Markus Schmalzl[1]

RESEARCH DATA MANAGEMENT - A COMMON TASK FOR ARCHIVES AND DATA PRODUCERS

Abstract

Purpose: *The purpose of the research was to examine to what extent public archives are affected by the archiving of research data and which challenges and opportunities exist here.*

Method/Approach: *The approach taken in this research builds on the evaluation of different research data management initiatives and cooperation projects of archives for archiving research data.*

Results: *Research data management is often only designed for the publication of the data and short subsequent periods of long-term storage of around 10 years. However, various branches of science require research data over much longer periods of time. To make this possible, close cooperation between data producers, research data managers and archivists is necessary*

Conclusions/findings: *Many public archives are already actively involved in archiving research data and have extensive skills for this and for the FAIR management of this data. The goal of sustainable content preservation can only be achieved from a practical point of view if data producers, data managers and responsible archives cooperate as early as possible in the data life cycle and ideally at an institutional level. Large-scale initiatives such as the National Research Data Infrastructure (NFDI) also enable the necessary standards and best practices to be jointly developed with the help of competence networks in the spirit of open archival science and to support the necessary cultural change towards sustainable data management.*

Keywords: *research data management, long term archiving, content preservation*

[1] Markus Schmalzl, Dr. phil., Staatsarchiv München, e-mail: Markus.Schmalzl@stam.bayern.de.

GESTIONE DEI DATI DI RICERCA: UN COMPITO COMUNE PER ARCHIVI E PRODUTTORI DI DATI

Abstract

Scopo: *lo scopo della ricerca era esaminare in che misura gli archivi pubblici sono interessati dall'archiviazione dei dati di ricerca e quali sfide e opportunità esistono in questo caso.*

Metodo/approccio: *l'approccio adottato in questa ricerca si basa sulla valutazione di diverse iniziative di gestione dei dati di ricerca e progetti di cooperazione di archivi per l'archiviazione dei dati di ricerca.*

Risultati: *la gestione dei dati di ricerca è spesso progettata solo per la pubblicazione dei dati e brevi periodi successivi di archiviazione a lungo termine di circa 10 anni. Tuttavia, vari rami della scienza richiedono dati di ricerca per periodi di tempo molto più lunghi. Per rendere ciò possibile, è necessaria una stretta cooperazione tra produttori di dati, gestori dei dati di ricerca e archivisti.*

Conclusioni: *molti archivi pubblici sono già attivamente coinvolti nell'archiviazione dei dati di ricerca e hanno ampie competenze per questo e per la gestione FAIR di questi dati. L'obiettivo della conservazione sostenibile dei contenuti può essere raggiunto solo da un punto di vista pratico se i produttori di dati, i gestori di dati e gli archivi responsabili collaborano il prima possibile nel ciclo di vita dei dati e idealmente a livello istituzionale. Iniziative su larga scala come la National Research Data Infrastructure (NFDI) consentono inoltre di sviluppare congiuntamente gli standard e le best practice necessari con l'aiuto di reti di competenze nello spirito della scienza archivistica aperta e di supportare il necessario cambiamento culturale verso una gestione sostenibile dei dati.*

Parole chiave: *gestione dei dati di ricerca, archiviazione a lungo termine, conservazione dei contenuti*

UPRAVLJANJE RAZISKOVALNIH PODATKOV - SKUPNA NALOGA ZA ARHIVE IN PROIZVAJALCE PODATKOV

Izvleček

Namen: *Namen raziskave je bil preveriti, v kolikšni meri arhiviranje raziskovalnih podatkov vpliva na javne arhive in kakšni izzivi in priložnosti se pojavijo zaradi tega.*

Metoda/pristop: *Pristop, uporabljen v tej raziskavi, temelji na vrednotenju različnih pobud za upravljanje raziskovalnih podatkov in projektov sodelovanja arhivov pri arhiviranju raziskovalnih podatkov.*

Rezultati: *Upravljanje raziskovalnih podatkov je pogosto zasnovano le za objavo podatkov in kratka nadaljnja obdobja dolgotrajnega shranjevanja okoli 10 let. Vendar pa različne veje znanosti zahtevajo raziskovalne podatke v veliko daljših časovnih obdobjih. Da bi bilo to mogoče, je potrebno tesno sodelovanje med proizvajalci podatkov, upravljavci raziskovalnih podatkov in arhivisti*

Sklepi/ugotovitve: *Številni javni arhivi se že aktivno ukvarjajo z arhiviranjem raziskovalnih podatkov in imajo obsežna znanja za to in za upravljanje s temi podatki po principu FAIR (Findable, Accessible, Interoperable and Reusable / Najdljivo, dostopno, interoperabilno in ponovno uporabno). Cilj trajnostnega ohranjanja vsebine je s praktičnega vidika mogoče doseči le, če proizvajalci podatkov, upravljavci podatkov in odgovorni arhivi sodelujejo čim prej v življenjskem ciklu podatkov, in najbolje na institucionalni ravni. Obsežne pobude, kot je nacionalna raziskovalna podatkovna infrastruktura (NFDI), prav tako omogočajo skupni razvoj potrebnih standardov in najboljših praks s pomočjo kompetenčnih mrež v duhu odprte arhivske znanosti in podpirajo potrebne kulturne spremembe v smeri trajnostnega upravljanja podatkov.*

