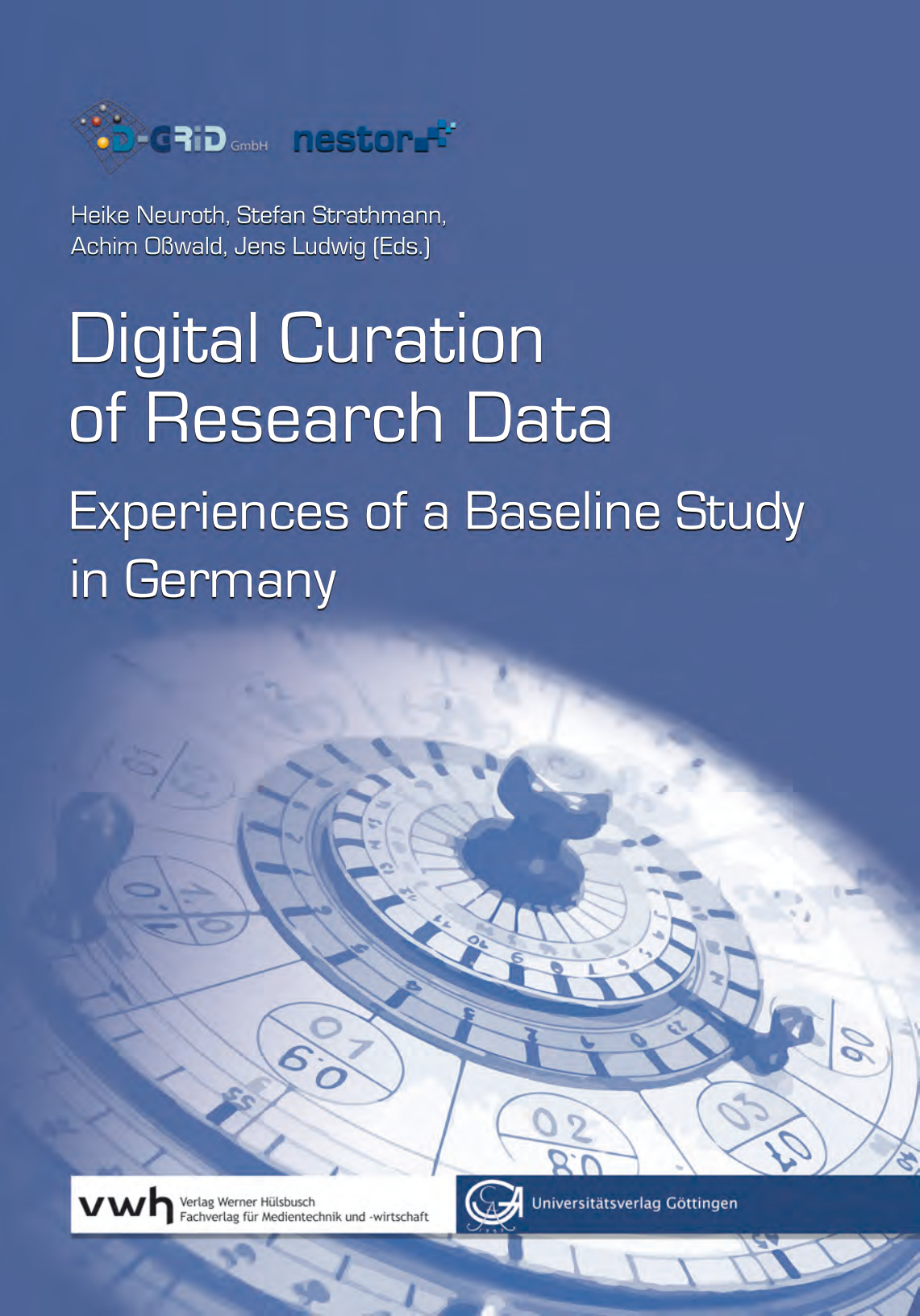
Heike Neuroth, Stefan Strathmann,
Achim Oßwald, Jens Ludwig (Eds.)

Digital Curation of Research Data

Experiences of a Baseline Study in Germany



vwh Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

Universitätsverlag Göttingen

Neuroth, Strathmann, Oßwald, Ludwig (Eds.)
Digital Curation of Research Data

Kontakt / Contact: editors@langzeitarchivierung.de

c/o Niedersächsische Staats- und Universitätsbibliothek Göttingen, Dr. Heike Neuroth, Forschung und Entwicklung, Papendiek 14, 37073 Göttingen

c/o Göttingen State and University Library, Dr. Heike Neuroth, Research and Development, Papendiek 14, 37073 Göttingen, Germany

**Heike Neuroth, Stefan Strathmann,
Achim Oßwald, Jens Ludwig (Eds.)**

Digital Curation of Research Data

Experiences of a Baseline Study in Germany

vwh
Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

Digital Curation of Research Data

Herausgegeben von Heike Neuroth, Stefan Strathmann, Achim Oßwald und Jens Ludwig · im Rahmen des Kooperationsverbundes nestor – Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland · http://www.langzeitarchivierung.de/

Edited by Heike Neuroth, Stefan Strathmann, Achim Oßwald and Jens Ludwig · within the context of nestor – Network of Expertise in the Long-Term Storage of Digital Resources for Germany · http://www.langzeitarchivierung.de/

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter http://www.d-nb.de abrufbar.

Bibliographic information of the German National Library
The German National Library lists this publication in the German National Bibliography; detailed bibliographic data is available online at http://www.d-nb.de.

Die Inhalte dieses Buches stehen auch als Onlineversion über die Website von nestor zur Verfügung / This work is available as an Open Access version at the nestor website: http://nestor.sub.uni-goettingen.de/bestandsaufnahme/index.php?lang=en

vwh Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

in Kooperation mit dem Universitätsverlag Göttingen
in cooperation with the Universitätsverlag Göttingen

Printed in Poland · ISBN: 978-3-86488-054-4

Content

Foreword

The relevance of research data today and for the future is well documented and discussed, in Germany as well as internationally. In addition, more and more policy makers are aware of the meaning of research data and the possibilities to access, share, and re-use them. Recently the government of the United Kingdom decided that publicly funded research publications and research data must be freely available and accessible to the public in all situations in which there are no copyright issues or other legal aspects that would prevent it. It is expected that the European Commission will publish similar requirements in the context of the next framework program Horizon 2020 starting at the end of 2013. The newly funded Research Data Alliance (RDA) is one example of the increasing interest in these topics worldwide. Ensuring that research data are accessible, sharable, and re-usable over time makes several further steps possible:

- Research data are documented and could therefore be validated.
- Research data could be the basis for other and new research questions, since they could be an integral part of the (digital) research lifecycle from the very beginning.
- Research data could be re-analysed by using new, innovative digital methods which were unknown at the moment of data acquisition.
- Research data could be used by other disciplines, therefore encouraging interdisciplinary research.

For all of these reasons, it is essential that research data are curated, which means that they are kept accessible and interpretable over time. A standardized questionnaire was developed in order to understand whether the approaches and methods of dealing with research data within the academic disciplines are different or whether there are similarities in terms of solutions as well as challenges and problems. This questionnaire was distributed to representatives from those disciplines in Germany that were identified as familiar with or already expert in research data curation.

The results of this survey have been published in German in 2012 in the handbook *"Langzeitarchivierung von Forschungsdaten – Eine Be-*

standsaufnahme".[1] This publication is the English-language translation of the main chapters of this handbook. The original German version also contains detailed analyses of the situation regarding the curation of research data of eleven disciplines ranging from humanities and social sciences to the natural sciences and medicine. Colleagues from these eleven disciplines were asked to describe the state-of-the-art regarding their methods of handling and experiences with research data curation in the questionnaire. These chapters have not been included in this English publication.

The last chapters of this English-language publication analyse the responses from all disciplines, compare the similarities as well as differences, and conclude with some overall implications and recommendations for stakeholders, policy-makers, key-players, and scholarly societies.

The editors of this volume, as well as the additional editors of the German version, have been working together closely for many years, such as in the context of nestor – the German competence network for digital preservation. When we started the data curation discussion in Germany at universities and in research disciplines, the terms "long-term preservation" and "digital preservation" were already established in Germany. Today we prefer the term "digital curation". As a result we used these earlier terms in cases where we are following the original German handbook or for citations. In all other cases the more modern term "data curation" is used.

We would like to express our special thanks to Hanna-Lena Stolz and Dr. Kathleen M. Smith for their valuable support in translating the main chapters of the German handbook. Without their help, we would never have been able to share our experiences, thoughts, and conclusions on this important and urgent topic. We are eager for exchange with the broader data community, across geographic and linguistic borders, across academic disciplines, across funding agencies, and many other levels.

With best regards,
Heike Neuroth, Stefan Strathmann, Achim Oßwald and Jens Ludwig

1 Neuroth et al. (2012).

1 Digital Curation of Research Data: An Introduction

Achim Oßwald, Heike Neuroth, Regine Scheffel

Particularly since it was reported in the media that NASA would only be able to recover the data from the first manned flight to the moon with a significant investment of resources, it has been clear that major efforts are necessary to preserve digital research data for the future.[2] Other large-scale breakdowns in the preservation of data confirm that this need applies to additional fields of study.[3] In addition, there have been repeated incidents of deliberate research data manipulation by researchers.[4]

The scholarly community requires reliable long-term access to research data for several reasons. For example, the scandal involving the cell biologist Tae Kook Kim has made clear the importance of keeping research data available and verifiable, especially data upon which current scholarly publications are based.[5] Digital research data – today the essential foundation of scholarship – are often irreproducible. If they are lost, they are gone forever and therefore no longer verifiable. Measurement data in the field of climate research from the last few decades serves as a clear example. In such cases, the curation and long-term availability ensures the verifiability, interpretability, and reusability of the research data that has been collected. The forms of subsequent use are determined by these expanded possibilities for access. The integration of digital data in new disciplinary contexts provides new opportunities in a way that old research questions can be answered in new ways and entirely new research questions can be generated. By including this data long-term studies in climate science or in the social sciences become possible at all. E.g. in astronomy, (analogous)

2 Schmundt (2000); Hammerschmitt (2002).

3 See *Spiegel Online* (2007).

4 See Heinen (2010).

5 See Kennedy; Alberts (2008).

photography has been used since the end of the nineteenth century to permanently preserve astronomical data.[6] One of the most comprehensive data collections is the archive of the Harvard College Observatory with over 500,000 photographic plates taken within more than 100 years, ending in 1989.[7] Another example is the Sonneberg Observatory archive, which includes approximately 300,000 photographic plates taken over seventy years, by which more than 10,000 variable stars have been discovered.[8] These huge data archives are gradually being digitized to preserve them for posterity and to make it possible to analyse them with computerized techniques. They are an indispensable resource, particularly for studying the changes in brilliancy and in the position of stars over dozens of years.

The interdisciplinary use of data is made possible by free access to and the citability of research data. A new form of re-use developed in the USA is the trend of crowdsourcing, in which the general public, or a clearly-defined subsection of the disciplinary population (such as graduate students), participates in the creation or qualitative enrichment of research data.[9] The *Galaxy Zoo* project is an example of *citizen science*[10] or *crowdsourcing*, in which interested laymen are involved in the research process.[11] Modern sky mapping creates countless images of galaxies. These galaxy shapes show a great variety and complexity. There is still no good computerized classification method available for this kind of data. For this reason, American astrophysicists decided to involve members of the general public in this process in July 2007. They invited amateur astronomers to participate in the classification of these galaxies and offered special training sequences so that new participants could learn the classification

6 We are grateful to Prof. Wambsgans at the Astronomisches Rechen-Institut (ARI) of the Zentrum fuer Astronomie at the University of Heidelberg (ZAH; http://www.zah.uni-heidelberg.de/zah/) for this information.

7 See Harvard College Observatory, http://www.cfa.harvard.edu/hco/.

8 See Sternwarte Sonneberg, http://www.stw.tu-ilmenau.de/.

9 See Website “Crowdsourcing” (2013).

10 See Website “Citizen Science” (2013).

11 See Galaxy Zoo, http://www.galaxyzoo.org

criteria. One structurally similar example in the humanities is the *Collaborative Manuscript Transcription* project.[12]

Digital curation, after all, is about making research data digitally available for the long term – sometimes even as independent publications in their own right.[13] The intention is to make them verifiable, interpretable, and re-usable, and to cross-link research data using research infrastructures, especially in order to increase the potential for interdisciplinary reuse. At the same time, more emphasis has been placed on a new vision of research environments which was provided in October 2010 as the Vision 2030 for research data by the *High Level Expert Group on Scientific Data*, a European Commission panel of experts:

> Our vision is a scientific e-infrastructure that supports seamless access, use, re-use, and trust of data. In a sense, the physical and technical infrastructure becomes invisible and the data themselves become the infrastructure – a valuable asset, on which science, technology, the economy and society can advance.[14]

The realization of this vision is still associated with a number of open questions and challenges, starting with the term *research data* itself. What are research data? For example, this term could refer to data from instruments such as a telescope or raw data from a mass spectrometer, and to digital maps or full-text documents such as those used in the creation of critical editions. The term research data must always be viewed in relation to a particular subject discipline. Similarly, all requirements for the management and long-term availability of research data must be differentiated from each other in regard to both general and discipline-specific aspects and solutions.

Thus far, there is no general agreement on the definition of digital curation, not only in Germany, but on international levels as well. E.g. nestor, the German competence network for digital preservation, which has been dealing intensively with this subject for years, offers no definition on its homepage.[15] The following explanation is found in the intro-

12 See Brumfield (2011).

13 See, for example, PANGAEA, http://www.pangaea.de.

14 See High Level Expert Group on Scientific Data (2010).

15 See nestor, http://www.langzeitarchivierung.de.

duction to the nestor reference work *nestor Handbook: A Small Encyclopaedia of Digital Preservation / nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*):[16]

> Preservation in this context means more than simply compliance with legal requirements concerning the duration of time in which data tables that are relevant for tax purposes must be kept available. "Long-term" refers to an undefined period of time in which important and unpredictable technological and socio-cultural changes occur: changes which could completely revolutionize the form and the use scenarios of digital resources. It is important, therefore, to develop strategies for specific digital collections that protect the long-term availability and reuse of digital objects, depending on individual needs and future use scenarios. "Long term" does not mean a guarantee for the preservation of digital resources over five or over fifty years, but rather the responsible development of strategies that could deal with the constant changes caused by the information market.[17;18]

By digital preservation, we mean the period of time as defined on an individual basis according to the context of the preservation of digital objects, beyond basic technological and socio-cultural processes of change. Long-term preservation makes it possible to secure access to and re-use of research data for the future.

The subsequent challenges are clear: Since we cannot preserve all research data, what are the selection criteria for the data to be preserved, and who defines them? Who can safely estimate at the present time what kinds of research data will be of interest to future researchers? How do we deal with research data that cannot be reproduced (for example, climate data and the astronomical observations mentioned earlier)? It is clear that *bitstream preservation*,[19] which means preserving only the bits and bytes of

16 Please see the printed edition 2.0 of the *nestor Handbuch* (Neuroth et al. 2009) as well as the updated online edition 2.3 from 2010 (Neuroth et al. 2010).

17 See the German version of this definition at Liegmann; Neuroth (2010), p. 1:2.

18 In this context, the question arises as to whether Schwens and Liegmann's original explanation of long-term archiving and long-term availability as published in 2004 can be adopted by the academic community. See Schwens; Liegmann (2004), p. 567.

19 See Ullrich (2010).

the *physical object*,[20] can only be a first step at best. The requirements for long-term availability, meaning the future interpretability and usability of scholarly data, are much more difficult because the nature of future technological interfaces cannot be predicted. Therefore, digital objects that are placed in a long-term archive must be described by metadata.[21] The technical and organizational context in which the data were created must also be maintained and documented in a standardized form. Only this offers the chance of using these data (possibly based on emulation[22] or migration[23]) in the future.[24] In the near future, however, descriptive, technical, and administrative metadata will be required, as demonstrated by the factsheet *Keeping Research Data Safe* (KRDS),[25] a combination of two studies about the costs of digital curation of research data.[26] As the follow-up report noted, the research results from studies completed even a few years ago could not be re-used by participating researchers because the methods used to collect the data were not documented in sufficient detail.[27] This is particularly the case where research data should be preserved for re-use in ways that cannot be anticipated at the present time, e.g. those data that reflect fundamental socio-cultural changes. For example, today the gender aspects of old church registers are a topic of analysis, an aspect which surely was not anticipated in the past. In order to maintain today's administrative files and databases, which include comparable data, usable for

20 For digital objects, Thibodeau differentiates between the level of the conceptual object, which is deemed worthy of preservation; the logical object of the realization in the form of data that are bound to a particular hard- and software environment; and the physical object of the pure bitstream; see Thibodeau (2002).

21 We assume a long-term archive based on the OAIS model. See the discussion about the updated version of the standard at "Reference Model for an Open Archival Information System" (OAIS) (2009) and the discussion based on it. For an overview of OAIS see OAIS (2010).

22 See Funk (2010a).

23 See Funk (2010b).

24 The legal conditions under which this would be feasible are still unclear.

25 See Charles Beagrie Ltd & JISC (2010).

26 See Beagrie; Chruszcz; Lavoie (2008); Beagrie; Lavoie; Woollard (2010).

27 See Beagrie; Lavoie; Woollard (2010), p. 2.

future research questions, appropriate metadata must be created, archived, and kept available. In this context, diverse future use scenarios and potential user groups (the designated communities) and their expectations for the description of the surviving data should be considered in preservation concepts and considerations.

