing research data be integrated into the scientific reputation system?
- Standards: How can standards be developed and established in order to ensure simple and effective organization of reviewing of research data?

Research data repositories:
The contribution made by research data repositories to quality assurance finds widespread recognition. The specific access modalities and services of the repositories can support quality assurance of data in many ways. However, since e-infrastructures are often in an early stage of development, many questions concerning e-infrastructures still need to be answered. To support the development, standardization and networking of repositories, there is need for research in the following fields:

- Audit and Certification: How can audit and certification of research repositories be supported in order to drive forward the development of reliable locations to secure data?
- Service: How can research data management be designed in line with scientific requirements? Which services are required to avoid "data bureaucracy"?
- Networking: How can the networking of existing research data repositories be supported in order to create a network of reliable data repositories?
- Reviewing: Which innovative measures and tools can be supplied by data repositories to support the reviewing of research data?
- Must these tools be preserved and if so, how?
- Cooperation with journals: How can cooperation between repositories and journals be ideally created, in organizational and technical terms, to enable and support quality assurance of data?

Journals:
The inclusion of statements concerning accessibility of research data in editorial policies shows that publishers and editors are willing to deal with the issue of modern handling of research data. This is illustrated in the positive reactions to the new phenomenon of setting up data journals. On account of the early stage of development of this type of publication, but also to be able to handle underlying research data more effectively in the context of traditional journals, research is required in the following fields:

- Effectiveness: How can peer review of research data be organized effectively?
- Criteria: How can criteria for the peer review of research data be formulated in specific disciplines?

- Cooperation with repositories: How can cooperation between repositories and journals be ideally created, in organizational and technical terms, to enable and support quality assurance of data?
- Standards: which standards can journals implement for reproducible publication of research data?
- Integration: How can research data be integrated more seamlessly into publications?

Tool builders and e-infrastructures:

If data are to be incorporated more deeply - or even seamlessly - into publications, a whole dimension of research is opened. If there need to be "executable papers", then research must in particular address the problems associated with the fact that journals – and now data repositories – provide the records of science. At first glance, "providing the records" and the dynamisms implied by the term executable paper, appear to be mutually exclusive. A more mundane, but absolutely essential question is that of simplicity: The concepts as well as the tools themselves must be easily understood and operated to ensure uptake as well as reliable use.
Some of the questions would appear to be:

- If articles enable a range of viewpoints and parameters, does peer review pertain to the whole range or just to the initial viewpoint and parameters offered? How to reference an article with a specific (non-initial) set of viewpoints/parameters?
- Who will be responsible for the reliable and authentic operation of executable elements within articles? The publisher, the data repository or a third party?
- How to preserve the records of science? Current preservation techniques do not reach much beyond a fair number of text formats, falls short on most data formats – not to speak of keeping alive and "authentic" an executable environment.
- Should there be a sensible time limit on the (authentic) preservation of the executable elements, e.g. 1 or 10 years? After that, what remains to be preserved of those elements - a static view?
- Can there be generic tools (toolkits) and answers to all those questions or are they dependent on discipline / community of practice?

7 CONCLUSIONS

The discussion concerning modern handling of research data and its permanent accessibility is being conducted on many levels. Science organizations and sponsors are becoming more involved in the way research data is handled. Their aim is to support verifiability and re-usability of data. In the course of producing data management plans which are especially demanded by sponsor organizations, scientists are required to describe the handling of data after a project has been completed. Infrastructure facilities are involved in the setting up of research data repositories, on which data can be made accessible in a reliable environment. Publishers are starting to found data journal publication organs which, in cooperation with repositories, support publication of research data in a reusable form.

A precondition for quality control of research data is accessibility of such data. According to the study, "To Share or not to Share" of the Research Information Network (RIN), quality assurance of research data can be divided into three categories:

Quality assurance in the data creation process: To secure quality of data during data collection, scientists are required to apply methods and tools in a qualified and professional manner.

Data management planning: Infrastructure facilities such as libraries and data centres can contribute to quality assurance of data via measures of research data management, by providing repositories on which data can be made accessible in a reusable form.

Quality assessment of datasets: Publishers and journals can support quality assurance of data by demanding specific handling of data which form the basis of an article (e.g. within editorial policies). In addition, they can contribute to quality assured publication of research data by operating data journals in cooperation with repositories.

The following section summarizes the documented activities of the three participants and their contributions to quality assurance. In addition, future challenges are also described:

a) Scientists

Scientists are of the opinion that reviewing of data supports quality. However, the majority fear that the reviewing of data within Peer reviews is not feasible on account of the work load for the reviewer. It is ascertained that reviewing is only possible if data is accessible in a reusable form. Scientists are sceptical about obligatory measures of data management because they fear "data bureaucracy". Innovative publication strategies such as Data Publications are considered to be a positive contribution.

