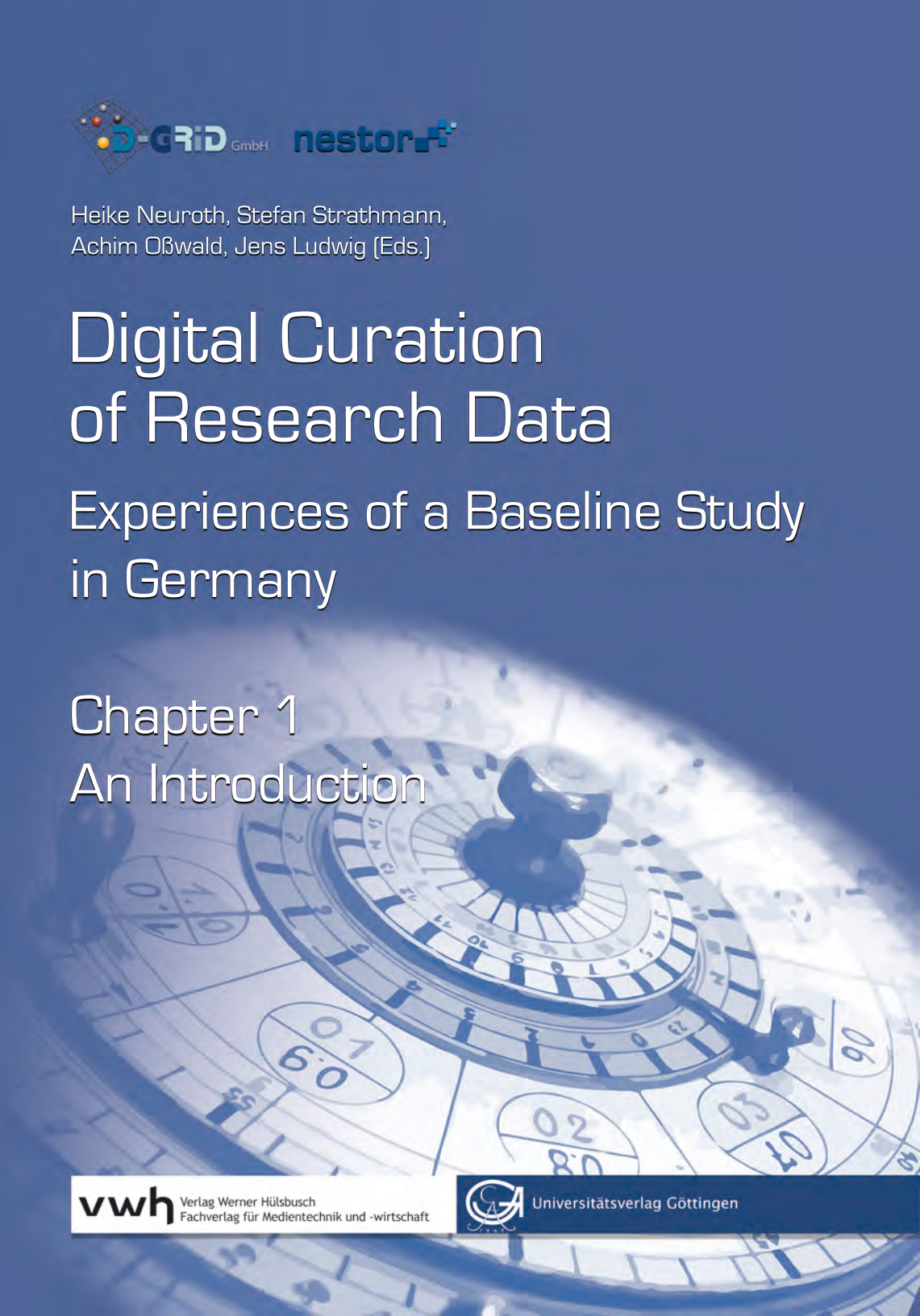
Heike Neuroth, Stefan Strathmann,
Achim Oßwald, Jens Ludwig (Eds.)

Digital Curation of Research Data

Experiences of a Baseline Study in Germany

Chapter 1
An Introduction



vwh Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

Universitätsverlag Göttingen

**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

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.