Consequently, descriptive metadata are particularly important. This is especially true for metadata providing systematically differentiated details, which shed light on the criteria used in selecting the object of investigation, the methods of examination, measurement and surveying, their application as well as the results of the examination. The overview of the current situation provided in this survey investigates general and discipline-specific standards relevant to the curation of research data and the establishment of research infrastructures throughout Germany.

In general, it is clear that this type of *digital curation*[28] of research data already offers advantages for current research activities regarding digital preservation and long-term availability. Accessibility to published research data ensures the quality of academic activities and facilitates academic publishing.[29] It also has the secondary effect of increasing research standards and productivity. This can be seen, for example, in a very pragmatic aspect such as maintaining the continuity of research work over several generations of researchers. Another advantage of the systematic documentation and maintenance of research data during their production is the long-term savings in costs. The retrospective correction of erroneous metadata can be more expensive by a factor of 30 than the original creation of the data itself.[30]

Research organizations in Germany have long been responding to this situation with guidelines for data preservation. The German Research Foundation (Deutsche Forschungsgemeinschaft [DFG]), one of the major funding agencies for academic research in Germany, requires projects to

28 The term "digitales Kuratieren," a translation of the English term digital curation, is beginning to establish itself in German-speaking areas to refer to the systematic planning, creation, evaluation and transformation and reuse of digital research data and – in a further sense – all digital objects (see Digital Curation Centre [2011b]).

29 Charles Beagrie Ltd & JISC (2010), p. 2.

30 Ibid.

ensure that the data on which their findings are based must be kept available for at least ten years.[31] The Alliance of German Research Organizations (Allianz der deutschen Wissenschaftsorganisationen)[32] is working to improve the creation and re-use of research data by developing standards, archive structures, and incentive systems.[33] Even the German Council of Science and Humanities (Wissenschaftsrat [WR]) has taken a clear position in this regard in its "Comprehensive Recommendations for Information Infrastructures" ("Übergreifende Empfehlungen zu Informationsinfrastrukturen")[34] in January 2011, which called for the sustained funding of corresponding research infrastructures and long-term archival concepts. The identification of research data with persistent identifiers (such as URN[35], DOI[36], and EPIC[37]) is a significant step towards the permanent citability of these data and data collections. However, the DFG's ten-year perspective is only a contribution to data curation; the subsequent re-use of research data presupposes long-term preservation and long-term availability.

Cooperation plays a central role in the success of curation of research data. Cooperative efforts are found on various levels: on a local or institutional level. Advantages can be experienced by researchers immediately because they have an unmediated influence on the process. On a regional, and certainly on a national level, institutional and/or legal measures can be put in place. On the European and international level, structures and processes (ideally standardized) can be established to accommodate the increasingly global research activities which are taking place. Discipline-specific data centers, which already ensure efficient data management,

31 See DFG (1998), p. 12.

32 See Alliance of German Science Organisations http://www.allianzinitiative.de/en/start/

33 See Alliance of German Science Organisations (2010) or http://www.allianzinitiative.de/en/core_activities/research_data/.

34 See Wissenschaftsrat (2011b).

35 See Schöning-Walter (2010).

36 See Brase (2010).

37 See EPIC, http://www.pidconsortium.eu.

could become points of intersection in a long-term archival network.[38] Together, they could form a long-term archival infrastructure based on maintaining the long-term availability of scholarly research data.

Although there have been extensive preparations and concepts for the sustainable management of research data in the recent past, their implementation is still in its infancy. One important factor appears to be that the solutions that have previously been tested cannot be integrated well enough in research activities and workflows. A SURF Foundation study examined the results of 15 projects studying the use of research data.[39] In particular, the study focused on researchers' requirements for research data infrastructures and which requirements were essential in order for researchers to use these infrastructures for research data. In the summary of the cases examined in this study, there were two different roles: the researcher as a producer of data and the researcher as a consumer of data. It turned out that the needs of these two roles were almost diametrically opposed. While the data consumer expected a central point of access with a variety of possible combinations of data and tools, the data producer required a locally managed, customized work environment. In addition, formal regulations, data management plans, and their verification were perceived as obstacles. Bridging the contradictions between these roles remains a significant challenge. A major concern must therefore be to examine the causes of this ambivalence more precisely and find out how to overcome them. Possibilities include providing an infrastructure which can be used intuitively, or establishing an incentive or sanction system, and, in doing so, promoting the development of a new publication culture for research data. The government, the academic community, and infrastructure institutions should address these challenges cooperatively. It is important to consider the subject-specific characteristics and requirements and to keep in mind that this process can only begin with the individual

38 In particular, the Helmholtz Foundation (HGF) is operating several subject-specific data centers, such as the Deutsches Fernerkundungszentrum at the German Aerospace Center (Deutsches Zentrum für Luft und Raumfahrt; see http://www.dlr.de/dlr/en/desktopdefault.aspx/tabid-10002/) and the World Data Center for Remote Sensing of the Atmosphere (WDC-RSAT). Homepage: http://wdc.dlr.de.

39 See Feijen (2011).

disciplines. A top-down approach or a standard solution for all disciplines will not be accepted and therefore will have little chance to be successful.

In recent years, the public debate in Germany about the curation of digital data (such as in relation to nestor) has focused on a more traditional interpretation of the field of cultural heritage. It is time for governmental policy makers and the general public to recognize research data as a national, scholarly cultural asset, and to provide support for the curation of research data by providing infrastructural measures.

2 Status of Discussion and Current Activities: National Developments

Stefan Winkler-Nees

The discussion in Germany about the handling und reuse of scholarly data from publicly funded projects is shaped by researchers who are relatively autonomous and independent. The basic principle of academic freedom, as laid out in article 5, paragraph 3 of the German constitution ("… Art and scholarship, research, and teaching shall be free. …" e.g. this freedom may include the right, not to publish), also determines attitudes about ways in which to deal with individual or collaboratively acquired scholarly findings, and, correspondingly, the willingness of researchers to make their data available. Claims, both existing and perceived, to knowledge produced by him or her lead to uncertainty and concerns regarding the unmanageable nature of subsequent re-use and the potential misuse of data once it has been provided. A constructive discussion about the potential and possibilities of the standardized preparation of research data is hindered by these uncertainties.

At the same time, it is generally acknowledged that a sustainable approach to project results was not adequately considered in the past, especially in publicly funded research in the light of the "virtualization of research."[40] In addition to the accelerating pace of transformation with simultaneous fundamental changes in academic research processes, and the potential value offered by professional information management, require organizational modifications to the current structural framework.

In the recent past, and in some cases already in the course of the last decade, several activities were developed independently of each other. These activities led to the development of infrastructures that were highly regarded on disciplinary levels. These "grassroots projects" were, however, unaccompanied by any universal interdisciplinary discussion, nor were they embedded in an overarching, consensually-agreed upon concept

40 See Horlings et al. (2006).

for data management. The heterogeneous, diverse research landscape in Germany, with its various research performing organizations, universities, federal structure and funding streams[41] and frequently ambiguously undefined responsibilities, has meant that the early impulses that were implemented were not coordinated with each other, and structural measures that planned for the future were not discussed comprehensively. The German Research Foundation (Deutsche Forschungsgemeinschaft [DFG]) report "Safeguarding Good Scientific Practice" was the first to contain a general, cross-disciplinary requirement to preserve data beyond a fixed period of time.[42] Correspondingly, there was originally no requirement to provide research data for scholarly purposes and for reuse. Research data first came to the forefront beginning with the intensifying debate about Open Access publishing and the corresponding change in the awareness of the necessity of access to digital information. In addition to academic researchers, stakeholders from the research organizations, from science policy, from the information infrastructure organizations, and from funding agencies began to take on more responsibility for developing, coordinating, and implementing appropriate measures.

2.1 Research Organizations

The strongest motivation to secure, archive, and make research data available according to professional criteria lies in recognizing and making use of the potential benefits that can be thus realized. Academic disciplines that work closely with digital data and, while doing so, engage in international collaboration, have in many cases built systems and infrastructures that suit their specific requirements. Due to the high acceptance rate on the part of the scholarly community, these systems represent successful flagship projects for the meaningful and effective use of research data.

41 See http://www.research-in-germany.de/main/2866/research-landscape.html for further information.

42 See DFG (1998).

One example of this engagement is the situation in the field of marine and environmental sciences, which led to the creation of the PANGAEA information system more than two decades ago. The network of the ICSU[43] World Data Center has established an internationally recognized information infrastructure through cooperation between one of the leading German marine research facilities, the Alfred Wegener Institute for Polar and Marine Research in Bremerhaven, and the Marum Center for Marine Environmental Sciences at the University of Bremen. In addition to the sustained commitment of those involved, the increasing integration of PANGAEA by scientists in their research activities as a source of information and as a data repository contributed to its success. Furthermore, funding organizations were successfully convinced of the significance and use of this system, so that a long-term and sustainable operation became feasible. The involvement of the Helmholtz Association of German Research Centres (Helmholtz-Gemeinschaft Deutscher Forschungszentren e.V.) in the discussion about access to research data also contributed to this development, particularly in the context of their open access activities. Thus, many research centers in the association are committed to the principle of free access to research data and an additional ICSU World Data Center (WDC) could be established (the WDC for Remote Sensing of the Atmosphere).[44]

Until now, the research institutions within the Leibniz Association (Leibniz-Gemeinschaft) have been reviewing individual approaches to deal with research data that demonstrate success in several initiatives at the disciplinary level. For example, the Leibnitz Institute for Psychology Information and Documentation (Leibniz-Zentrum für Psychologische Information und Dokumentation [ZPID])[45] has built a unique reference database for psychological literature, testing procedures, and various other materials. As a service provider for the Leibniz Association, the German National Library of Science and Technology (Technische Informations-

43 See ICSU, http://www.icsu.org.

44 See World Data Center for Remote Sensing of the Atmosphere, http://wdc.dlr.de.

45 See Leibniz Institute for Psychology Information (ZPID) http://www.zpid.de.

bibliothek Hannover [TIB])[46] in Hannover assumed the task of making research data available in a more efficient way for research. Research data, which form the basis of publications, are not hosted by the TIB itself as a general rule, but are registered and made searchable by assigning a DOI in order to create lasting citability. This service is intended to be generally available for all disciplines and institutions, including those outside the Leibniz Association.

The issue of dealing with research data at universities has as yet no particular nationwide resonance in Germany. In contrast to international approaches, German universities do not view themselves as under the obligation or even capable of initiating measures to improve the situation. However, several university libraries have seized the initiative, such as the Göttingen State and University Library in Göttingen (Niedersächsische Staats- und Universitätsbibliothek Göttingen [SUB]). The nestor project,[47] conducted under the aegis of the German National Library and carried out in cooperation with other libraries, museums, and institutions of higher education, has laid the basic groundwork for the ongoing discussion with a series of studies. On a disciplinary level, however, researchers predominantly use existing options available from research organizations or they attempt to make digital content available using solutions from their own institutes. In many cases, this type of approach does not possess the required sustainability and availability on a supra-regional level for research data repositories. Researchers at German research academies are in a comparable situation. The TELOTA system, which was established at the Berlin-Brandenburg Academy of Sciences and Humanities (Berlin-Brandenburgische Akademie der Wissenschaften [BBAW]) in 2002, is one example.[48] Here, the research data in particular, such as editions, dictionaries, bibliographies and documentation, which are produced in large quantities by academy programs, are intended to be preserved and made available in a suitable manner. The wide range of subjects, from ancient inscriptions to medieval glass painting and records about silk road docu-

46 See German National Library of Science and Technology, http://www.tib.uni-hannover.de/en.html.

47 See nestor, http://www.langzeitarchivierung.de.

48 See TELOTA (2011).

ments, from ancient inscriptions to the treatment of electronic "heritage assets," seems to be well-suited for the development of a system with services for the entire range of disciplines in the humanities.

In 2007, the Max Planck Society (Max-Planck-Gesellschaft [MPG]) established a new service unit, the Max Planck Digital Library [MPDL]), which "is intended to assist researchers at the Max Planck Society in organizing the research information process" („… den Forschern der Max-Planck-Gesellschaft helfen soll, den wissenschaftlichen Informationsablauf zu organisieren …").[49] The primary focus of the MPDL is on providing research information, the dissemination of information, eScience services, and providing support for the Max Planck Society in implementing open access. With the help of internally funded projects and the engagement of externally funded projects, researchers, particularly those of the Max Planck Society, should be able to have the best possible access to digital resources.

2.2 Recommendations and Policies

Before 1998, the handling of research data, according to the current understanding, did not play a significant role in shaping research policies in Germany. It was only with the increasing prominence of cases of academic misconduct within the bodies of the DFGDFG, that more attention was paid to the necessity of access to research results. Paramount was not the aspect of scholarly reuse and future use of data, but instead the possibility of verifying the accuracy of data after publication. The implementation of this commitment to accessibility was intended to be carried out by all recipients of funding from the DFG who would pledge themselves to "ensure good scholarly practices." In a memoranda published by the DFG in 1998, recommendation seven stated that: "primary data, as a basis of publications, are to be saved on durable and safe media at the same institu-

49 Max Planck Digital Library, http://www.mpdl.mpg.de.

tion in which they arose for at least ten years".[50] The motivation for this recommendation was to make results in publications more reproducible with the help of the data, and to provide evidence of scholarly misconduct, retrospectively as well. The form, formats, and responsibilities were not further specified, so this practice does not satisfy the requirements of modern data management.

The "Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities" ("Berliner Erklärung über offenen Zugang zu wissenschaftlichem Wissen")[51], which was signed by many German and international research organizations on October 22, 2003, can be considered an integral milestone on the path to improved accessibility to scholarly knowledge and research results. Although the focus lay on open access to research literature, the Berlin Declaration expanded the definition of open access to research information "as a comprehensive source of human knowledge and cultural heritage" as a response to the diverse possibilities of information access via the internet.

The significant changes in research and the provision of research information led to intensive discussions about strategic modifications to funding options for research information infrastructures. Accordingly, the DFG Committee on Scientific Libraries and Information Systems (Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme [AWBI]) prepared a position paper in 2006 that recommended modified priorities for funding measures.[52] This paper focused on the changing research requirements caused by increasing digital networking. With respect to research data, it is essential to develop new "structures for storing, referencing and making data available"[53] („Strukturen zur Speicherung, Referenzierung und Verfügbarmachung"). Established information institutions like libraries, archives, and museums are considered major actors

50 DFG (1998): „Primärdaten als Grundlagen für Veröffentlichungen sollen auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, für zehn Jahre aufbewahrt werden."

51 See Website "Berliner Erklärung über offenen Zugang zu wissenschaftlichem Wissen" (2006).

52 DFG (2006).

53 Ibid.

whose technical expertise in the field of information management and preservation should be connected to the new requirements for providing access to information. At the same time, it is necessary to factor in the varying requirements in different academic disciplines in this discussion and to create opportunities for referencing and making research data available. It is of primary importance to achieve increased willingness on the part of researchers in working together on the establishment and use of information infrastructures. This need was also highlighted in a study commissioned by the German Federal Ministry for Education and Research (Bundesministerium für Bildung und Forschung [BMBF]).[54] According to this study, researchers for the most part decide for themselves what they want to use for their work, which is why information infrastructure options must be attractive.

In 2008, the Alliance of German Research Organizations (Allianz der deutschen Wissenschaftsorganisationen), which includes some of the biggest research organizations in Germany, established the priority initiative "Digital Information" with a funding period until 2012.[55] One of the foundations of this initiative was the awareness that there are no systematic approaches and methods and no sustainable infrastructures for the backup, provision, and archiving of research data. Organizational and technical aspects as well as legal and financial aspects are largely unclear. Therefore, the main objective in the area of research data is to develop and implement coordinated measures that take discipline-specific differences into account to ensure the efficient and professional handling of research data in Germany. These activities should focus on three areas: (1) development of a common policy agreed upon by all alliance and partner organizations; (2) encouragement for the development of individual information infrastructures, which should be designed and developed as pilot projects by subject experts and information specialists working in close cooperation; and (3) at a later stage, the definition and characterization of the different use scenarios in the various academic disciplines. These measures should include all groups of stakeholders: researchers as data producers and users, research institutions and universities, infrastructure institutions as well as

54 See Horlings et al. (2006).

55 See Alliance of German Science Organisations (2008).

relevant parties in the state and federal government. The Alliance Initiative working group on "research data" produced a paper on "Grundsätze zum Umgang mit Forschungsdaten" ("Principles for the Handling of Research Data"), which was adopted on July 24, 2010, by the board of directors of all Alliance partner organizations.[56] This paper supported as a basic principle the free and open access to data from publicly-funded research in accordance with legal requirements. The rights of researchers are to be respected at the same time. Equally, the differences between academic disciplines and their respective requirements should be considered. This paper also recommended placing a higher value on making research data available, along with the necessary investment of resources that are connected, and to establish them as an integral part of a scholarly reputation. Furthermore, the recommendations stressed the necessity of integrating the management of research data and their methods and mechanisms into certification programs for infrastructure experts as well as into academic curricula. Research data should be collected, saved, and archived according to standards, both those currently existing and those yet to be developed. Finally, the document recommended developing suitable infrastructure in combination with sustainable research data management. After these recommendations are signed, appropriate measures between partner organizations should be coordinated to ensure that these suggestions are implemented.

Independently of the Alliance Initiative "Digital Information," the DFG subcommittee on information management, part of the committee for scientific libraries and information systems, published the "Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsdaten" ("Recommendations for the secure preservation and provision of digital research data").[57] These recommendations were the result of various workshops and roundtable discussions among experts at the initiative of the DFG. They included a definition of primary data for research,[58] as

56 See Alliance of German Science Organisations (2010).

57 See DFG (2009b).

58 In these recommendations, the term "primary research data" ("Forschungsprimärdaten"), which was used in earlier papers in this form, is used. Since this usage

well as the recommendation to adequately factor in subject-specific aspects. Research data should be preserved and secured under widely accepted standards and identification of the data producer should be provided. In this paper, researchers are requested to allocate research data make access free, unrestricted by local or national borders, if possible. Generally, data are to be provided with metadata under existing standards or those currently under development. Representatives of all academic disciplines are called upon to develop mechanisms and methods for suitable quality control.