- To support quality assurance of data standards have to be developed in many disciplines which enable exchange of data in a reproducible form.
- Interdisciplinary exchange of methods of quality assurance of research data can help in disciplines which do not have fixed methods of establishing processes for quality assurance.
- Quality assurance of data is a time-consuming activity which is not accordingly recognized within scientific reputation systems. The development of incentive and reward systems can help to increase recognition for such work.

b) Research Data Repositories

Data repositories make an important contribution to quality assurance of stored data. The storing of research data on reliable repositories is seen as an important measure of quality assurance. Inclusion of these e-infrastructures in everyday scientific work as well as their contribution to quality assurance varies according to the respective discipline.

- To support scientists in quality assurance of data it is necessary to establish discipline-specific services of data management which are in line with scientific requirements. Cooperation with publishers in developing data journals is of immense importance.

- The contribution of individual repositories to quality assurance of data has not yet been widely identified. A comprehensive appraisal of the situation regarding data repositories can make a contribution here.[94]
- The selection and verifiability of data in standardized form is attributed great importance within data management. The continued development of this process is very important
- Audit and Certification secure the quality of data repositories and affect the quality assurance of data. It is therefore very important that the development of audit and certification be further developed and advanced.

c) Journals

When reviewing a classical article, reviewing of underlying data plays a supporting role at best. A lack of reference standards impedes the reviewing of data in some disciplines. Journals are increasingly including statements on accessibility on their editorial policies and thus support the opportunities of their reviewing. By means of developing data journals, publishers and editors make an innovative contribution to quality assured publication of research data.

- To organize reviewing of data effectively, standards and criteria of quality assurance have to be developed. They may need to be discipline- and community- specific. Journals can make an important contribution here by formulating requirements of the quality of data in the editorial policies.
- Cooperation between journals and data repositories is becoming of central importance. To support such cooperation, dialog between publishers and repositories needs to be intensified.[95]
- Data publications provide a variety of opportunities of supporting the sharing of research data in a quality assured form. On account of the early stage of development of this type of publication, best practice and dialog among the participants needs to be supported and further developed.

[94] Projects such as the Registry of Research Data Repositories (Re^{3}Data) currently being developed can make an important contribution: http://www.re3data.org

[95] Initiatives such as DataCite make an important contribution: http://datacite.org

REFERENCES

Alliance of German Science Organisations. (2010). *Principles for the Handling of Research Data*. Retrieved from http://www.allianzinitiative.de/en/core_activities/research_data/principles/

American Geophysical Union. (1996). *Policy on Referencing Data in and Archiving Data for AGU Publications*. Retrieved from http://www.agu.org/pubs/authors/policies/data_policy.shtml

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. (2003). Retrieved from http://oa.mpg.de/files/2010/04/berlin_declaration.pdf

Borgman, C. L. (2007). *Scholarship in the Digital Age. Information, Infrastructure, and the Internet*. Cambridge, Massachusetts: MIT Press.

Brammer, G. R., Crosby, R. W., Matthews, S. J., & Williams, T. L. (2011). Paper Mâché: Creating Dynamic Reproducible Science. *Procedia Computer Science, 4*, 658-667. doi:10.1016/j.procs.2011.04.069

Breure, L., Voorbij, H., & Hoogerwerf, M. (2011). Rich Internet Publications: "Show What You Tell." *Journal of Digital Information, 12*(1). Retrieved from http://journals.tdl.org/jodi/article/view/1606/1738

Cell. (n.d.). *Information for Authors. Editorial Policies*. Retrieved December 28, 2011, from http://www.cell.com/authors#ed_policies

Chavan, V., & Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics, 12*(Suppl 15), S2. doi:10.1186/1471-2105-12-S15-S2

Chen, C. (2011). ISMPP 2011 – what's next for F1000? *Naturally Selected*. Retrieved December 28, 2011, from http://blog.f1000.com/2011/11/24/ismpp-2011-%e2%80%93-what%e2%80%99s-next-for-f1000/

Committee on Publication Ethics. (2008). *A Short Guide to Ethical Editing for New Editors.* Retrieved from http://www.publicationethics.org/files/short_guide_to_ethical_editing_for_new_editors.pdf

Dallmeier-Tiessen, S. (2011). *Strategien bei der Veröffentlichung von Forschungsdaten*. Berlin. Retrieved from http://www.ratswd.de/download/RatSWD_WP_2011/RatSWD_WP_173.pdf

