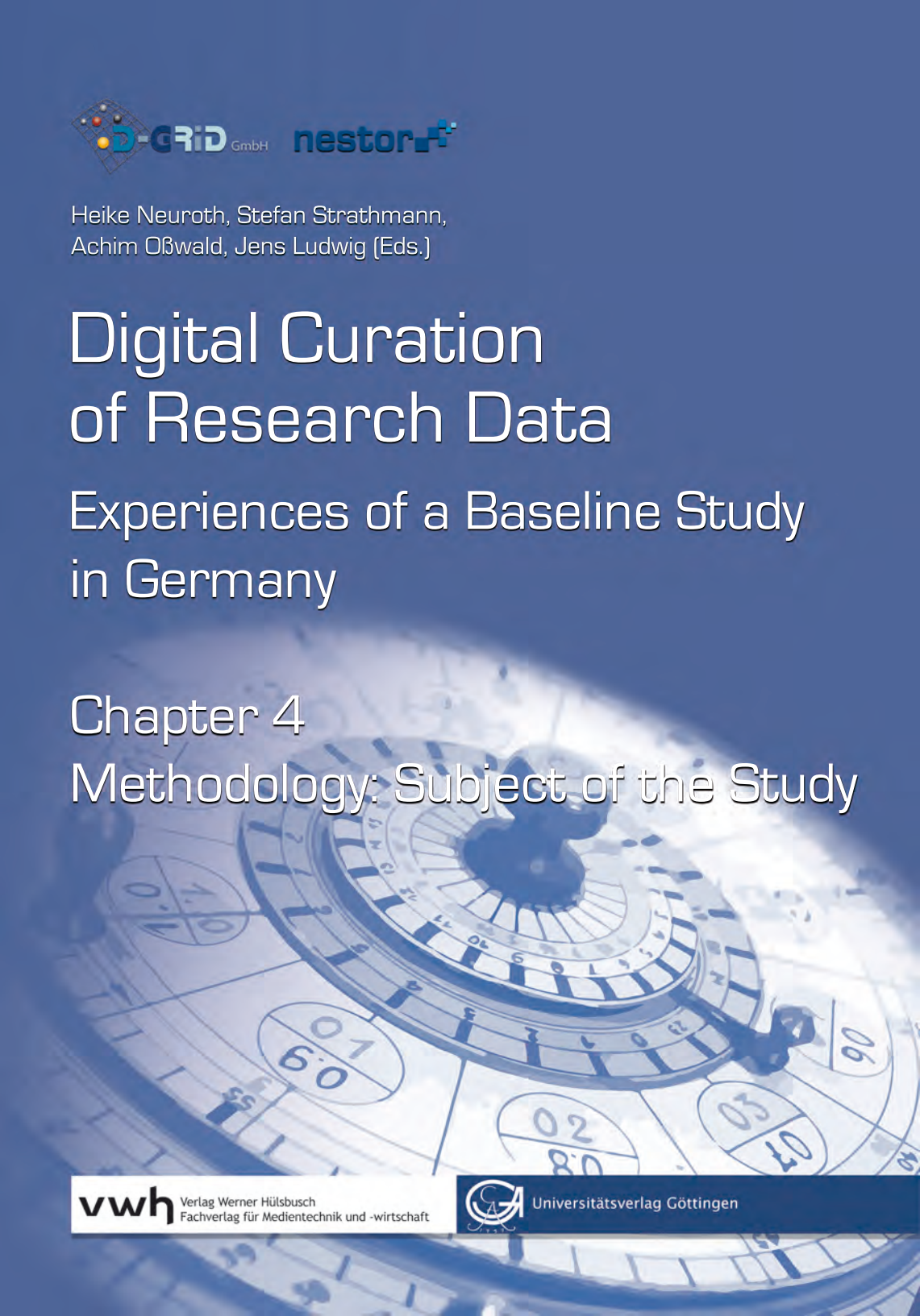
Heike Neuroth, Stefan Strathmann,
Achim Oßwald, Jens Ludwig (Eds.)

Digital Curation of Research Data

Experiences of a Baseline Study in Germany

Chapter 4
Methodology: Subject of the Study



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

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.

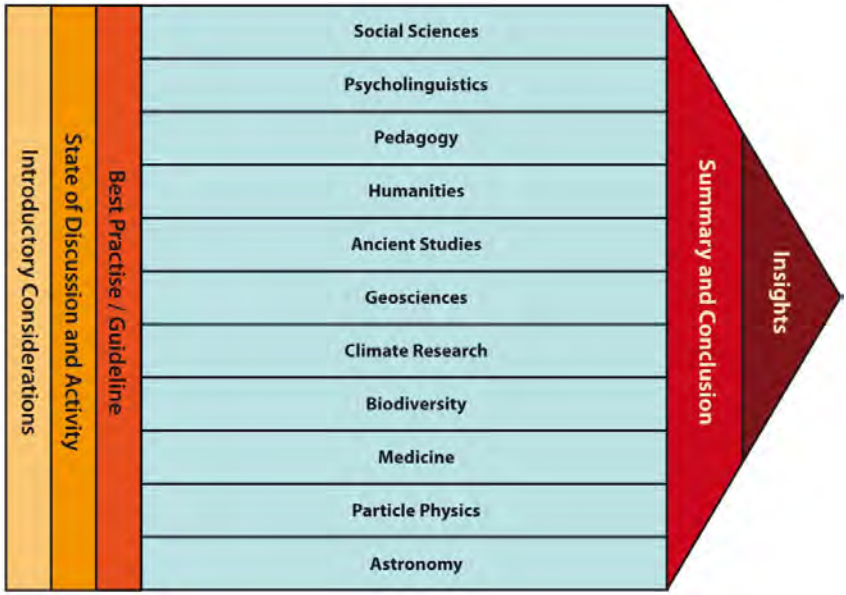


Figure 1: Structure of this volume

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)?