In 2009, the Joint Science Conference (Gemeinsame Wissenschaftskonferenz des Bundes und der Länder [GWK]) asked the Leibniz Association to develop a comprehensive concept for subject-specific information infrastructures in Germany.[59] The commission Future of Information Infrastructure (Zukunft der Informationsinfrastruktur [KII]) that was created as a result, prepared recommendations during 2010 in eight thematically-oriented working groups on the most important aspects of future information infrastructure. Most of these working groups were closely connected in content and organization with the complementary working groups of the Alliance Initiative "Digital Information." In contrast to the functional nature of the Alliance working groups, who were not limited by time restraints, the KII commission had a temporary and strategically limited appointment that lasted only until the completion of its assignment. The working group "Forschungsdaten" ("Research Data") stated that there is a strong need for action despite numerous individual activities. This overlaps with the tasks identified by the Alliance working group "Research Data": they listed organizational, technical, legal and in particular financial challenges. They explicitly recommended considering the disciplinary differences. This is also reflected in the recommendations that were developed, which were addressed to the key players in the management of research data. The fact that in those academic disciplines with well-developed international networks, already institutional structures are in place, demonstrate that accredited and trustworthy institutions are in a

combines an unnecessary restriction with a definition of primary data that has not been generally agreed upon, the term "research data" is preferable.

59 See Kommission Zukunft der Informationsinfrastruktur (2011).

position to take a leadership role in the process of developing a sustainable way of providing research data. The working group stated that universities and research institutions are responsible for participating in raising awareness. This type of outreach activity could be achieved by measures such as the establishment of central structures, the implementation of data management plans, and providing options for securing data while following good scientific practice. Moreover, it is necessary to make the legal requirements and framework for dealing with data clearer. The research funding agencies are called upon to offer programs to support pilot projects and research projects in need of professional research data management, and to provide means for the development of discipline-specific organizational forms. Information infrastructures have the duty to establish local services and advising. Corresponding activities should be cross-linked nationally in close consultation with representatives from the individual disciplines. Likewise, both researchers and employees of research institutions should be able to take advantage of the existing expertise in information management with a focus on research data in the form of training courses and continuing education. The working group viewed as another essential item the linking of research data repositories with appropriate publication databases from publishers. Federal and state governments, as institutions funding research with public money, were advised to view research data as a national cultural asset and therefore to create possibilities for its sustainable preservation and future access. This must be accompanied by a clear definition of responsibilities and the installation of appropriate organizational structures as well as cooperation in clarifying legal frameworks. In addition, they expressed the need to provide immediate resources for a fundamental establishment of an appropriate research data infrastructure.

The paper "Comprehensive approach for information infrastructure in Germany" ("Gesamtkonzept für die Informationsinfrastruktur in Deutschland"), prepared by the commission, was presented to the GWK in May 2011. The concept as a comprehensive planning document and the process of structural cooperation between all key players in information infrastructure was welcomed and the German Council of Science and Humanities (Wissenschaftsrat [WR]) was asked to incorporate the results by mid-2012 in its paper "Recommendations for research infrastructures."

In an extensive position paper that was published in January 2011, the German Council of Science and Humanities addressed the significance of information infrastructures for research in Germany.[60] In this paper, research collections, libraries, archives, and data collections in the broader sense were subsumed under the term information infrastructures. In line with the recommendations of other research policy organizations, the German Council of Science and Humanities attributed great importance to the development of information infrastructures for research in Germany. Special emphasis was placed on the role of universities, which needs to be significantly expanded, funding on the state level, meticulously coordinated planning, and the close integration of the academic community into the design process. The German Council of Science and Humanities advocated developing a comprehensive national strategy for information infrastructure ("nationale Gesamtstrategie für Informationsinfrastrukturen") for Germany by 2020.

2.3 Information Infrastructure Institutions

In 1996, the German Rectors' Conference (Hochschulrektorenkonferenz [HRK]) declared in its recommendations from the 179th plenum session that „in der Informationsgesellschaft […] Methoden und Techniken der Erzeugung, Verbreitung und Vermittlung von Wissen grundlegend verändern [werden]" ("there [will be] a fundamental change […] in the methods and techniques of generating, distributing and transmitting knowledge in the information society").[61] As an immediate need for action, „zentrale Einrichtungen […] innerhalb der Hochschulen […] Rechenzentren, Medienzentren und Bibliotheken verstärkt Dienstleistungsfunktionen für die Fachbereiche übernehmen [sollten]" ("data centers, media centers, and libraries as central institutions […] within the universities […] [should]

60 See Wissenschaftsrat (2011b).

61 German Rectors' Conference (1996).

take on more service functions for the individual subject areas").[62] In principle, this need still exists today. Fifteen years ago, the emphasis was on the use of information technology in instruction and the dissemination of set course material, rather than on the provision of research data. However, the requirements in information science and organizational matters hardly differ in both areas.[63] The increasing use of digital information systems in research is an irreversible process[64] accompanied by specific challenges. The operators of information infrastructures have a special responsibility. Discipline-specific information infrastructures were discussed above and in certain ways, they reflect the self-organized response to a subject-specific need for digital information services. The needs of researchers are foregrounded here, which in the past, owing to insufficient consideration of information science/technical requirements, led to solitary structures that were not networked in many cases.

The need for action on the part of the "central institutions" – in the sense of assuming responsibility – as identified by the HRK in 1996, is now increasingly being taken up by individual shareholders. The TIB is one example, as already mentioned. In addition to the basic functions of a library in the field of providing access to literature, the active design and development of systems for the information supply of digital content is becoming increasingly important. Crucial at this point is the employment of internal information/technical expertise in information services that reflects research requirements. In addition to organizational and information/technical aspects, the user's perspective must be examined and considered above all. Of great importance in this context is the creation and communication of added value in the sense of providing incentives not only to use digital information systems through *data retrieval*, but also to enrich these digital information systems by providing information. The TIB in Hannover, for example, established itself as an agency for the awarding of Digital Object Identifiers (DOI) and is one of the founders of

62 Ibid.

63 See Alliance of German Science Organisations (2008); Alliance of German Science Organisations (2010); and Kommission Zukunft der Informationsinfrastruktur (2011).

64 See Horlings et al. (2006).

the international DataCite Consortium.[65] On the federal level, the German National Library of Medicine (Deutsche Zentralbibliothek für Medizin [ZB MED]), the Leibniz Institute for the Social Sciences (Leibniz-Institut für Sozialwissenschaften [GESIS]) and the German National Library of Economics (Deutsche Zentralbibliothek für Wirtschaftswissenschaften [ZBW]) are also members of DataCite. The use of DOI allows for persistent referencing to digital content on the internet. This is a vital prerequisite not only for the reliable retrieval of specific content, but also of potential benefit to the reputation of the original data producer. Contents can be linked to the data creators in this way. In addition, the system supports interlinking academic publications and research data stored in individual repositories. At the same time, persistent identification enables the development and progress of innovative publishing activities for research data[66] and has in the meantime led to new cooperations between research infrastructure organizations and publishers.[67]

Such initiatives are examples of how "traditional" information infrastructure institutions could integrate themselves into the discussion about the improved handling of research data. Institutions above and beyond research libraries should also be involved in this discussion because the long-term preservation of research data carries diverse challenges that cannot be solved by individual actors. The discussion about preservation periods, for example, is not about a set, clearly-defined period of time, but in fact should be guided by the need to archive research data for the foreseeable future.[68] Concerning research data and the decision which data should be preserved and how, general questions must be considered about

65 See DataCite (2011), http://datacite.org.

66 See, for example, Earth System Science Data (ESSD).

67 See, for example, that when using DOI, there is alternation between Elsevier and the information system PANGAEA and between data sets at PANGAEA and digitally available publications at Elsevier.

68 Stefan Luther from the German Pension Fund (Rentenversicherung) referred to, for example, in his presentation in 2010 at the 11th Oracle Bibliotheken Summit in Weimar about the "Interaction between High-Volume Archives and Storage Platforms" („Zusammenspiel von hochvolumigen Archiven und Storageplattformen") that the problem is not the data volume in itself, but the deletion of data. In this concrete case, the unclear lifecycle of the data was the setting.

long-term data curation and, linked to that, sustainable long-term availability. In this domain, the archives with their subject expertise should be more involved since there is a need to develop appropriate measures and standards which comply with research requirements. Several state and university archives are already interested in addressing this challenge. In addition, the need to professionally preserve, archive, and provide research data in a professional way requires an IT environment which is matched with these tasks. At this point, the expertise from data centers could doubtlessly be incorporated. The primary task of data centers is in ensuring the undisrupted operation of the IT infrastructure of an organization such as a university. However, the tasks of data centers are changing. Increasingly, providing services for users, in addition to the purely technical maintenance of the IT infrastructure, plays a role.[69] It is still a topic for discussion as to which of these services could be offered by data centers concerning research data management. However, it seems to be obvious that the professional expertise of data centers is needed, for example, in dealing with high-throughput technologies in life sciences.

Research collections and museums with research departments have only recently begun to participate in the discussion about the provision of comprehensive digital information. The installation and the systematic use of digital systems in such institutions, above and beyond the strictly management and inventory duties of collections, are still in the early stages. On an international level, the need for a coordinated approach has been a topic of discussion among major natural science museums within the framework of the Scientific Collection International (SciColl) Initiative since 2006.[70] At the national level, this topic has already been addressed, and, for example, the DFG's available funding options, which previously focused on manuscript and printed documents, were expanded in 2011 with the announcement of a call for pilot projects dealing with the "classification and digitization of object-based research collections" („Erschließung und Digitalisierung von objektbezogenen wissenschaftlichen Samm-

69 See, for example, the description of tasks of the computing center at the University of Stuttgart: http://www.hlrs.de/.

70 See Scientific Collections International (SciColl), http://scicoll.org.

lungen“).[71] The intention of this broad call for applications was to investigate the need for and interest in this field, and to gain experience for the development of future funding measures. In addition to the specific technical and organizational challenges in the field of developing and digitization of objects, with the backup, archiving, and retrieval of digital information in this context, information specific technical requirements arise, that also play a role in research data. Therefore, all measures should be accompanied by the appropriate relevant professional expertise in the field of information management. Accordingly, in the recommendations of the German Council of Science and Humanities (“Wissenschaftsrat”) from January 28, 2011, it is stated that these measures require a high degree of initiative and self-organization on the part of the collections and their host institutions.[72]

As the basis of most research results, research data are an integral part of scholarly publications. Therefore, creating connections between scholarly articles in digital form and their underlying research data is not only useful but is virtually obligatory. This interlinking is made possible by the assignment of DOI’s and in some cases has already been implemented.[73] There is still no standard process for creating this link, however. Similarly, there is no clear agreement regarding the corresponding responsibilities. In this area, it is obvious that academic publishers could play an important role and should have the responsibility for actively supporting this interlinking. As a prerequisite for independent research and for the best possible access to research results, a free and unrestricted access to relevant data – linked via research publications or research data repositories – needs to be guaranteed.

71 See DFG (2010a).

72 See Wissenschaftsrat (2011a).

73 See, for example, ESSD (2011).

2.4 Funding Organizations

As major sources of financial support for research, funding organizations play a prominent role in this general discussion. The requirement that funding recipients deal in a professional manner with research data and participate actively in measures to develop suitable systems must be accompanied directly by the funding necessary for doing so. Funding agencies have already recognized, as far as possible, the need for action and their corresponding duties. However, there is no consensus about which specific and long-term supporting measures should be taken since there are numerous, still unresolved questions about the details. It is imperative to first develop basic frameworks, particularly against the background of not insignificant financial needs, of which the concrete amount is difficult to ascertain at present, and the overlap of institutional funding independent of third-party sources. The actual needs of the individual academic disciplines and interdisciplinary fields take precedence here, as do the accompanying responsibilities and the implementation of the widest possible acceptance of a sustainable handling of research data and the associated changes to scholarly research processes. These profound changes in the culture of research make the participation of researchers at an early stage in this process mandatory.

As part of its funding priorities up to 2015, the DFG has announced a call in 2010 for "information infrastructures for research data"[74] for the development and optimization of information infrastructures, trying to achieve the efficient and effective handling of research data.[75] The aim of this announcement was to encourage new measures for discipline-specific forms of organizations and to provide options for existing research data repositories to expand their services or to professionalize them. As a result of this call, projects in which information infrastructure institutions are cooperating closely with researchers emerged from a variety of academic disciplines. At the same time, since April 2012 the application procedure for research projects has required applicants to include a plan for handling

74 See DFG (2010b).

75 See DFG (2006).

research data in their research project proposal. The application guidelines state:

> If data measurements that are suitable for reuse are to be collected systematically using project resources, please explain what measures have been taken, or will be taken, during the duration of the project, to secure the data sustainably and to make them available for possible reuse. Please also consider, if possible, the existing standards in your subject discipline and the options available from data repositories.[76]

This requirement is intended to support the result that applicants – if necessary – will formulate concrete plans for the handling of digital research findings in projects funded by the DFG, and possibly align themselves with appropriate partners from the field of information infrastructures. This approach also creates the possibility for existing research data repositories to adapt or expand their services specifically in accordance with the requirements of the academic community and, in doing so, achieve greater acceptance and sustainability. With a similar objective in the context of funding Priority Programs (Sonderforschungsbereiche [SFB]), in 2007 it became possible to apply for funding for central subprojects that deal, in cooperation with appropriate infrastructure institutions, with data management and the sustainable availability of research data compiled in the SFB.[77] Currently, more than twenty Priority Programs have made use of this option. A similar adjustment in the application process took place in the second round of the German Excellence Initiative[78], in which the application form for excellence clusters was

76 DFG (2010c), p. 32: „Wenn aus Projektmitteln systematisch (Mess-) Daten erhoben werden, die für die Nachnutzung geeignet sind, legen Sie bitte dar, welche Maßnahmen ergriffen wurden bzw. während der Laufzeit des Projektes getroffen werden, um die Daten nachhaltig zu sichern und ggf. für eine erneute Nutzung bereit zu stellen. Bitte berücksichtigen Sie dabei auch – sofern vorhanden – die in Ihrer Fachdisziplin existierenden Standards und die Angebote bestehender Datenrepositorien."

77 See DFG (2009a).

78 The German Excellence Initiative aims to support top-class research and to advance the quality of universities and research institutions in Germany. The Excellence Initiative was initiated by the German federal and state governments and the DFG

amended with the addition of the requirement to provide information about data management. This addition was designed to ensure that the handling of data should be covered as a central issue and appropriate measures considered at an early stage.

The German Federal Ministry for Education and Research is also pursuing the goal of ensuring the future (re)usability of project results with the requirement for funding applications to provide information in the context of a "plan for realization." In addition to a description of the scholarly, economic and technical prospects for success, it is about the project's adaptability in terms of the sustainable use of project results in subsequent academic projects, in applied research, or in the context of commercial applications.[79] For the preparation of projects, the BMBF provides an overview of subject information centers and nationwide information centers, which generally differ from research data repositories.[80] At present, a description of essential measures for securing, archiving, and re-using research data and other digital research outcomes is not included. This approach is strongly influenced by different strategies in the individual research areas and it is exclusively project-orientated.

With the D-GRID initiative[81], the BMBF has supported a national IT infrastructure since 2005 that has the goal of providing a high-performance computing and storage structure for academic research as well as for industry. Numerous individual projects within the D-GRID initiative address scholarly objectives as well as operational and commercial applications in addition to projects to guarantee the operation of the grid on different levels. D-Grid is characterized by intensive cooperation between industrial partners and academic and research institutions. In the scholarly context, a number of disciplines created specific grid-based initiatives that

---

was commissioned to run the initiative in cooperation with the German Council of Science and Humanities (Wissenschaftsrat).

79 See BMBF (2011a).

80 See BMBF (2011b).

81 The D-Grid Initiative (German Grid Initiative) was a federally funded project with the objective to develop computer infrastructure for research and education. Using and implementing grid computing technology the initiative started in 2005 with seven projects including an integrational project and several partner projects.

addressed the characteristic requirements for data processing and made available corresponding options. However, the focus has often been placed on applications and services that are comparable to virtual research environments. The subjects of sustained data and information management, data preservation and the support for and provision of research data have only been emphasized in a few projects up to this point (such as AstroGrid, C3Grid, TextGrid, and WissGrid).[82]

At present [2012; Ed.], there is no systematic financial support from other funding agencies, foundations, and other donors for projects dealing with the management of research data.

82 See D-GRID, http://www.d-grid-gmbh.de.

3 Status of Discussion and Current Activities: The International Perspective

Stefan Strathmann

German issues of research data management are essentially identical to international issues in this field: It is often organizational adjustments to the structural framework that are lacking in order to address the changing requirements of a research which is based on the extensive use of information technology.[83] As in Germany, it is primarily grassroots projects and a few flagships that are actively addressing the challenges of sustainable research data management. However, there are some organizations and institutions that have been trying to guarantee long-term archiving and availability for several years. The organizations and institutions discussed briefly below represent only a selection of those working in this area.[84]

3.1 International Organizations

The choice of institutions and organizations presented here takes account of the wide variety of different approaches concerning the digital curation of research data. Therefore, the selected institutions are to be seen representatively for a larger group of similar acting facilities.