Data Archiving and Networked Services. (2011). *Data Reviews. Peer-reviewed research data*. Retrieved from http://www.dans.knaw.nl/en/content/categorieen/publicaties/dans-studies-digital-archiving-5

Digital Curation Centre. (n.d.). *Data Management Plans*. Retrieved December 28, 2011, from http://www.dcc.ac.uk/resources/data-management-plans

e-Infrastructure Reflection Group, & European Strategy Forum on Research Infrastructures. (2009). *e-IRG Report on Data Management*. Retrieved from http://www.e-irg.eu/images/stories/e-irg_dmtf_report_final.pdf

EUROHORCs & ESF. (2009). EUROHORCs and ESF Vision on a Globally Competitive ERA and their Road Map for Actions. Retrieved from http://www.era.gv.at/attach/EUROHORCs-ESF_Vision_and_RoadMap.pdf

Earth System Science Data. (n.d.). *About this Journal*. Retrieved December 28, 2011, from http://www.earth-system-science-data.net/general_information/about_this_journal.html

Earth System Science Data. (n.d.). *Repository Criteria*. Retrieved December 28, 2011, from http://www.earth-system-science-data.net/general_information/repository_criteria.html

Earth System Science Data. (n.d.). *Review Guidelines – Step by Step*. Retrieved December 28, 2011, from http://www.earth-system-science-data.net/review/ms_evaluation_criteria.html

Ecological Society of America. (n.d.). *Data papers, supplements, and digital appendices for ESA journals*. Retrieved December 28, 2011, from http://www.esapubs.org/archive/default.htm

GRDI2020. (2011). *Global Research Data Infrastructures: The GRDI2020 Vision*. Retrieved from http://www.grdi2020.eu/Repository/FileScaricati/6bdc07fb-b21d-4b90-81d4-d909fdb96b87.pdf

Hey, A. J. G., & Trefethen, A. E. (2003). The Data Deluge: An e-Science Perspective. In F. Berman, G. Fox, & A. J. G. Hey (Eds.), *Grid Computing - Making the Global Infrastructure a Reality* (pp. 809-824). Chichester: Wiley and Sons. Retrieved from http://eprints.ecs.soton.ac.uk/7648/

House of Commons. Science and Technology Committee. (2011). *Peer review in scientific publications. Report, together with formal minutes, oral and written evidence*. London. Retrieved from http://www.publications.parliament.uk/pa/cm201012/cmselect/cmsctech/856/856.pdf

International Association of STM Publishers. (2007). *Brussels Declaration*. Retrieved from http://www.stm-assoc.org/brussels-declaration/

Klump, J. (2011). Criteria for the Trustworthiness of Data Centres. *D-Lib Magazine*, *17*(1/2). doi:10.1045/january2011-klump

Kronick, D. A. (1990). Peer Review in 18th-Century Scientific Journalism. *JAMA: The Journal of the American Medical Association*, *263*(10), 1321-1322. doi:10.1001/jama.1990.03440100021002

Lawrence, B., Jones, C., Matthews, B., Pepler, S., & Callaghan, S. (2011). Citation and Peer Review of Data: Moving Towards Formal Data Publication. *International Journal of Digital Curation*, *6*(2). doi:10.2218/ijdc.v6i2.205

Ligue des Bibliothèques Européennes de Recherche. (2009). *Making the case for European research libraries. LIBER Strategic Plan 2009-2012*. Retrieved from http://www.libereurope.eu/sites/default/files/d5/LIBER-Strategy-FINAL.pdf

Mark Ware Consulting. (2008). *Peer review in scholarly journals: Perspective of the scholarly community - an international study*. Retrieved from http://www.publishingresearch.net/documents/PeerReviewFullPRCReport-final.pdf

Morris, C. (Ed.). (1992). *Quality*. *Academic Press Dictionary of Science and Technology*. London: Academic Press.

Müller, U. T. (2008). *Peer-Review-Verfahren zur Qualitätssicherung von Open-Access-Zeitschriften – Systematische Klassifikation und empirische Untersuchung*. Berlin. Retrieved from http://nbn-resolving.de/urn:nbn:de:kobv:11-10096430

National Institutes of Health. (2003). *Final NIH Statement on Sharing Research Data*. Retrieved from http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

National Science Foundation. (2011). *Proposal and Award Policies and Procedures Guide. Chapter VI - Other Post Award Requirements and Considerations. Dissemination and Sharing of Research Results*. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4

Nature. (2008). *Nature Special: Big Data*. Retrieved from http://www.nature.com/news/specials/bigdata/

