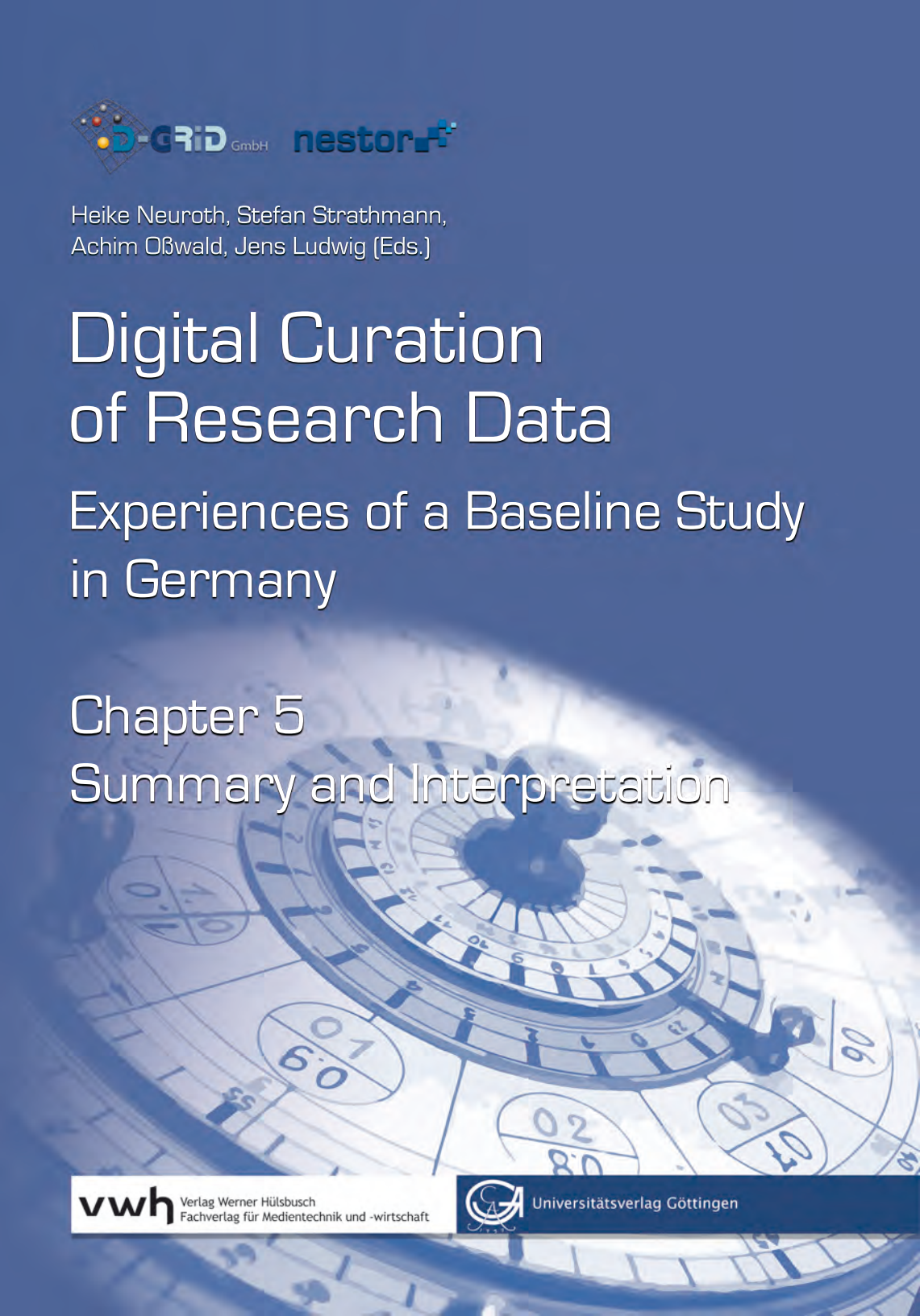
Heike Neuroth, Stefan Strathmann,
Achim Oßwald, Jens Ludwig (Eds.)

Digital Curation of Research Data

Experiences of a Baseline Study in Germany

Chapter 5
Summary and Interpretation



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

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.