---

83 See Chapter 2.1.

84 In addition to the institutionally-based activities and discussions outlined in this chapter, there are naturally international discussions taking place. In addition to the community-based discussions, the e-mail list Research-Dataman is particularly worthy of mention (https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=RESEARCHDATAMAN). It was initiated by the Digital Curation Center (DCC) in Britain on behalf of the Joint Information Systems Committee (JISC). http://www.jisc.ac.uk.

3.1.1 United Nations Educational, Scientific and Cultural Organization (UNESCO)

The "Charter on the Preservation of Digital Heritage,"[85] which was adopted on October 17, 2003, at the 32th UNESCO General Conference,[86] also includes scholarly research data:

> The digital heritage consists of unique resources of human knowledge and expression. It embraces cultural, educational, scientific and administrative resources, as well as technical, legal, medical and other kinds of information created digitally, or converted into digital form from existing analogue resources. Where resources are "born digital," there is no other format but the digital object.[87]

Digital research data are considered as part of the digital cultural heritage and the UNESCO member states are directed to preserve this heritage in order to "[…] ensure that it remains accessible to the public."[88] In reference to research data, the charter states that "Measures should be taken to […] encourage universities and other research organizations, both public and private, to ensure preservation of research data."[89] With this charter, the United Nations called attention at an early stage to the necessity of comprehensive measures for the preservation of the cultural and scholarly heritage and committed member states to conserving this heritage.

3.1.2 Organisation for Economic Co-Operation and Development (OECD)

Considering the enormous costs involved in the creation of research data, research data management is a topic for the OECD as well.[90] As early as 2004 archiving and providing access to publicly funded research data were

85 UNESCO (2003).

86 See UNESCO Homepage: http://www.unesco.org.

87 UNESCO (2003).

88 UNESCO (2003).

89 UNESCO (2003).

90 See Organisation for Economic Co-Operation and Development Homepage: http://www.oecd.org.

the subject of the final document[91] of the "OECD Committee for Scientific and Technological Policy at Ministerial Level." The research ministers came to the following conclusion:

> Co-ordinated efforts at national and international levels are needed to broaden access to data from publicly funded research and contribute to the advancement of scientific research and innovation. To this effect, Ministers adopted a Declaration entrusting the OECD to work towards commonly agreed Principles and Guidelines on Access to Research Data from Public Funding.[92]

The access to publicly funded research data is also the subject of an annex (Annex I: "Declaration on Access to Research Data from Public Funding") to this document. On the basis of these ministerial decisions, recommendations for managing publicly funded research data were developed.[93] These "Principles and Guidelines for Access to Research Data from Public Funding" were adopted and published by the OECD. OECD recommendations are not legally binding, but "are considered to have a great moral force" nevertheless.[94] The final section is dedicated to sustainability:

> Due consideration should be given to the sustainability of access to publicly funded research data as a key element of the research infrastructure. This means taking administrative responsibility for the measures to guarantee permanent access to data that have been determined to require long-term retention. This can be a difficult task, given that most research projects, and the public funding provided, have a limited duration, whereas ensuring access to the data produced is a long-term undertaking. Research funding agencies and research institutions, therefore, should consider the long-term preservation of data at the outset of each new project, and in particular, determine the most appropriate archival facilities for the data.[95]

---

91 See Organisation for Economic Co-Operation and Development (2004).

92 See Organisation for Economic Co-Operation and Development (2004).

93 See Organisation for Economic Co-Operation and Development (2007).

94 This type of instrument is often referred to as "soft law"; see Organisation for Economic Co-Operation and Development (2007), p. 7.

95 Organisation for Economic Co-Operation and Development (2007), p. 22.

It is becoming more and more evident that especially the recommendation to prepare for research data curation from the very beginning of data creation – both on the part of the researchers and of those providing research funding – is one key to successful research data management.

3.1.3 European Union (EU)

In the research framework programs of the EU, the issue of research data infrastructures has been of special interest for several years now. Just as projects for the preservation of cultural objects, projects for the digital curation of research data have been, and continue to be, promoted. Examples include the completed project PARSE.Insight[96] or the ongoing project APARSEN[97], both of which were initiated by the "Alliance for Permanent Access."[98] Currently, the development of research infrastructures, which includes data infrastructures, is one of the main priorities of the EU Seventh Framework Programme (FP7).[99] That is partly due to the implementation of the guidelines of the European Strategy Forum on Research Infrastructures (ESFRI).[100] The forum is a strategic instrument for the development of research infrastructures in Europe. The integration and strategically defined development of policies for research infrastructures are the main focus of this group, whose members are nominated by the research ministers of the member states.

A vision of an EU strategy for dealing with research data can be found in the report "Riding the Wave: How Europe can gain from the rising tide of scientific data" from the High-Level Group on Scientific Data.[101] A central component of the vision outlined in this report is a collaborative

96 See PARSE.Insight Homepage: http://www.parse-insight.eu.

97 See European Commission, http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri.

98 See Alliance for Permanent Access Homepage: http://www.alliancepermanentaccess.org.

99 See European Commission (2011).

100 See European Commission, http://cordis.europa.eu/fp7/capacities/research-infrastructures_en.html.

101 See High Level Expert Group on Scientific Data (2010).

data infrastructure in which data represent the actual infrastructure and the physical and technical infrastructure is in the background.

In addition the report of the "Comité des Sages"[102] of January 2011 draws a direct connection between increasing digitization and the corresponding need for research data curation.

3.1.4 World Health Organization (WHO)

With the participation of the World Health Organization (WHO)[103] and the Wellcome Trust,[104] a number of organizations supporting proposals in the field of health care have agreed on common funding goals, visions, and principles in recent years and in early 2011 published a code of conduct in a joint statement.[105] In particular, this statement describes important requirements concerning data management and access to data created by means of this funding. The long-term goals include the following:

- Data collected for health research are made available to the scientific community for analysis which adds value to existing knowledge and which leads to improvements in health […]
- To the extent possible, datasets underpinning research papers in peer-reviewed journals are archived and made available to other researchers in a clear and transparent manner[106]

The long-term availability of research data in the medical sector is thus a clearly stated goal of the funding organizations involved.

3.1.5 Knowledge Exchange

Knowledge Exchange[107] is a collaborative consortium of four European funding institutions to achieve their funding goals more efficiently through

102 See Niggemann; De Decker; Levy (2011).

103 See World Health Organization (WHO), http://www.who.int.

104 See Wellcome Trust Homepage: http://www.welcome.ac.uk.

105 See Wellcome Trust (2011b).

106 See Wellcome Trust (2011a).

107 See Knowledge Exchange Homepage: http://www.knowledge-exchange.info.

the coordinated use of resources. The partners are Denmark's Electronic Research Library (DEFF),[108] the German Research Foundation (Deutsche Forschungs-gemeinschaft [DFG]),[109] the British Joint Information Systems Committee (JISC),[110] and the SURFfoundation[111] in the Netherlands.

Stable and long-term access to research data is one focus of their joint activities.[112] The website of the cooperation states:

> In the future of academic and scholarly communication compound publications in the means of article and research data will play an ever increasing role. Research data have to be accessible both as open access and in the long term but also in different environments and tools.[113]

With the requirement that research data must not only be openly available in the long term, but also made accessible for various environments and with different tools, Knowledge Exchange continues the discussion about a life cycle model in research data curation.[114]

3.2 Model Realizations

3.2.1 National Science Foundation (NSF)

Since the beginning of 2011, the NSF[115] in the United States has extensively revised the regulations for submitting a funding proposal.

> Proposals submitted or due on or after January 18, 2011, must include a supplementary document of no more than two pages labelled "Data Man-

108 See Denmark's Electronic Research Library (DEFF), http://www.deff.dk/english/.

109 See Deutsche Forschungsgemeinschaft (DFG); Chapter 2.1.

110 See Joint Information Systems Committee (JISC), http://www.jisc.ac.uk.

111 See SURFfoundation, http://www.surffoundation.nl.

112 Among other things, Knowledge Exchange established a separate working group for research data (see Knowledge Exchange (2011b)).

113 Knowledge Exchange (2011a).

114 See, e. g., Digital Curation Centre (2011a) and Australian National Data Service (ANDS): http://ands.org.au/about-ands.html.

115 See NSF Homepage: http://www.nsf.gov.

> agement Plan". This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results.[116]

All applications for funding now require a "data management plan (DMP)," in which the data that will be produced during the funding period must be described and which explains the way in which data will be handled.[117] This includes the publication as well as the archiving of data:

> The DMP should describe data formats, media, and dissemination approaches that will be used to make data and metadata available to others. Policies for public access and sharing should be described, including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements. […]
>
> The DMP should describe physical and cyber resources and facilities that will be used for the effective preservation and storage of research data. These can include third party facilities and repositories.[118]

A proposed plan for ensuring sustainability beyond the duration of the project is emphasized by information about the plans for research data curation as well as by the inclusion of a report in all subsequent proposals and applications regarding the obligation to maintain and manage the data inherited from previously funded projects.[119]

3.2.2 Australian National Data Service (ANDS)

The ANDS[120] is establishing an Australian Research Data Commons (ARDC).[121] This commons will include research data from all disciplines,

---

116 NSF Data Management Plan (2011).

117 This requirement applies generally to all applications. The specific requirements for each directorate, funding line, and so on, are regulated individually; see NSF Data Management for NSF SBE Directorate Proposals and Awards (2010).

118 NSF (2010), p. 3.

119 Specifically, this means that "data management must be reported in subsequent proposals by the PI and Co-PIs under 'Results of prior NSF support'."; see NSF (2010), p. 4.

120 See ANDS Homepage: http://www.ands.org.au.

121 See Australian National Data Service (2007).

all universities and all research institutions that are supported by public funding in Australia. Australian research data are to be transformed as a whole into a strategic national infrastructure. To achieve these goals, the ANDS is meeting various organizational and technical challenges and promoting the development of an infrastructure. Its objectives include the following:

- Support for research data managers
- Promotion of the transfer of research data in stable and accessible research data curation environments
- Provision of training opportunities in the field of data management that are independent of institutions or communities.
- The ability of researchers to access the Australian data commons and to work with it.
- Support for the integration of Australian research data in international as well as national and multi-disciplinary research groups.[122]

For a practical implementation of these goals, a number of infrastructure components are currently being created:

- Data Capture Infrastructure (especially the integration of existing infrastructure)
- Research Metadata Store Infrastructure (the goal is a combination of "data stores" and "metadata stores")
- Automatic Public Data Publication Infrastructure (for the publication of descriptions of data collections e. g. from public authorities [meteorology, statistics, etc.] or other providers of research data aggregations [libraries, museums, etc.]).
- Australian Research Data Commons Core Infrastructure (persistent identification, authority files, controlled vocabulary, retrieval options, etc.).
- Australian Research Data Commons Applications Infrastructure (possibilities for data integration, data visualization, and data analysis).[123]

122 See Australian National Data Service: http://ands.org.au/about-ands.html.

123 See Australian National Data Service: http://ands.org.au/ardc.html.

The current budget (2012/2013 financial year) for these activities from the National Collaborative Research Infrastructure Strategy (NCRIS) and the Education Investment Fund (EIF) sums up to $71,037,566.45[124].

As part of the efforts to build this comprehensive research data infrastructure, some of the functions of the Australian Research Data Commons (ARDC) have already been implemented and are accessible via the ANDS website.[125] For instance, it is already possible to register research data with accompanying metadata and persistent identifiers, and to make data accessible. The actual storage and archiving is the responsibility of the institutions and researchers who provide the data. The ANDS serves as a "metadata store," meaning that it does not provide a way to archive research data. However, it does provide a search function and access to the extensive data resources which have already been registered. These options are supplemented by a series of helpful guides to best practices in research data management, research data curation, and the legal implications for the publication of research data. In addition, there are a variety of information and training sessions that familiarize (potential) users with the services of the ARDC and explain the basics of research data management. Such an ambitious national approach is just one way to ensure long-term access to research data. In addition, there are several other approaches, by far more common, to ensure the digital curation of research data. They may be based on individual projects, on national or international efforts or on community-specific or institutional activities.

124 See ANDS (2012–13), p. 14.

125 See ANDS Homepage: http://www.ands.org.au.

4 Methodology: Subject of the Study

Heike Neuroth

Even if many questions in connection with research data curation and availability of research data will remain to be solved, it is an important first step to investigate the status quo and the needs of different disciplines and their actors. In this way, it will be possible to derive requirements for research data infrastructures and develop strategies to realize them. Until now, this type of mapping of the research landscape has not been carried out in reference to sustainable research data management in Germany. Such a task is difficult to carry out comprehensively, since there is currently no concept that would reliably ensure the visibility of these individual approaches – regardless of questions about the potential gain in knowledge that would come from a comprehensive survey of the research landscape. As long as there is no comprehensive overview available, the inventory of the individual fields of research presented below can serve as an initial orientation in this area. There was no systematic process for selecting subject areas; instead, some disciplines were selected to serve as case studies in order to represent various research areas which have emerged in the context of nestor, the eScience Initiative of Germany, and the German Grid Initiative. These disciplines were selected for the following reasons (see figure 1):

- the subjects of their research are digitally available (e.g. a 3-D scan of a museum object) or their research generates digital research data;
- research data are frequently published together with research findings;
- these data are intended for long-term archiving and to be kept available for subsequent re-use;
- They are actively contributing to the creation of a (sustainable) research data infrastructure, and therefore their initial deliberations and experiences on this subject are available.

The subject areas presented here cover a wide range of disciplines, from the humanities to the natural sciences, including medicine. Unfortunately, it was not possible to identify exemplary approaches in Germany in all

disciplines, and in several academic disciplines (such as life sciences or engineering), existing practices and solutions could not be considered. For this reason, the subject overview presented here is a sample and does not claim to represent a comprehensive representation of the situation in all disciplines in Germany.

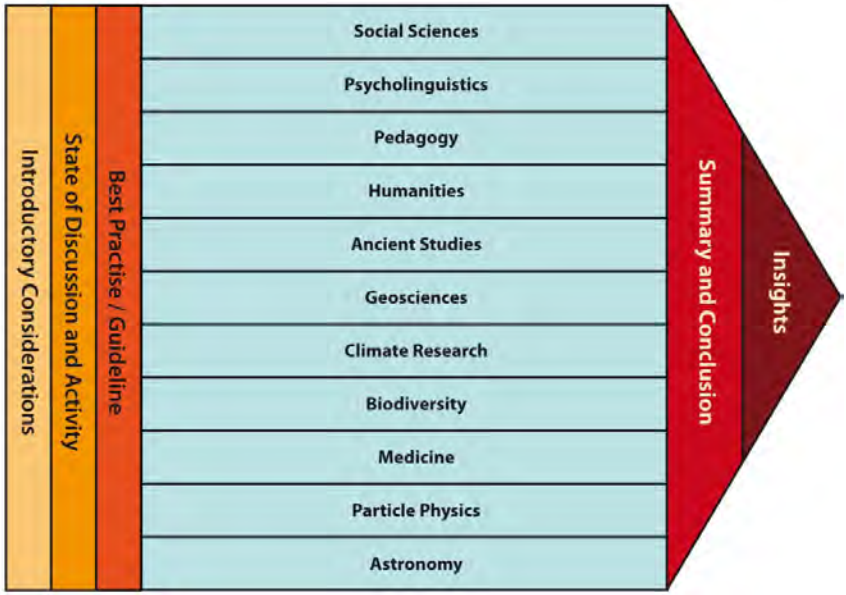


Figure 1: Structure of this volume

4.1 Structure of this Volume

Against the background of a brief overview of national and international discussions and developments (see the two previous chapters regarding the "Status of Discussion and Current Activities"), this chapter presents the list of key questions submitted to members of the individual disciplines in Germany to collect information about their data management practices and approaches. The next chapter provides a comparison of the differences and similarities of the approaches that were observed. From these compari-

sons, conclusions can bedrawn about factors that promote the creation and operation of both universal and subject-specific research data infrastructures as well as research data curation. These conclusions can serve as a model for other disciplines or provide information and recommendations about further developments. The final chapter summarizes our findings and describes certain areas of action that are currently topics of discussion on national and international levels.

4.2 Key questions for mapping research disciplines

The following list and description of key questions, with explanatory background information as necessary, was provided to researchers from the individual disciplines. Using this standardized structure to gather information about subject-specific levels of development, it is possible to compare the status quo regarding research data curation practices among different disciplines. Furthermore, comparative analyses can feasibly be used to support the deve-lopment of research data infrastructures in Germany. In March 2011, at the invitation of the D-Grid GmbH at the Technical University (TU) Dortmund, a workshop was held with all stakeholders, during which initial comprehensive solutions and strategies were discussed.

4.3 Introduction to the Research Area

Following is a characterization of the disciplinary environment, description of the research field and existing structures within it.

4.3.1 Background

Disciplinary differentiation such as more (inter)nationally coordinated collaborative projects versus heterogeneous project landscape etc.; in the case of a project, relevant information includes e.g. structure, funding, partners, as well as objectives, background information, results of the project, etc.

4.3.2 Cooperative Structures

- Is research *collaboration among institutions* the rule or rather the exception? Background: Cooperative collaboration increases the need for methods of data exchange. This pressure has a positive effect on the development of standards for data exchange and standardized data models. Scholarly cooperation creates a stronger incentive to prepare data for re-use.
- Is there an institution that is already in charge of providing (centralized) research data curation services for the whole discipline or that collects and documents research findings for the entire field? Background: An already existing central facility could handle and coordinate (central) research data curation services Germany-wide.
- Is there usually collaboration with internal or external service infrastructure institutions (e.g. an ICT department, library, computing center, etc.)? Background: Examples of how to describe this collaboration could include: Collecting and making available research data and/or publications. Such an institution could take on research data curation tasks, for example. In this case, the (future) designation of responsibilities, roles/duties, etc. in the area of data management or research data curation would also be easier. If such approaches are already being used, please describe them here. If not, please state this as well.