Nature. (2009). *Nature Special: Data Sharing*. Retrieved from http://www.nature.com/news/specials/datasharing/

Nature. (2011). *Guide to Publication Policies of the Nature Journals*. Retrieved from http://www.nature.com/authors/gta.pdf

Nature Cell Biology. (n.d.). *Editorial Policies*. Retrieved December 28, 2011, from http://www.nature.com/ncb/about/ed_policies/index.html

Nowakowski, P., Ciepiela, E., Harężlak, D., Kocot, J., Kasztelnik, M., Bartyński, T., Meizner, J., et al. (2011). The Collage Authoring Environment. *Procedia Computer Science*, *4*, 608-617. doi:10.1016/j.procs.2011.04.064

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

PLoS ONE. (n.d.). *PLoS ONE Editorial and Publishing Policies. Sharing of Materials, Methods, and Data*. Retrieved December 28, 2011, from http://www.plosone.org/static/policies.action#sharing

Pampel, H., & Bertelmann, R. (2011). „Data Policies“ im Spannungsfeld zwischen Empfehlung und Verpflichtung. In S. Büttner, H.-C. Hobohm, & L. Müller (Eds.), *Handbuch Forschungsdatenmanagement* (pp. 49-61). Bad Honnef: Bock + Herchen. Retrieved from http://opus.kobv.de/fhpotsdam/volltexte/2011/228/

Pfeiffenberger, H., & Carlson, D. (2011). “Earth System Science Data” (ESSD) — A Peer Reviewed Journal for Publication of Data. *D-Lib Magazine*, *17*(1/2). doi:10.1045/january2011-pfeiffenberger

Reilly, S., Schallier, W., Schrimpf, S., Smit, E., & Wilkinson, M. (2011). *Report on Integration of Data and Publications*. Retrieved from http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-ReportOnIntegrationOfDataAndPublications-1_1.pdf

Research Councils UK. (2011). RCUK Common Principles on Data Policy. Retrieved from http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx

Research Information Network. (2008). *To Share or not to Share: Publication and Quality Assurance of Research Data Outputs. Annex: detailed findings for the eight research areas*. Retrieved from http://www.rin.ac.uk/system/files/attachments/To-share-data-outputs-annex.pdf

Research Information Network. (2008). *To Share or not to Share: Publication and Quality Assurance of Research Data Outputs. Main report*. Retrieved from http://www.rin.ac.uk/system/files/attachments/To-share-data-outputs-report.pdf

Research Information Network. (2011). *Data centres: their use, value and impact*. Retrieved from http://www.jisc.ac.uk/news/stories/2011/09/~/media/Data Centres-Updated.ashx

Savage, C. J., & Vickers, A. J. (2009). Empirical Study of Data Sharing by Authors Publishing in PLoS Journals. PLoS ONE, 4(9), e7078. doi:10.1371/journal.pone.0007078

Science. (2011). *Special Online Collection: Dealing with Data*. Retrieved from http://www.sciencemag.org/site/special/data/

Science. (2011). *Science Special: Data Replication and Reproducibility*. Retrieved from http://www.sciencemag.org/site/special/data-rep/

Science and Technology Facilities Council. (2011). *STFC scientific data policy*. Retrieved from http://www.stfc.ac.uk/Resources/pdf/STFC_Scientific_Data_Policy.pdf

Sense about Science. (2009). *Peer Review Survey 2009: Full Report*. Retrieved from http://www.senseaboutscience.org/data/files/Peer_Review/Peer_Review_Survey_Final_3.pdf

Van Gorp, P., & Mazanek, S. (2011). SHARE: a web portal for creating and sharing executable research papers. *Procedia Computer Science, 4*, 589-597. doi:10.1016/j.procs.2011.04.062

Ware, M. (2011). Peer Review: Recent Experience and Future Directions. *New Review of Information Networking, 16*(1), 23-53. doi:10.1080/13614576.2011.566812

Wellcome Trust. (2010). *Policy on data management and sharing*. Retrieved from http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm

Wikipedia. (2011). Standing on the shoulders of giants. Retrieved from http://en.wikipedia.org/w/index.php?title=Standing_on_the_shoulders_of_giants&oldid=458721848

Woutersen-Windhouwer, S., Brandsma, R., Verhaar, P., Hogenaar, A., Hoogerwerf, M., Doorenbosch, P., Dürr, E., et al. (2009). *Enhanced Publications. Linking Publications and Research Data in Digital Repositories*. (M. Vernooy-Gerritsen, Ed.). Amsterdam: Amsterdam University Press. Retrieved from http://dare.uva.nl/aup/nl/record/316849

FIGURES