4.3.3 Data and Metadata

This section of questions does not focus on the general research data curation of publications, such as institutional (document) repositories, but on the classification of data sources or different types of research data in

digital form. The emphasis is in particular on research data that are (intended to be) published.

- What *types of data* are generated in the research field, e.g. by large instruments (telescopes, accelerators, etc.), simulations, laboratory experiments, field measurements and surveys or digitized objects (digital documents, digital archive items, digital research findings or digital museum objects, etc.)? Background: Depending on the type of data available, specific research data curation strategies must be defined and developed. If there are strategies, please list them here.
- How are data *published and made available long-term* in the research field? Are there any established data centers or data archives? Background: If a culture of publishing research data already exists, it is very important to establish management structures for research data curation.
- Are there any minimum requirements for the *integration of research data* into a data center (such as format specifications, quality control, metadata, persistent identifiers, and so on)? Are there any *management plans* for research data? Background: If these exist, please describe them here as well, since they could be (partially) re-used by other disciplines/research fields.
- What *volume of research data* is produced each year? What is the growth rate? Background: Large amounts of data, such as many petabytes, may require different research data curation strategies than smaller volumes of data, which are significantly more heterogeneous in data type and/or data format.
- What are the *standardized formats* for research data? Are there any recommendations for specific formats? Background: Formats are essential for; if a discipline or field of research has already agreed on a specific (standardized) format, this is particularly important to list.
- Are research data subject to *limited-use restrictions*, e.g. through data privacy protection, legal requirements, individual rights, copyrights, etc.? Background: Restrictions in the use of research data directly affect research data curation and must therefore be taken into account right from the beginning (in matters such as policies, technology, etc.)
- How important is it to re-use *older* research material, research reports and research data, etc.? The term “older” refers to digital research data

that are no longer easily accessible. For example, these data could be stored on internal servers without adequate metadata documentation and therefore could not be interpreted correctly anymore. Background: If only the latest articles and research findings are of significance in a particular discipline, there may be no need to curate these data. Otherwise, procedures must be developed and established to maintain these research data in the research cycle.

- Are *metadata* (possibly even standardized metadata) used for descriptive, structural and administrative description, including metadata for the persistent addressing of research data? Which persistent identifier system (e.g. DOI, Handle, URN) is used? Which (subject-specific) metadata schemes are used? Are metadata, along with research data, also documented (perhaps even standardized) and stored with the description of the technical requirements (e.g. to document or even archive hardware and software frameworks)? Background: If there is no (standardized) metadata describing the research data and no persistent identification is assigned, this tends to suggest that the research data can (or possibly should) only be used by a small circle of researchers. Research data cannot be interpreted without descriptive, technical, and administrative metadata. Data that have been exported from the system in which they were produced, and whose structure and origin were not documented, can only be used in new contexts with a significant investment of resources.

4.3.4 Internal Organization

- Is research data curation implemented according to *established processes and rules*? Are there any established strategies, policies, procedures, and implementations? Are there any specific cooperation with other partners (national, international)? Background: Without this type of structural framework, it is not possible to run a data archive sustainably.
- How is the repository *funded*? Are there any fixed resources for research data curation allocated in the budget? Is there any secure (or potential) funding by the federal government / the federal states? Is a flat rate charge for data (for example, per a defined amount of data)

collected from (external) projects? Background: Only after long-term funding there can be regulated, institutionalized long-term preservation.

- What are the estimated *costs* for the development (one-time initial costs) and the operation of your data archive? Background: For some data archives in Germany, there are already rough estimates available. There can only be a "national strategy" and therefore sustainability in the area of research data curation when policy makers, funders, research fields, and service infrastructure institutions etc. know about the estimated costs.
- Are there any *specially trained staff and data specialists* (such as researchers, data managers, information professionals, IT experts, etc.), who deal primarily with research data curation? Background: If this is not the case, individual researchers would have to learn about many important aspects, which is not conducive to a homogeneous approach in this area. In addition, researchers alone are thus responsible for research data curation, which might be less sustainable.
- Are researchers and/or those in charge of data management employed *on a permanent basis at an institution*, or is the need for staff predominantly handled by temporary staff positions? Background: In case of a high level of staff turnover, the need for standardized data archiving and standardized documentation of the data is increasing. At the same time, however, staff motivation to perform high-quality data archiving could decrease if researchers know that they will work on a particular problem only for a short period of time.
- Are *third party services* being used for research data curation? Background: For smaller institutions or departments, outsourcing these tasks to larger institutions can be a possible solution for the long-term preservation of research findings.

4.3.5 Perspectives and Visions

- Are there any *specific issues and challenges* that have not previously been addressed and that are relevant to research data curation?
- What are the *possibilities* for *initiating and supporting* the universal and long-term use of research data (data sharing, data re-use, data

publication)? Examples include supporting different stakeholders (researchers, trained IT / data curation experts, etc.) and certain infrastructure fields (such as persistent identification, authentication, technical maintenance of data repositories) as well as research data infrastructures, funding organizations, EU guidelines, incentive systems, training programs, etc.

- What are the *desires/visions* for research data curation, and who can help in their implementation? What is lacking, for example, and how can external support be utilized in the best way (e.g. on a national level)?

5 Summary and Interpretation

Jens Ludwig

In a synopsis of the different disciplines, one might get the impression that the situation with research data is similar to that of the animals in a "Chinese encyclopedia" described by Borges:

> [...] that "animals are divided into: (a) belonging to the Emperor, (b) embalmed, (c) tame, (d) sucking pigs, (e) sirens, (f) fabulous, (g) stray dogs, (h) included in the present classification, (i) frenzied, (j) innumerable, (k) drawn with a very fine camelhair brush, (l) et cetera, (m) having just broken the water pitcher, (n) that from a long way off look like flies".[126]

People have very different understandings of the term "research data" and the infrastructure associated with it, and very different expectations as well as dimensions to describe data conflict with each other. These differences may seem unnecessarily complicated and could generate the need for a single, clear definition for the field of research as a whole. However, one goal of the following comparison of essential characteristics of research data, as reported by researchers from various disciplines, is to demonstrate that this diversity does not in general represent a deficit, error or lack of development in the disciplines, but is the necessary result of the differentiation of academic research.

The eleven academic disciplines surveyed in this study are in themselves inherently complex areas of research. With one or two exceptions, all indicated that their areas of research are either interdisciplinary activities, in which different disciplines work together to investigate a topic (such as biodiversity) or that research is carried out in a highly differentiated discipline in which very different topics are examined (e. g., in the geosciences). The two disciplines in which the above-mentioned aspects are considered to be less important are particle physics and astrophysics, which, as their names suggest (with no claim to accuracy in terms of sci-

126 Foucault (2006), p. xvi.

ence history), could be regarded as independent branches falling under the parent discipline of physics.

One difficulty in comparing the approaches dealing with research data in this broad selection of disciplines is that even individual disciplines are comprised of very different subsets and therefore require not just one, but many types of data management. A more precise definition of the disciplines would not overcome this difficulty, however, because research areas will always require different types of data management. It is one of the characteristics of research that the methods, tools, and requirements for research data in the study of the same subject are diverse and that they change as part of the research progress. Just as the use of research data is diverse within a discipline, whether it be broadly or narrowly defined, thus correspondingly the methodology and the means of dealing with research data is rarely unique to one discipline, but rather can be observed in a variety of disciplines. Although disciplines do have specific characteristics, no successful research field is so specialized that others will not adopt its procedures for the treatment of their topics (such as the medical magnetic resonance imaging to visualize the brain that is used in psycholinguistics), and thereby the specific characteristics of the other discipline will be put into perspective.

5.1 Cooperative Structures

Research thrives on an intensive exchange of information and on cooperation. All eleven disciplines represented here reported on cross-institutional collaboration, although to different degrees and for different reasons. Driving factors are in particular the research instruments and the objects of investigation. Most notable are perhaps particle accelerators and telescopes that can neither be funded nor operated efficiently by individual institutions. But even if instruments the size of a building are not needed, data collection can require such a considerable investment of resources that it is no longer manageable individually but only through cooperation.

This was, e. g., the case in the social sciences and education; the major surveys carried out by these disciplines require a coordinated approach.

The object of investigation itself can be a reason to cooperate as well. The instruments can be relatively small, unspectacular and manageable by individuals or individual institutions, but the size, distance or distribution of the object under investigation could make cooperation necessary (e. g., in the fields of climate research and classical studies). Finally, for some areas, interdisciplinarity and the differentiation of individual disciplines are reasons to enter into cooperation in order to pool the diverse types of expertise necessary (e. g., in bio-diversity, medicine, and classical studies), which individual scientists cannot possibly master as a whole anymore.

This is primarily about collaboration in the context of research questions, but also the management of research data is collaboratively organized. Among all the disciplines surveyed here, the social sciences and the climate sciences have implemented the centralization of tasks on an institutional level most comprehensively. The German Leibniz-Institute for the Social Sciences (Leibniz-Institut für Sozialwissenschaften [GESIS]) and the German Climate Computing Center (Deutsches Klimarechenzentrum [DKRZ]) are independent institutions whose core function is to offer these services. In the geosciences, education and to a certain degree also in psycholinguistics, established institutions such as the World Data Centers, the German Institute for International Educational Research (Deutsche Institut für Internationale Pädagogische Forschung [DIPF]), and the Max Planck Institute for Psycholinguistics (Max-Planck-Institut für Psycholinguistik [MPI PL]) have taken on these tasks for others in addition to their own research activities. In all other disciplines, research data management is carried out in federations or through individual solutions at the institutions where the data is created.

Whether or not these individual solutions are sensible and efficient is difficult to judge. Data management comprises both subject-specific and general tasks;[127] and it is frequently stated that the former cannot be handled competently by interdisciplinary institutions without expert knowledge of the field. These subject-specific areas indicate that individual solutions might be better positioned. But an institution with a broader

127 See, e.g., Kommission Zukunft der Informationsinfrastruktur (2011), p. B125.

scope can often make use of economies of scale for generic services and ideally has specific expertise in order to provide certain subject-specific services efficiently. It is not possible to clarify easily or in general which scope makes sense for centers – what kinds of disciplinary granularity, disciplinary knowledge and tasks they should have – or whether individual solutions are preferable.

Accordingly, the relationships between the institutions who manage research data and those who produce or use it can vary widely. A classic interdisciplinary information institution such as a library or an archive usually exists to serve several other institutions or research groups and operates a variety of information systems, each of which contains, in turn, several data collections. In the area of research data, the mapping between institutions, data collections and users can be totally different. In the social sciences, for example, data centers often are facilities that manage mainly the data collection of just one source, such as a public authority. In the case of particle physics, the data from one source (such as the LHC accelerator) are not maintained by a single institution but are instead preserved, made available, and analyzed by an entire federation.

Even though funding agencies and research organizations are increasingly requiring the involvement of traditional, cross-disciplinary information institutions such as libraries and computer centers in collaboration for research data management,[128] it is not clear, from the outset, that these institutions will play a role in the collaboration. In the case of libraries whose future role in the field of digital information is often regarded as uncertain, this is less surprising than in the case of data centers. In the various collaborations for research data management presented here, information infrastructure institutions are mentioned in about half of the cases (humanities, medicine, geosciences, psycholinguistics, education, and biodiversity). Computer centers and libraries are mentioned equally often, and the German Institute for Medical Documentation and Information (Deutsches Institut für Medizinische Dokumentation und Information [DIMDI]), is often referred to as an example of a documentation center. The libraries mentioned by the respondents are major institutions, in particular the German National Library (Deutsche Nationalbibliothek [DNB])

128 See, e.g., Neuroth (2012), Chapter 2.1.3; DFG (2009a), (2009b).

and the Technical Information Library (Technische Informationsbibliothek Hannover [TIB]). The survey showed that the tasks carried out by infrastructure institutions usually include data hosting or assignment of persistent identifiers for research data, such as DOIs by the TIB as part of the DataCite consortium, and URNs by the DNB or handles by the Göttingen Society for Scientific Data Processing (Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen [GWDG]). Whether these basic services can provide multidisciplinary information institutions with a permanent role in collaboration for research data management, and what other services (such as advising) they could offer, remain open questions, as well as many other organizational issues.

5.2 Data and Metadata

There is a wide range of answers, as already indicated, to the question: "What types of research data are to be found in a research field?" The most common responses include video, audio, simulation data, photos, quantitative/qualitative data, digitized images/scans, markup/annotations, observation data, statistics, documents, experiment data, time series and remote sensing data. Categorizing these responses, two main types of data that are recognized as research data can be differentiated: in about sixty percent of the responses, research data are defined extrinsically, i.e., they are defined by their role in research or by the method used to create or make use of them, such as simulation data, observation data, experiment data, time series, interviews, and so on. However, the internal structure and the technical format of these different types of data can even be identical. In another thirty percent of the responses, data are instead intrinsically characterized by the type of media such as video, audio, mark-up, 3-D models, etc. In these cases, the term research data is used to express the distinction between data and documents, or at least between data and documents that are only used as publications or articles and not for recording measured data or interview transcripts. A certain percentage of the

responses is hard to categorize, such as biomaterial data, which could perhaps be defined as a data type characterized by the object of investigation.

The reason for the variety responses is that, depending on the methods and objects of investigation in the individual discipline, different criteria are relevant for the differentiation of research data. In the natural sciences, the difference between observation, experiment, and simulation data can be minimal in terms of encoding and the technical requirements for individual data records, but this distinction is essential when it comes to deciding whether the data is worthy of preservation and what background information will be necessary. In contrast, observation data alone is created by the analyses of society and of the individual that are carried out in the field of social science. The basic decision about which tasks are necessary in research data management work processes in this discipline is determined by the difference between quantitative and qualitative data, which have to be handled completely differently e. g., in terms of data privacy protection. However, in the fields of the arts and humanities, focused research on an individual case-by-case basis is much more frequent and important. Technologies are much more heterogeneous, and media categories such as photos, videos, and documents often provide the best basis for categorizing data.

The diversity of this characterization of research data draws attention to its context dependence and to the vagueness of the term research data. There is little point in restricting research data as regards content or its sources, because in principle, everything can serve as an object of investigation in scholarly research. The statement that data represents “research data” refers more to their methodological use in a particular scholarly context. If literary scholars read and analyzed a digitized book in the same way as its analog counterpart, this book would normally not be considered as research data merely because it is digital. If, however, the same book is analyzed by humanities scholars as part of a large digital linguistic corpus regarding particular patterns and word frequencies, these activities are possibly not only superficially similar to the interpretation of measured data in the natural sciences; potentially they use the same statistic proce-

dures and technologies for pattern recognition as well. In this case, this book should obviously be considered as part of the research data.[129]

This context dependence has an additional temporal dimension. Data that were created specifically for scholarly analysis could be defined as research data from the beginning. However, data that were not created specifically for research could nevertheless become research data at a later point, e. g., if a scholarly interest in studying them arises later on and they will be used in this function only then.[130] As a consequence of both the context-dependent nature of research data and the difficulty in determining fixed definitions for them, decisions and classifications that rely solely on the basis of whether or not data can be regarded as research data can be highly problematic. A (hypothetical) university guideline to the effect that all research data should be archived at a data center, and all other digital resources should be preserved by a library, would certainly require some arbitrary definitions and produce many exceptions.

Research data formats are so numerous and diverse that it is hardly possible to map them in an appropriate way. All disciplines seem to have one thing in common: the use of subject-specific formats in addition to the generally established formats. However, the various disciplines handle the heterogeneity and diversity of formats very differently. Four basic approaches can be distinguished: 1. the formats are limited through policies. 2. The formats are effectively restricted. 3. The formats are effectively not

129 See Michel et al. (2011) for an example of the use of a digital collection of books as research data.

130 See, for example, ship's logs from the time of the First World War, which were transcribed by internet users in the Old Weather Project, in order to make the historical weather observations in them useable for climate research (see Old Weather Homepage: htttp://www.oldweather.org). Conversely, David Rosenthal has drawn attention to the fact that interesting data, such as internet advertisements, are not being collected and archived on account of the technical and legal difficulties associated with it. The few institutions that are taking part in archiving the internet leave ads out, and therefore they will not be available to future researchers. A similar situation can be found in the US presidential election campaigns in 2008, in which blog entries and YouTube videos were central documents (Rosenthal 2011).

restricted. 4. The formats cannot be restricted. In the first approach, there are explicit format specifications defined by the institution that manages the data. In the second approach, the institution in charge of data management does not make any format specifications, but the instruments used for research produce only certain formats, partly because the scholarly community has already agreed upon a standard. This is especially the case in disciplines in which researchers are dependent upon huge research instruments that are shared among many others (e. g., particle physics and astronomy). However, if the diversity of formats in a discipline is not limited (as in approach 3), this could be the case because a standardization has not yet been implemented or is viewed as principally unfeasible. In interdisciplinary research areas, in innovative research areas that have a corresponding need for new formats, or in the case of newly established research archives that have to build up a collection of research data in the first place, very restrictive format requirements could be such an obstruction that standardization is regarded as principally not feasible (approach 4).

Similar to data formats, there are also a large number of metadata formats. Each discipline has its own metadata formats and many are based on XML. In comparing the disciplines, it is noteworthy that particularly in biodiversity and archeology, where researchers are said to demonstrate a lack of awareness for the importance of standards, there are not one at all, but instead a multitude of metadata formats. Indeed, these are fields with a strong interdisciplinary orientation which use a variety of analytical techniques; their description necessitates perhaps a variety of different metadata formats.

The lack of standardization in data formats and metadata formats does not necessarily have serious consequences and is not automatically the indicator of a deficit. As indicated above, standardization can imply restrictions that are principally not or not yet appropriate for some fields. After all, research means exploring new fields that are challenging the standards. For research data curation on the basic level of the *bitstream preservation*, a lack of standards does not present a problem at all. The integrity of the data can be ensured regardless of file formats and metadata formats. The various formats begin to cause difficulties only at levels of technical and content re-usability. Limiting file formats and metadata

formats makes sense in order to reduce the number of technical environments (e. g., hardware and software) and the amount of contextual information which is necessary for data interpretation. Data archives will have to consider both if they wish to support re-usability in a proper way. However, this is not about format obsolescence, which is incorrectly or imprecisely considered to be a fundamental problem in long-term preservation. Modern file and metadata formats rarely become completely obsolete – in the sense that a technical environment and documentation for them can no longer be found. With a certain degree of effort, therefore, data and metadata that would not be found in a discipline with a higher degree of standardization can also be used. The obsolescence of file formats and metadata formats relative to the technical and content requirements of the target groups is far more significant. Principally, the data could be used with old software and emulators, though not according to the requirements of the target groups and therefore the data are useless or inefficient in practice. To ensure the technical and content re-usability of research data in accordance to these requirements, continuous attention has to be paid to both the requirements of the target group and the developments in technology, and corresponding adjustments have to take place. The use of standards is important to efficiently ensure the proper re-usability and to reduce the amount of technology and metadata from the target groups that must be observed and supported.

Corresponding requirements concerning the submission of research data to data centers appear to be explicitly made in rare cases only. There are indeed some format specifications (such as in climate research) and a general awareness of the importance of open formats. However, for many fields these requirements seem to take care of themselves since researchers have to use standard formats anyway as their software and tools are based on them. Some individual cases report on additional measures for quality control, such as plausibility tests or tests of the completeness of the metadata.

Especially large amounts of data generated through mass production tend to be much easier to handle because they are more standardized and homogeneous. The number of records/data objects, which generate more work as logical management units, is more problematic than the byte size of a single record, which is primarily a technical challenge. Moreover,

general statements about the volume of research data can neither be made on a multidisciplinary level nor for individual areas of research. The volume of data can frequently be specified for only one project and one research instrument and can range from tera- to petabyte levels per year.

Nearly all disciplines, with only a few exceptions, state that it is of fundamental importance to keep older research data available for re-use. This is especially true for disciplines in which researchers conduct long-term observations and examine conditions and changes of the environment, in space, or in society. The observational data related to an event and the measured values at a certain point in a time series are not reproducible if lost. In archeology, the situation is similar, although this case does not involve an observation at a certain point in time that cannot be repeated because of changing conditions, but a destructive or manipulating analysis after which the object of investigation may no longer exist in its original form.

The situation can be very different in the case of measured data from laboratory experiments. Researchers in the earth sciences and psycholinguistics reported that some data will quickly become obsolete because the experiments can be reproduced with higher precision thanks to improved measurement technologies. Regarding these and other kinds of reproducible data, the storage costs will have to be compared with the costs of reproducing the data. Such data might be stored only for a limited period of time in order to verify research methodology. However, the reproducibility of experimental measurements using large instruments can likely be of a theoretical nature only. Particle accelerator experiments in physics are repeatable in principle, but if none of the particle accelerators currently operating can perform the experiment, reproducibility is impossible for practical and financial reasons. Therefore, storing data that could, in principle, be reproduced, is important as research questions in particle physics change over time.

According to the survey, research data are stored for different purposes which vary from serving the needs of research groups for internal use only (as in particle physics and medicine) to providing and making data available to groups that can demonstrate legitimate research interest (such as the social sciences and education) to predominantly publicly available data publications (the earth sciences and climate sciences). There are two dis-

tribution channels for providing data. First, the institution or federation that manages the data permanently also makes it available via portals and databases (such as the GESIS research data centers, the World Data Center, and open collaboration/federations). Technically, these are proprietary solutions; standards are not mentioned, apart from the OAI-PMH metadata interface. The second distribution channel, that has been mentioned several times, consists of publishers and data publications.

The target groups and distribution channels are reported to be closely linked to the question of control of the data and data rights. Approximately half of the responses indicated that the re-use of data is subjected to restrictions. The most common reason given for that is that these are sensitive data that are subject to data privacy protection. This is no surprise in medicine, the social sciences, and education because the objects of investigation of these fields are humans. But also disciplines such as biodiversity, which may seem somewhat unexpected, have sensitive data: e. g., the breeding grounds of endangered species. The approaches used for managing sensitive data are restricting the target groups via authentication and authorization mechanisms and use sophisticated anonymization and pseudonymization techniques (especially in the case of the medical field).

In addition to the sensitivity of the data, another main reason for usage restrictions reported by the respondents lies in the fact that data producers have a right to privileged access to the data. Information as a digital commodity has the advantage that it is not depleted by use. In principle, one institution or person generating and maintaining research data is enough to provide the possibility for any number of persons to use the data efficiently. On the one hand, this makes it relatively easy to provide open access, especially since it is part of the statutes of the World Data Centers (in the fields of earth sciences and climate sciences). On the other hand, data producers would have little incentive to invest time and effort in this area if they did not receive any benefits from doing so, and if there was no participation or acknowledgment from other users. For this reason, a number of agreements have been made in the various disciplines. For example, in the earth sciences and biodiversity, rights of first use and moving walls are used, which guarantee data producers the exclusive right of use for a set period of time. In some disciplines, such as psycholinguistics, permission to access must be explicitly obtained from the data producer, depend-

ing on the dataset. However, the universal method with which users recognize the work on the part of data producers remains citation. Efforts to establish the citation of research data therefore represent an important contribution to the proper management of research data, because they can create a powerful incentive to do so.

A key tool in encouraging the citation of research data is the use of persistent identifiers. Even non-persistent identifiers are an important tool for data management. Since identification requires stable and clearly identifiable data objects, a number of important data management topics, such as the precise definition of objects, are already involved in the process of developing a concept for identifiers.[131] Persistent identifiers must also be stable in the long term to permit permanently valid citation, even if the location of the data changes. The various disciplines mention three approaches for persistent identifiers, of which several can be used in one discipline: DOIs, assigned by the DataCite network that is involved especially in setting standards for research data citation (climate science, medicine, education, biodiversity, and the social sciences); handles, which are the technical and syntactical basis of DOIs and are assigned by the EPIC Consortium for research data (humanities and psycholinguistics); and URNs, which are a subset of the URIs used in the internet and which are particularly prevalent in the library field (humanities and climate research).

5.3 Internal organization

Set rules and processes for research data curation have, according to the survey, only partly been established and are often still in development. Particularly in medicine, there are standard approaches in which the handling of data is determined by a variety of legal requirements. In general terms, it can be noted that work flows are well-established in institutions that centralize data management for disciplines or subsets of disciplines

131 See PILIN (2008).

(e. g., the World Data Centers in climatology). This is most likely due to the fact that established processes are a requirement for operating a central organization.

Another requirement for a data archive is funding. Funding is only partly ensured through core funding provided by institutions (like in the social sciences, climatology, earth sciences, and psycholinguistics) and is often still based on a project (humanities, biodiversity, medicine, and particle physics). However, in some cases, these projects are of very long duration, which creates a certain degree of security. Project funds do indeed have an appropriate place in the financial concept of established data archives, where this revenue can be used for short-term tasks, such as the further development of services (psycholinguistics) or the one-time ingest of very large or complex data collections (geosciences and climate research). Few disciplines provide information about the actual costs. For the centralized data management of the priority program "Biodiversity exploratories," staff expenses amount to two and a half full-time employees. These employees are responsible for operation, consulting, and development, but they are not able to perform quality control in terms of content. Psycholinguistics is the only field to provide numbers for monetary costs: nearly one hundred thousand euros per year for equipment and approximately three hundred thousand euros for staff to operate the system and maintain the software. This ratio of technological expenses to staff costs is completely within the normal range, as the literature of the field of data management states that seventy percent or more of the total costs are assigned to staff.[132] The total cost is quite low compared with the usual operating costs for data centers, in the amount of 3.5 million euros per year, as stated by the "Commission on the Future of Information Infrastructure" (KII) in their "General Strategy for Information Infrastructure in Germany."[133]

Considering the high percentage of staff costs in the total costs, it is not surprising that the staff situation is similar to the financial situation. Although there is usually staff who is mainly in charge of research data curation, these people are predominantly paid by project funds and employed

132 See Charles Beagrie Ltd. (2010), p. 14.

133 See Kommission Zukunft der Informationsinfrastruktur (2011), p. B122.

on a temporary basis. The exceptions are the social sciences and psycholinguistics, where permanent staff is employed, at least in part. In all disciplines, staff members usually obtained qualifications in practice and not through professional training.

In general, the area of organization presents a poor impression. Even the larger approaches and developments that deal with organizational aspects, such as the catalogue of criteria for trusted digital archives,[134] the Data Seal of Approval[135] or the nestor Ingest Guide [136], are very rarely mentioned, if at all. A number of studies about cost and funding issues have been published that provide methodological foundations and a variety of case studies. In this context, it seems apparent that the difficulty in clarifying the organizational aspects and the costs of the management of research data does not result from a lack of theoretical and methodological knowledge. This knowledge and technology are more important for the areas of data and metadata. In contrast, the difficulties of organization and costs are the practical realization in each specific situation.

5.4 Perspectives and Visions

It is notable that the importance of research data is heavily emphasized in all disciplines, but there are still many open questions, the least important of which are related to technical matters. The particular challenges with which the disciplines are confronted can be categorized into three groups. The first group of challenges involves communication regarding the importance and usefulness of research data management. Many disciplines are faced with a lack of awareness on the part of individual researchers about the value of archiving and sharing research data (such as the humanities and social sciences), and they wish to improve the re-use of exist-

134 See Online Computer Library Center & the Center for Research Libraries (2007); nestor (2008a).

135 See Data Seal of Approval (2011).

136 See Into the Archive (2009).

ing data collections. Climate research data, for example, could be made available for commercial purposes (such as the tourism industry). In addition to awareness of the importance of data management, the disciplines face several other challenges related to the perception of archives: in order to be perceived as reliable and valuable institutions by the scholarly community, questions dealing with how they handle research data must be clarified and communicated. Are the archives reliable? Are researchers' intellectual achievements and rights taken into account (as in medicine)? Is the necessary personal privacy protection guaranteed and verified as part of certification processes (education)? Lastly, the third group of challenges mentioned relates to the qualitative and quantitative improvement of data collections, such as long-term preservation of processing software, which is necessary for data use (such as in particle physics), or the improvement of metadata standards and data quality (in education, archeology, and biodiversity).

The practical options that are currently available for promoting archiving and re-use of research data can be described in a broader interdisciplinary sense as a greater integration of research data management into research workflows. This ranges from providing support staff in the form of data management specialists for researchers (social sciences) to the technical integration of individual data services such as automatic quality control and data repositories into research workflows (humanities, biodiversity) to the creation of virtual research environments (earth sciences). A more sustainable option and necessity is to integrate the topic of research data curation into scientific training and to create awareness for this issue at that point (climate science, social science, and particle physics).

In light of the heterogeneous situation described here, there is an astonishing unity and clarity about the ultimate model for the management of research data. The establishment of professional competence centers for research data that are responsible for long-term research data curation, developing standards, and providing consulting services for researchers in a centralized or decentralized network is considered optimal (social sciences, humanities, particle physics, classical studies, psycholinguistics, education, and biodiversity). As stated above, it will not be an easy task to determine the characteristics of such centers and to decide how much of this cross-disciplinary ideal of subject-specific centers can be realized

using a multidisciplinary infrastructure. There is no doubt, however, that centers are regarded as the best way to improve the availability and efficient use of research data.

6 Implications and Recommendations on Research Data Curation

Heike Neuroth, Achim Oßwald, and Uwe Schwiegelshohn

On the basis of the comparative survey of approaches to the management of research data in the eleven academic disciplines that have been analysed in the course of this survey, the following results and theses can be formulated. They emphasize the importance of research data curation from a research perspective and refer to conceptual and operational circumstances that should be considered as an initial result and looked at more closely. However, a number of aspects in terms of science and social policy have to be considered as well.

General Issues:

- The importance of research data and its long-term storing and provision is emphasized by all academic disciplines surveyed here.
- The different approaches to research data curation research data curation in these disciplines do not indicate a lack of cooperation across disciplinary boundaries but are a logical consequence of the different requirements and methods practiced within every single discipline.
- Cooperative structures within a discipline are the rule rather than an exception in the field of research data curation.
- Infrastructure facilities such as libraries or data centers are often included as cooperation partners in research data curation. However, their role and current function has not been clearly defined yet.
- In many academic disciplines, researchers are still confronted with a lack of appreciation for the value of long-term archiving and a low acceptance for data sharing and the re-use of data. The awareness of academic disciplines as well as of society and other stakeholders (e.g. libraries, data centers etc.) for the value of data is an important precondition for further discussions and developments.
- Data management, one of the first steps of the actual research data curation, comprises subject-specific as well as generic tasks. The

close cooperation of the various interest groups and stakeholders allows the exact definition of tasks and areas of responsibility.

- It is not possible to make any reliable statements about the data volumes and the number of digital objects that are to be stored and provided, neither for single academic disciplines nor for disciplines in general. All in all, however, a rapid increase of the data volume of digital research data can be recognized across all academic disciplines.

Research Data Centers:

- The processes for ensuring research data curation are already better established in those disciplines in which central structures for data management have emerged than in other academic disciplines.
- Many disciplines consider data centers as an ideal solution for improving and securing the availability and the efficient re-use of research data in the long term. They can be organized centrally or in a decentralized network. They may also play an important role in the development of standards and in providing advice within the relevant academic disciplines.
- There is a need for clarification regarding the reliability of data centers and what criteria have to be met to ensure it. Questions about how to evaluate a data center's trustworthiness (e.g. through the external certification of data centers) and who is responsible for doing so are still open.

Metadata and Formats:

- Nearly every academic discipline uses its own metadata formats. Most of these formats are based on XML. Many academic disciplines have developed subject-specific metadata formats in recent years.
- Research data are available in an almost unmanageable number and variety of data formats. Almost all disciplines share the common trait of using numerous subject-specific and proprietary formats.
- The individual disciplines handle the diversity and heterogeneity of formats very differently. The different formats are either specified by a policy or otherwise restricted, or the choice of format is open or rather cannot be limited because of discipline-related reasons.

- Overall, academic disciplines use open formats wherever they can. However, this can be severely restricted by the given software or hardware. Considering established industrial and commercial processes is helpful when implementing standardization.

Technical Backup of Data:

- The technical backup of data is a first step to research data curation. Through the purely technical storage of research data, the integrity of the data can be preserved, independent of file and metadata formats. However, this does not guarantee the effective re-use of research data.
- Limiting the variety of data and metadata formats reduces the number of technical environments (regarding both hardware and software) necessary for reproducing the data. This makes its re-use easier.
- To ensure the technical and intellectual re-usability of research data, continuous technology watch, the observation of requirements and technical equipment and community watch are needed.

Re-use of Research Data:

- Research data are made available for re-use for various reasons, e.g. for cooperation within research projects, for external researchers or for the general (professional) public upon publication.
- Academic disciplines, their funding agencies and the general public are following the debates about the re-use of research data and the regulations concerning it. The results are insistent requests to make research data accessible and to guarantee their subsequent re-use in the long term.
- Providing and thus encouraging the re-use of research data is mostly prevented for the following reasons: the threat of loss of control over research data, unsolved legal rights issues and conditions of use concerning data, and data protection restrictions. Potential scenarios for re-use are also influenced by the financial effort involved in generating the data.
- The possibility of long-term citation and referencing of research data is one of several motives for research data curation. Therefore, persistent identifiers play an important role.

Costs, Financing, Efficiency and Institutionalization:

- Since research in general is a responsibility of society as a whole, the cost of research should be paid by public funding. In return, society has the right to expect an efficient use of the resources provided. Regarding research data and their subsequent re-use, there are two approaches:
  - Preserving research data after its creation for re-use, or
  - Reproducing or recreating research data. It must be noted that some processes cannot be repeated, e.g. the collection of climate data.
- If both approaches are roughly equal regarding the quality of research data itself, the most cost-efficient approach is to be preferred. For evaluating these approaches and decision-making, very well-informed cost estimates must be available.
- Currently, there are only limited amounts of reliable information about the costs and cost factors involved in research data curation available. In this respect, it is not possible to make any specific statements about cost structures yet. Previous studies indicate that staff costs represent the largest part of the total costs. Up to now, this staff has been paid mainly from project funds.
- (Proportional) financial coverage for research data curation in the form of institutionally-based funding could only be established in some of the academic disciplines studied here. Most disciplines are still using project funds to finance these activities, although some of these projects have extremely long terms.
- There is an urgent need to clarify the costs and cost factors that arise in the context of research data curation. This is the only way to develop and implement sustainable organizational and business models (including financing models) in the different disciplines.
- Securing and maintaining research data is part of scholarly work. Necessary resources for this must be included in cost estimates for research projects.
- The founding of data centers can help to increase the effectiveness of research data curation. This can lead to new organizational structures that extend beyond institutional boundaries.

Qualification:

- There is an urgent need for training in the field of research data curation, especially in theoretical and conceptual areas. Apart from the nestor activities, there are currently few or no systematic training opportunities available in Germany, neither for researchers on a disciplinary level nor for information specialists.
- The exchange of research results relevant to research data curation or examples of best practices occurs only on a limited basis owing to a lack of systematic opportunities for transferring knowledge. Case-by-case decision-making and the focus on one specific discipline and its perceived uniqueness rather than on shared interdisciplinary characterristics have been hindering the establishment of multidisciplinary evaluation criteria and training measures.
- The integration of research data curation into the methodological principles of degree programs or major study courses (such as data librarian and data curator) and research contexts should be a long-term objective. Additional educational opportunities, such as core subjects or interdisciplinary master's programs, are an important source of support for infrastructure activities.

Social Significance:

- Research results are increasingly driving socio-political decisions concerning issues like nuclear power, pre-implantation diagnostics (PID), pandemics, and health risks. It is necessary to preserve the research data on which these decisions are based so that full transparency will be possible in future evaluations.
- The preservation of our cultural heritage is recognized as a social responsibility. Research data are part of this cultural heritage.
- Investigations into violations of scholarly best practices or the detection of methodological errors require that those research data which formed the basis of the specific publication or research paper continue to be available.

All in all, the results and theses presented in this survey demonstrate how significant the (future) role of research data curation is. The EU expert

group "High Level Expert Group on Scientific Data"[137] gave a similar statement: data are infrastructure and serve as a guarantee for innovative research.

Recommendations for future activities that address this topic must be developed on science policy level and implemented through political and funding programs on a national and international level. Some areas of activity have already been identified by researchers and policy-makers.

The above-mentioned "High Level Expert Group on Scientific Data" gave six recommendations[138] which include the establishment of an international framework for the development of collaborative data infrastructures, increased funding for the development of data infrastructures and the development of new approaches and methods to measure and evaluate the value, the importance and the quality of data use. In addition, the importance of training a new generation of "data scientists" is emphasized as well as the establishment of educational opportunities in the new degree programs. The creation of incentive systems in the field of "green technologies" to meet the increasing demand for resources such as energy plays an important role under environmental aspects as well. Lastly, the recommendations suggest the establishment of an international expert panel to promote and manage the development of data infrastructures.

The report from the Commission on the Future of Information Infrastructure (Kommission Zukunft der Informationsinfrastruktur [KII])[139] emphasizes from a national perspective the need for data management plans and data policies as a prerequisite for the exchange and re-use of research data. These plans and policies need to include clear definitions of the responsibilities, the functions and the roles of all stakeholders. Additionally, specific funding programs are recommended for the various aspects in the field of research data curation, making a distinction between development costs for the construction or upgrading of infrastructure and operating costs for permanent operation, including data maintenance.

137 High Level Expert Group on Scientific Data (2010).

138 Ibid.

139 Kommission Zukunft der Informationsinfrastruktur (2011).

In the EU GRDI 2020 report[140], it is assumed that over the next ten years, global research data infrastructures will have to be built in order to operate beyond linguistic, political and social boundaries. These infrastructures are to make research data available and support discovery, access and use. In this context, the model "Digital Ecosystems Science" will be introduced in which the following (new) stakeholders are involved: digital data libraries, digital data archives, digital research libraries and communities of research. This model implies a sometimes entirely new distribution of roles and tasks for the current stakeholders and calls for the creation of newly defined areas of responsibility. The main focus is always on the backup and re-use of research data, allowing the re-use of data also above and beyond disciplinary boundaries. To achieve these objectives, the report formulates eleven recommendations and courses of action which include among others the establishment of new professions and qualification processes. In addition, it is recommended to develop new tools (e.g. in the areas of data analysis or data visualization) and services (e.g. for data integration, for data retrieval or ontology services) for the management and use of data and to take "open science" and "open data" concepts into account.

The present comparative survey of the eleven academic disciplines in Germany confirms the validity of the above-mentioned statements. The overall picture reveals an urgent need for action, especially in the following areas:

- National and international programs have to be initiated to meet the new major challenges in the field of research data.
- A redefinition of roles and responsibilities is necessary to deal with the different areas of activity involved in the accessibility and re-use as well as the long-term preservation of research data.
- New careers and educational opportunities need to be developed and research data management has to be present in (new) degree programs and major study courses to ensure the professional handling of research data.

140 See GRDI 2020 Roadmap Report (2011).

- The publication of research data has to be regarded as an indispensable part of research processes in order to support the verification and further development of research results.

In conclusion, research data are both the result and the indispensable basis of scholarly work. They must be understood as resources that are continuously growing in importance for future generations of researchers as well as across disciplines. In this respect, they are part of the international cultural heritage. Therefore it is needed to curate and maintain research data throughout their entire life cycle.

Although there are already highly promising international approaches and some national developments and discussions have taken place in Germany, it will require a large, nationally coordinated effort before the vision of a "data infrastructure" can become reality. This process will involve both discipline-specific and interdisciplinary aspects and is to be embedded in international efforts. Legal, financial and organizational aspects should not hinder but support these developments. At this point, policy makers should get involved.

References[141]

Alliance of German Science Organisations (2008): Priority Initiative "Digital Information" by the Alliance of German Science Organisations. http://www.allianzinitiative.de/fileadmin/user_upload/keyvisuals/atmos/Allian Initiative_englisch.pdf

Alliance of German Science Organisations (2010b): Priority Initiative "Digital Information" of the Alliance of Science Organizations in Germany. Working Group Research Data: "Principles for the Handling of Research Data". http://www.allianzinitiative.de/fileadmin/user_upload/Home/Principles_Research_Data_2010.pdf

Australian National Data Service (ANDS) (2007): Technical Working Group, Australian Government Department of Education, Science and Training: Towards the Australian Data Commons: A Proposal for an Australian National Data Service. http://ands.org.au/towardstheaustraliandatacommons.pdf

Australian National Data Service (ANDS) (2012–13): Australian National Data Service (ANDS). BUSINESS PLAN 2012–13. http://ands.org.au/resource/ands-business-plan-2012-13-web.pdf

Beagrie, N., Chruszcz, J., & Lavoie, B. (2008): Keeping Research Data Safe: A Cost Model and Guidance for UK Universities. Final Report April 2008. http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf

Beagrie, N., Lavoie, B., & Woollard, M. (2010): Keeping Research Data Safe 2. Final Report April 2010. http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf

BMBF (2011a): Der BMBF-Formularschrank. Übersicht über Fachinformationszentren und überregionale Informationseinrichtungen als Anlage zu den Richtlinien und Hinweisen. Vordrucknummern 0027, 0047, 0067a, 0087. https://foerderportal.bund.de/easy/module/easy_formulare/download.php?datei=163

BMBF (2011b): Der BMBF-Formularschrank. z.B. Richtlinien für Zuwendungsanträge auf Ausgabenbasis. Vordrucknummer 0027. https://foerderportal.bund.de/easy/module/easy_formulare/download.php?datei=135

141 All reference URLs have been checked as of 2013/08/30.

Brase, J. (2010): "Der Digital Object Identifier (DOI)". In: Neuroth, H. et al. (Eds.) (2010): *nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung: Version 2.3*, pp. 9:57–9:65: http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-20100305186

Brumfield, B. (2011): Collaborative Manuscript Transcription. http://manuscript-transcription.blogspot.com/2011/02/2010-year-of-crowdsourcing.html

Bundesministerium für Bildung und Forschung, BMBF (2007): Das 7. EU-Forschungsrahmenprogramm. http://www.forschungsrahmenprogramm.de/_media/7-EU_FRP.pdf

Charles Beagrie Ltd & JISC (2010): Keeping Research Data Safe Factsheet: Cost Issues in Digital Preservation of Research Data. http://www.beagrie.com/KRDS_Factsheet_0910.pdf

Charles Beagrie Ltd. (2010): User Guide for Keeping Research Data Safe: Assessing Costs/Benefits of Research Data Management, Preservation and Re-Use. Version 1.0 – December 2010. http://www.beagrie.com/KeepingResearchDataSafe_UserGuide_v1_Dec2010.pdf

DFG (1998): Proposals for Safeguarding Good Scientific Practice. Recommendations of the Commission on Professional Self-Regulation in Science. http://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/self_regulation_98.pdf

DFG (2006): DFG-Positionspapier: Wissenschaftliche Literaturversorgungs und Informationssysteme – Schwerpunkte der Förderung bis 2015. http://www.dfg.de/download/pdf/foerderung/programme/lis/positionspapier.pdf

DFG (2009a): DFG-Merkblatt 60.06. Service-Projekte zu Informationsmanagement und Informationsinfrastruktur in Sonderforschungsbereichen INF.

DFG (2009b): Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten, Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme, Unterausschuss für Informationsmanagement. http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf

DFG (2010a): DFG-Ausschreibung „Erschließung und Digitalisierung von objektbezogenen wissenschaftlichen Sammlungen". http://www.dfg.de/download/pdf/foerderung/programme/lis/ausschreibung_ed_objekte.pdf

DFG (2010b): DFG-Ausschreibung „Informationsinfrastrukturen für Forschungsdaten". http://www.dfg.de/download/pdf/foerderung/programme/lis/ausschreibung_forschungsdaten_1001.pdf

DFG (2010c): Merkblatt für Anträge auf Sachbeihilfen mit Leitfaden für die Antragstellung und ergänzenden Leitfaden für die Antragstellung für Projekte mit

Verwertungspotenzial, für die Antragstellung für Projekte im Rahmen einer Kooperation mit Entwicklungsländern, Deutsche Forschungsgemeinschaft (DFG), Bonn (DFG-Vordruck 1.02 – 8/10). http://www.dfg.de/foerderung/programme/einzelfoerderung/sachbeihilfe/formulare_merkblaetter/index.jsp

D-Grid GmbH (2011): Die D-Grid Initiative im vierten Jahr: Zwischenbilanz und Vorstellung der Projekte. http://www.d-grid-gmbh.de/index.php?id=51

Digital Curation Centre (DCC) (2011a): DCC Curation Lifecycle Model. http://www.dcc.ac.uk/resources/curation-lifecycle-model

Digital Curation Centre (DCC) (2011b): What is Digital Curation? http://www.dcc.ac.uk/digital-curation/what-digital-curation

ESSD (2011): Earth System Science Data: The Data Publishing Journal. http://www.earth-system-science-data.net/

European Commission (2011): APARSEN: Alliance Permanent Access to the Records of Science in Europe Network. http://cordis.europa.eu/fetch?CALLER=PROJ_ICT&ACTION=D&CAT=PROJ&RCN=97472

Feijen, M. (2011): What Researchers want. SURF Foundation. http://www.surf.nl/nl/publicaties/Documents/What_researchers_want.pdf

Foucault, M. (2006): *The Order of Things: An Archaeology of the Human Sciences*. London: Routledge

Funk, S. (2010a): "Emulation". In: Neuroth, H. et al. (Eds.): *nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, Version 2.3*, pp. 8:16–8:23. http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-20100305134

Funk, S. (2010b): "Migration". In: Neuroth, H. et al. (Eds.): *nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, Version 2.3*, pp. 8:10–8:15. http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-20100617189

Gemeinsame Wissenschaftskonferenz des Bundes und der Länder. http://www.gwk-bonn.de/index.php?id=126

German Rectors' Conference (HRK) (1996): Moderne Informations- und Kommunikationstechnologien („Neue Medien") in der Hochschullehre. Empfehlung des 179. Plenums 1996. http://www.hrk.de/positionen/gesamtliste-beschluesse/position/convention/moderne-informations-und-kommunikationstechnologien-neue-medien-in-der-hochschullehre/ (in German only)

GRDI 2020 Roadmap Report (2011): Global Scientific Data Infrastructures, The Big Data Challenges – Short Version, September 2011.

http://www.grdi2020.eu/Repository/FileScaricati/fc14b1f7-b8a3-41f8-9e1e-fd803d28ba76.pdf

Hammerschmitt, M. (2002): Die Uhr läuft, Die NASA hat Probleme mit den Innovationszyklen in der IT-Industrie. *Telepolis*. http://www.heise.de/tp/artikel/12/12538/1.html

Heinen, N. (2010): Datenanalyse entlarvt „Schummelkultur“ in medizinischen Studien. *Heise Online*. http://www.heise.de/newsticker/meldung/Datenanalyse-entlarvt-Schummelkultur-in-medizinischen-Studien-1158102.html (30.08.2013)

High Level Expert Group on Scientific Data (2010): Riding the Wave: How Europe Can Gain from the Rising Tide of Scientific Data. Final report of the High Level Expert Group on Scientific Data. A Submission to the European Commission. October 2010. http://ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=6204 (30.08.2013)

Horlings, E., Ligtvoet, A., Cave, J. Frinking, E., Mildt, C., Shergold, M., & Kahan, J. (2006): Markets of Virtual Science, Report on the Economics and Policy Implications of an Emerging Scientific Phenomenon. PM-1976-BMBF. Prepared for the German Bundesministerium für Bildung und Forschung (BMBF), RAND Europe. http://www.innovationsundtechnikanalysen.de/projekte/abgeschlossene-projekte/chancen-und-risiken-virtualisierter-wissenschaft/maerkte-der-virtualisierten-wissenschaft/at_download/ita_projektbericht

Into the Archive – a Guide to the Information Transfer to a Digital Repositiory (2009): (Draft for public comment). http://files.d-nb.de/nestor/materialien/nestor_mat_10_en.pdf

Kennedy, D. & Alberts, B. (2008): Editorial Expression of Concern. In: *Science*, 319 (5868), 1335. http://dx.doi.org/10.1126/science.1157223

Knowledge Exchange (2011a): Research Data. http://www.knowledge-exchange.info/Default.aspx?ID=284

Knowledge Exchange (2011b): Research Data Working Group. http://www.knowledge-exchange.info/Default.aspx?ID=285

Kommission Zukunft der Informationsinfrastruktur (KII) (2011): Master plan for Germany’s information infrastructure approved; Recommendations by the Commission on the Future of Information Infrastructure; commissioned by the Joint Science Conference of the federal state and the German Länder. http://www.leibniz-gemeinschaft.de/en/infrastrukturen/kii/ (PDF in German only)

Liegmann, H.; Neuroth, H. (2010): “Einführung”. In: Neuroth, Heike et al. (Eds.): *nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivie-*

rung: Version 2.3, 2010, pp. 1:1–1:10. http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-2010030508

Michel, J. B. et al. (2011): "Quantitative Analysis of Culture Using Millions of Digitized Books." In: *Science* 331 (6014), pp. 176–182. http://dx.doi.org/10.1126/science.1199644

nestor (2008a): Arbeitsgruppe Vertrauenswürdige Archive – Zertifizierung: Kriterienkatalog vertrauenswürdige digitale Langzeitarchive. Version 2. Frankfurt am Main. http://files.d-nb.de/nestor/materialien/nestor_mat_08.pdf

Neuroth, Heike, Oßwald, Achim, Scheffel, Regine, Strathmann, Stefan, & Jehn, Mathias (Eds.) (2009): *nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung: Version 2.0*. Boizenburg: Hülsbusch

Neuroth, Heike, Oßwald, Achim, Scheffel, Regine, Strathmann, Stefan, & Huth, Karsten (Eds.) (2010): *nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung: Version 2.3*. http://nestor.sub.uni-goettingen.de/handbuch/nestor-handbuch_23.pdf

Neuroth, Heike, Strathmann, Stefan, Oßwald, Achim, Scheffel, Regine, Klump, Jens, & Ludwig, Jens (2012): *Langzeitarchivierung von Forschungsdaten – eine Bestandsaufnahme*. Boizenburg: Hülsbusch. http://nestor.sub.uni-goettingen.de/bestandsaufnahme

Niggemann, E., De Decker, J., & Levy, M. (2011): The New Renaissance: Report of the Comite des Sages. Reflection Group on Bringing Europe's Cultural Heritage Online. Luxembourg: Publications Office of the European Union. http://ec.europa.eu/information_society/activities/digital_libraries/doc/refgroup/final_report_cds.pdf

NSF Data Management for NSF SBE Directorate Proposals and Awards (2010): http://www.nsf.gov/sbe/SBE_DataMgmtPlanPolicy.pdf

NSF Data Management Plan (2011): NSF Data Management Plan Requirements. http://www.nsf.gov/bfa/dias/policy/dmp.jsp

OAIS (2009): Reference Model for an Open Archival Information System (OAIS). Draft Recommended Standard, from the Consultative Committee for Space Data Systems (CCSDS). CCSDS 650.0-P-1.1. Pink Book: August 2009. http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf

OAIS (2010): "Das Referenzmodell OAIS – Open Archival Information System". In: Neuroth, H. et al. (Eds.) (2010): *nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung: Version 2.3*, pp. 4:1–4:16. http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-2010061757

Online Computer Library Center (OCLC) & the Center for Research Libraries (CRL) (2007): Trustworthy Repositories Audit & Certification: Criteria and Checklist. Version 1.0. February 2007. Chicago: Center for Research Libraries; Dublin, Ohio. http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

Organisation for Economic Co-Operation and Development (OECD) (2004): Science, Technology and Innovation for the 21st Century. Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29–30 January 2004 – Final Communique. http://www.oecd.org/document/1/0,3343,en_2649_201185_25998799_1_1_1_1,00.html

Organisation for Economic Co-Operation and Development (OECD) (2007): OECD Principles and Guidelines for Access to Research Data from Public Funding. http://www.oecd.org/document/2/0,3746,en_2649_34293_38500791_1_1_1_1,00.html

PILIN (2008): Information Modeling Guide for Identifiers in e-research, University of Southern Queensland. http://www.linkaffiliates.net.au/pilin2/files/infomodellingeresearch.pdf

Rosenthal, D. (2011): Are We Facing a "Digital Dark Age?" [Blog post]. http://blog.dshr.org/2011/02/are-we-facing-digital-dark-age.html

Schmundt, H. von (2000): Im Dschungel der Formate. In: *Spiegel* 26 (2000). http://www.spiegel.de/spiegel/print/d-16748341.html

Schöning-Walter, C. (2010): "Der Uniform Resource Name (URN)". In: Neuroth, H. et al. (Eds.) (2010): *nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, Version 2.3*, pp. 9:46–9:56. http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-20100305176

Schwens, U. & Liegmann, H. (2004): "Langzeitarchivierung digitaler Ressourcen." In: Kuhlen, R., Seeger T., & Strauch, D. (Eds.): *Grundlagen der praktischen Information und Dokumentation. Handbuch zur Einführung in die Informationswissenschaft und -praxis* (pp. 567–570). Vol. 1. 5th Ed., München: Saur.

Spiegel Online (2007, December 17): Computer-Panne. Japan sucht die Rentendaten. http://www.spiegel.de/netzwelt/tech/0,1518,523769,00.html

TELOTA (2011): TELOTA – The Electronic Life of the Academy. http://www.bbaw.de/telota/

Thibodeau, K. (2002): Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years. Council on Library and Information Resources (CLIR). http://www.clir.org/pubs/reports/pub107/thibodeau.html

Ullrich, D. (2010): "Bitstream-Preservation". In: Neuroth, H. et al. (Eds.): *nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, Version 2.3*, pp. 8:3–8:9. http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-20100305123

UNESCO (2003): Charter on the Preservation of Digital Heritage. http://portal.unesco.org/en/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html

Website "Berliner Erklärung über offenen Zugang zu wissenschaftlichem Wissen" (2006): In: *Wikipedia, Die freie Enzyklopädie*. Bearbeitungsstand: 5. Juli 2013, 09:32 UTC. http://de.wikipedia.org/w/index.php?title=Berliner_Erkl%C3%A4rung_%C3%BCber_offenen_Zugang_zu_wissenschaftlichem_Wissen&oldid=120239872

Website "Citizen Science" (2013): In: *Wikipedia, Die freie Enzyklopädie*. Bearbeitungsstand: 7. August 2013, 09:11 UTC. http://de.wikipedia.org/w/index.php?title=Citizen_Science&oldid=121289680

Website "Crowdsourcing" (2013): In: *Wikipedia, Die freie Enzyklopädie*. Bearbeitungsstand: 11. Juli 2013, 16:46 UTC. http://de.wikipedia.org/w/index.php?title=Crowdsourcing&oldid=120449843

Wellcome Trust (2011a): Sharing Research Data to Improve Public Health. Full Joint Statement by Funders of Health Research. http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTDV030690.htm

Wellcome Trust (2011b): Sharing Research Data to Improve Public Health. Joint Statement of Purpose: http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTDV030689.htm

Wissenschaftsrat (2011a): Der Wissenschaftsrat: Empfehlungen zu wissenschaftlichen Sammlungen als Forschungsinfrastrukturen. http://www.wissenschaftsrat.de/download/archiv/10464-11.pdf (in German only)

Wissenschaftsrat (2011b): Der Wissenschaftsrat: Übergreifende Empfehlungen zu Informationsinfrastrukturen, Drs. 10466-11, Berlin (28.01.2011), Berlin. http://www.wissenschaftsrat.de/download/archiv/10466-11.pdf (in German only)

Abbrevations

ANDS	Australian National Data Service
APARSEN	Alliance Permanent Access to the Records of Science in Europe Network
ARDC	Australian Research Data Commons
AWBI	DFG Committee on Scientific Libraries and Information Systems (Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme)
BBAW	Berlin-Brandenburg Academy of Sciences and Humanities (Berlin-Brandenburgische Akademie der Wissenschaften)
BMBF	German Federal Ministry for Education and Research (Bundesministerium für Bildung und Forschung)
CRL	Center for Research Libraries
DCC	Digital Curation Centre
DEFF	Denmark's Electronic Research Library
DFG	German Research Foundation (Deutsche Forschungsgemeinschaft)
D-GRID	German Grid Initiative
DIMDI	German Institute for Medical Documentation and Information (Deutsches Institut für Medizinische Dokumentation und Information)
DIPF	German Institute for International Educational Research (Deutsche Institut für Internationale Pädagogische Forschung)
DKRZ	German Climate Computing Center (Deutsches Klimarechenzentrum)
DMP	Data Management Plan
DNB	German National Library (Deutsche Nationalbibliothek)
DOI	Digital Object Identifier
EIF	Education Investment Fund

EPIC	European Persistent Identifier Consortium
ESFRI	European Strategy Forum on Research Infrastructures
ESSD	Earth System Science Data
EU	European Union
GESIS	Leibniz Institute for the Social Sciences (Leibniz-Institut für Sozialwissenschaften)
GWDG	Göttingen Society for Scientific Data Processing (Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen [GWDG])
GWK	Joint Science Conference (Gemeinsame Wissenschaftskonferenz des Bundes und der Länder)
HGF	Helmholtz Foundation
HRK	German Rectors' Conference (Hochschulrektorenkonferenz)
ISCU	International Council for Science
JISC	Joint Information Systems Committee
KE	Knowledge Exchange
KII	Commission on the Future of Information Infrastructure (Kommission Zukunft der Informationsinfrastruktur)
KRDS	Keeping Research Data Safe
LHC	Large Hadron Collider
MPDL	Max Planck Digital Library
MPI PL	Max Planck Institute for Psycholinguistics (Max-Planck-Institut für Psycholinguistik)
NASA	National Aeronautics and Space Administration
NCRIS	National Collaborative Research Infrastructure Strategy
NSF	National Science Foundation
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OAIS	Open Archival Information System
OCLC	Online Computer Library Center
OECD	Organisation for Economic Co-Operation and Development

PI	Persisent Identifier
PID	Preimplantation Diagnostics
RDA	Research Data Alliance
SCICOLL	Scientific Collections International
SFB	DFG Collaborative Research Centres (Sonderforschungsbereiche)
SUB	Göttingen State and University Library (Niedersächsische Staats- und Universitätsbibliothek Göttingen)
SURF	SURFfoundation
TELOTA	The Electronic Life of the Academy
TIB	German National Library of Science and Technology (Technische Informationsbibliothek Hannover)
TU	Technical University Dortmund
UNESCO	United Nations Educational, Scientific and Cultural Organization
URI	Uniform Resource Identifier
URN	Uniform Resource Name
USA	United States of America
WDC-RSAT	World Data Center for Remote Sensing of the Atmosphere
WHO	World Health Organization
WR	German Council of Science and Humanities (Wissenschaftsrat)
ZBMed	German National Library of Medicine (Deutsche Zentralbibliothek für Medizin)
ZBW	Leibniz Information Center for Economics (Deutsche Zentralbibliothek für Wirtschaftswissenschaften)
ZPID	Leibniz Centre for Psychological Information and Documentation (Leibniz-Zentrum für Psychologische Information und Dokumentation)

Directory of Authors

Ludwig, Jens
Niedersächsische Staats- und Universitätsbibliothek Göttingen
http://rdd.sub.uni-goettingen.de
ludwig@sub.uni-goettingen.de

Neuroth, Dr. Heike
Niedersächsische Staats- und Universitätsbibliothek Göttingen
http://rdd.sub.uni-goettingen.de
neuroth@sub.uni-goettingen.de

Oßwald, Prof. Achim
Fachhochschule Köln
http://www.fbi.fh-koeln.de/en-index.htm
achim.osswald@fh-koeln.de

Scheffel, Prof. Regine
Hochschule für Technik, Wirtschaft und Kultur Leipzig
http://www.htwk-leipzig.de/en/about-us/faculties/media/
scheffel@fbm.htwk-leipzig.de

Schwiegelshohn, Prof. Uwe
TU Dortmund
http://www.irf.tu-dortmund.de/cms/en/Institute/index.html
uwe.schwiegelshohn@udo.edu

Strathmann, Stefan
Niedersächsische Staats- und Universitätsbibliothek Göttingen
http://rdd.sub.uni-goettingen.de
strathmann@sub.uni-goettingen.de

Winkler-Nees, Dr. Stefan
Deutsche Forschungsgemeinschaft (DFG)
http://www.dfg.de/en/
stefan.winkler-nees@dfg.de

Weitere ausgewählte Titel aus dem vwh-Verlag

nestor Handbuch Eine kleine Enzyklopädie der digitalen Langzeitarchivierung
Version 2.0, hrsg. v. Heike Neuroth, Achim Oßwald, Regine Scheffel, Stefan Strathmann, Mathias Jehn im Rahmen des Kooperationsverbundes nestor
2009, ISBN 978-3-940317-48-3, 620 S., 24,90 € (D), 25,60 € (A), 37,90 CHF

Pressestimmen: VÖB-Mitteilungen 62 (2009), Heft 4, S. 80–81: *In 19 Kapiteln wird der Leserin/dem Leser das Feld der digitalen Langzeitarchivierung von organisatorischen Fragen, Workflows, Metadaten und Erhaltungsstrategien bis hin zu praktischen Maßnahmen Formate, Kosten und Zugriffsmodelle betreffend aufbereitet. Besonders gelungen der Bogen von der Theorie zur Praxis.*

Rundbrief Fotografie 16 (2009), Heft 4, S. 30–32: *Die Sammlung an Kompetenz ist beachtenswert, das Thema dringlich. Man kann sich nur wünschen, dass mit der Herausgabe dieses Handbuches auch in der Öffentlichkeit das Bewusstsein für die Bedeutung der Bewahrung des digitalen kulturellen Erbes gefördert wird. Für direkt Betroffene und Fachleute ist es eine unerlässliche Arbeitshilfe […].*

Langzeitarchivierung von Forschungsdaten Eine Bestandsaufnahme
hrsg. v. Heike Neuroth, Stefan Strathmann, Achim Oßwald, Regine Scheffel, Jens Klump, Jens Ludwig im Rahmen des Kooperationsverbundes nestor
2012, ISBN 978-3-86488-008-7, 382 S., 29,90 € (D), 30,74 € (A), 36,90 CHF

Forschungsdaten und der langfristige Zugriff auf sie sind für Wissenschaftler aller Disziplinen von großer Bedeutung: als Nachweis des Forschungsprozesses und seiner Ergebnisse, zur Nachnutzung im Rahmen inner- und interdisziplinärer Kooperationen oder im Hinblick auf vergleichende Analysen, bei denen mit neuen Methoden oder unter veränderten Rahmenbedingungen geforscht wird. Dies alles ist nur möglich, wenn Forschungsdaten gesichert, für eine langfristige Nachnutzung archiviert und zur Verfügung gestellt werden. Angesichts rasant anwachsender digitaler Datenmengen ist die Langzeitarchivierung von Forschungsdaten für alle Wissenschaftsdisziplinen eine begleitende Infrastrukturaufgabe. In dieser Untersuchung haben WissenschaftlerInnen aus elf Fachdisziplinen – Geisteswissenschaften, Sozialwissenschaften, Psycholinguistik, Pädagogik, Altertumswissenschaft, Geowissenschaft, Klimaforschung, Biodiversität, Teilchenphysik, Astronomie und Medizin – systematisch den Stand im Umgang mit der Langzeitarchivierung von Forschungsdaten in ihrer jeweiligen Disziplin aufgearbeitet.

Leitfaden zum Forschungsdaten-Management Handreichungen aus dem WissGrid-Projekt hrsg. von Jens Ludwig und Harry Enke
2013, ISBN 978-3-86488-032-2, 120 S., 15,80 € (D), 16,24 € (A), 19,50 CHF

Digitale Forschungsdaten sind eine unverzichtbare Grundlage moderner Wissenschaft. Mit ihnen sind eine Reihe von Datenmanagement-Fragen verbunden: Wie lange sollen die Daten aufbewahrt werden? Welche Kontextinformationen müssen erfasst werden, um die Daten zu einem späteren Zeitpunkt noch sinnvoll benutzen zu können? Wie viel kostet die Aufbewahrung? Das Buch stellt mit einem Leitfaden und einer Checkliste einfach handhabbare Instrumente bereit, um die wichtigsten Aufgaben und Fragen im Forschungsdaten-Management strukturiert beantworten und effizient planen zu können.

H. Kohle: Digitale Bildwissenschaft
2013, 16,80 €, ISBN 978-3-86488-036-0

U. Höbarth: Konstruktivistisches Lernen mit Moodle Praktische Einsatzmöglichkeiten in Bildungsinstitutionen
2013, 31,50 €, ISBN 978-3-86488-033-9

E. Blaschitz et al. (Hg.): Zukunft des Lernens Wie digitale Medien Schule, Aus- und Weiterbildung verändern
2012, 23,50 €, ISBN 978-3-86488-028-5

C. Lehr: Web 2.0 in der universitären Lehre Ein Handlungsrahmen für die Gestaltung technologiegestützter Lernszenarien 2012, 27,90 €, 978-3-86488-024-7

J. Wagner/V. Heckmann (Hg.): Web 2.0 im Fremdsprachenunterricht
2012, 27,50 €, ISBN 978-3-86488-022-3

J.-F. Schrape: Wiederkehrende Erwartungen Visionen, Prognosen und Mythen um neue Medien seit 1970
2012, 11,90 €, ISBN 978-3-86488-021-6

S. Brugner: Über die Realität im Zeitalter digitaler Fotografie
2012, 23,90 €, ISBN 978-3-86488-018-6

S. Felzmann: Playing Yesterday Mediennostalgie im Computerspiel
2012, 22,50 €, ISBN 978-3-86488-015-5

C. Carstens: Ontology Based Query Expansion Retrieval Support for the Domain of Educational Research
2012, 34,90 €, ISBN 978-3-86488-011-7

D. Appel (Hg.): Welt|Kriegs|Shooter Computerspiele als realistische Erinnerungsmedien? 2012, 28,50 €, 978-3-86488-010-0

R. Sonnberger: Facebook im Kontext medialer Umbrüche
2012, 29,50 €, ISBN 978-3-86488-009-4

M. Janneck/C. Adelberger: Komplexe Software-Einführungsprozesse gestalten: Grundlagen und Methoden Am Beispiel eines Campus-Management-Systems 2012, 26,90 €, 978-3-940317-63-6

M. Görtz: Social Software as a Source of Information in the Workplace
2011, 31,90 €, ISBN 978-3-86488-006-3

G. Franz: Die vielen Wikipedias Vielsprachigkeit als Zugang zu einer globalisierten Online-Welt
2011, 27,50 €, ISBN 978-3-86488-002-5

B. Blaha: Von Riesen und Zwergen Zum Strukturwandel im verbreitenden Buchhandel in Deutschland und Österreich
2011, 24,90 €, ISBN 978-3-940317-93-3

J.-F. Schrape: Gutenberg-Galaxis Reloaded? Der Wandel des deutschen Buchhandels durch Internet, E-Books und Mobile Devices
2011, 17,90 €, ISBN 978-3-940317-85-8

W. Drucker: Von Sputnik zu Google Earth Über den Perspektivenwechsel hin zu einer ökologischen Weltsicht
2011, 25,90 €, ISBN 978-3-940317-82-7

K. Huemer: Die Zukunft des Buchmarktes Verlage und Buchhandlungen im digitalen Zeitalter
2010, 24,90 €, ISBN 978-3-940317-73-5

R. Bauer: Die digitale Bibliothek von Babel Über den Umgang mit Wissensressourcen im Web 2.0
2010, 26,90 €, ISBN 978-3-940317-71-1

H. Frohner: Social Tagging
2010, 26,90 €, ISBN 978-3-940317-03-2

C. Russ: Online Crowds Massenphänomene und kollektives Verhalten im Internet 2010, 31,50 €, 978-3-940317-67-4

S. Sobczak/M. Groß: Crowdsourcing
2010, 24,90 €, ISBN 978-3-940317-61-2

T. Memmel: User Interface Specification for Interactive Software Systems 2009, 33,90 €, 978-3-940317-53-7

M. Maßun: Collaborative Information Management in Enterprises
2009, 28,90 €, ISBN 978-3-940317-49-0

J. Sieck/M. A. Herzog (Hg.): Kultur und Informatik: Serious Games
2009, 30,90 €, ISBN 978-3-940317-47-6

S. Mühlbacher: Information Literacy in Enterprises
2009, 32,90 €, ISBN 978-3-940317-45-2

M. Heckner: Tagging, Rating, Posting
2009, 27,90 €, ISBN 978-3-940317-39-1

vwh Aktuelle Ankündigungen, Inhaltsverzeichnisse und Rezensionen finden sie im vwh-Blog unter www.vwh-verlag.de.
Das komplette Verlagsprogramm mit Buchbeschreibungen sowie eine direkte Bestellmöglichkeit im vwh-Shop finden Sie unter www.vwh-verlag-shop.de.

The relevance of research data today and for the future is well documented and discussed, in Germany as well as internationally. Ensuring that research data are accessible, sharable, and re-usable over time is increasingly becoming an essential task for researchers and research infrastructure institutions. Some reasons for this development include the following:

- research data are documented and could therefore be validated

- research data could be the basis for new research questions

- research data could be re-analyzed by using innovative digital methods

- research data could be used by other disciplines

Therefore, it is essential that research data are curated, which means they are kept accessible and interpretable over time.

In Germany, a baseline study was undertaken analyzing the situation in eleven research disciplines in 2012. The results were then published in a German-language edition.

To address an international audience, the German-language edition of the study has been translated and abridged. The present publication provides a summary of the preconditions, results, and conclusions of the baseline study "Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme".

The editors of this edition hope that this publication will contribute to further discussion and increased awareness of this topic, resulting in expanded measures to improve the framework and the conditions for the digital curation of research data.

vwh Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

www.vwh-verlag.de

in collaboration with

Universitätsverlag Göttingen

12,80 € (D)
13,16 € (A)
16,80 CHF

ISBN: 978-3-86488-054-4

9 783864 880544