Ključne besede: *upravljanje raziskovalnih podatkov, dolgoročno arhiviranje, ohranjanje vsebine*

1. FOCUS ON RESEARCH DATA MANAGEMENT

For several years, various initiatives in Germany, as well as in other European countries and at EU level, have been trying to improve the management of research data. In recent years, many universities and research institutions have created positions to impart the relevant knowledge to employees, students and researchers and usually also make this available online. One example is the University of Konstanz, which offers an important information platform on the subject with the website Forschungsdaten.info. In addition, there are now a number of regional and supra-regional networks, including at state level, such as the Research Data Management Bavaria initiative. At the legislative level, work is currently underway on a research data law that is intended to significantly improve access to data for science (BMBF, 2024). The most important impetus in Germany was certainly given in 2020 with the promotion and development of a National Research Data Infrastructure (NFDI), in which over 270 universities, research and infrastructure institutions, including libraries and public archives, are currently participating. This major project also forms the German pillar for the European Open Science Cloud (EOSC), which pursues corresponding goals for a research data infrastructure at EU level. By building infrastructure and a large number of accompanying projects and initiatives, the aim is to significantly improve findability, accessibility, interoperability and reusability, in accordance with the so-called FAIR criteria. In the medium term, this is supposed to lead to nothing less than a change in the working culture of data producers. Research data should thus be verifiable and reusable for further research even a long time after it has been created, e.g. long after the completion of corresponding scientific projects. So-called data management plans are an important tool in this regard. These are intended to help data producers, e.g. the project staff, to clarify important questions about the filing, structuring, documentation, publication and long-term storage of the data before the data is created. The focus is usually primarily on the publication of the data and secondarily on its long-term storage, with a period of at least 10 years being assumed, as required by the German Research Foundation’s guidelines for ensuring good scientific practice (DFG, 2019).

However, further processing of the data in accordance with the requirements of long-term content preservation of the data is often not considered (Paul-Stüve et al., 2023; Markus et al., 2024). It became clear, not least during the development process of the NFDI, that a wide variety of scientific disciplines need to keep data interpretable for their own research over much longer periods of time than is currently taken into account in standard data management plans. For a wide variety of questions in earth system science, climate and biodiversity research, but also medical research and, of course, history, the periods of around 10 years that have been required so far are in no way sufficient. The same applies to cross-disciplinary research or disciplines such as the history of science. The need for sustainable data management is therefore obviously there. The German Council for Scientific Information Infrastructures (Rat für Informationsinfrastrukturen, RfII) has therefore declared longterm archiving to be one "of the most important tasks of a national research data infrastructure" (RfII, 2016).

2. RESEARCH DATA AND ARCHIVES

The extent to which the topic also affects public archives, in particular state archives, is often underestimated. Instead, it is often emphasized that these archives are not responsible for data from research processes but for the data of authorities and other public bodies and thus only for administrative data. However, this assessment is not correct in many cases. On the one hand, research institutions often also fall under the responsibility of these archives. In Bavaria, for example, there are several state authorities and their partly independent, partly affiliated subordinate bodies, such as the Bavarian State Office for Agriculture, the Bavarian State Office for Woods and Forestry, the Bavarian State Office for the Environment or the State Institute for Health and Food Safety. All these authorities mentioned here as examples, all of which are under the responsibility of the Bavarian State Archives, generate research data in the narrower sense, i.e. data from research projects that were partly funded with state funds, partly with third-party funds and produced by scientific employees of these authorities. The responsibility for the long-term archiving of this data, provided it is of lasting value for administrative or scientific purposes, undoubtedly lies with the Bavarian State Archives. However, it is debatable what is meant by the term research data. More narrowly defined definitions only include the interim results and results of scientific research (Kindling & Schirmacher, 2013). The definition currently used for the National Research Data Infrastructure of Germany (NFDI) by the expert commission appointed by the Joint Science Conference of the Federal Government and the States, the Council for Information Infrastructure (RfII), defined the term research data much more broadly as early as 2019. Accordingly, this includes not only any data that originates from research processes or provides information about the methods and research tools used, but also data from surveys, measurements and data that was not obtained by the researcher, e.g. from official statistics or government information, which science accesses for research purposes, e.g. as a methodological basis (RfII, 2019). The NFDI‘s statement on the German Federal Government‘s planned Research Data Act also points out the importance of older data, which is stored in archives, among other places (NFDI, 2023) and advocates the adoption of the definitions of the German Data Use Act (Datennutzungsgesetz - DNG, 2021) and the so-called PSI or Open Data Directive of the EU (PSI-Richtlinie, 2019), according to which evidence that is used as part of the research process or that is generally considered necessary in the research community for the validation of research findings and results also falls under research data. According to these definitions, public and especially state archives hold research data on a large scale and, in accordance with their area of responsibility, are also responsible for the acquisition and long-term archiving of data that will arise in the future.

3. EXPERTISE OF PUBLIC ARCHIVES IN DATA MANAGEMENT

In addition, legislation in Germany has assigned archives the task of not only archiving data of lasting value, but also of advising data producers on the administration and storage of this data for decades. State archives in Germany in particular have taken this task very seriously with regard to administrative data in recent decades, as shown by the involvement of the State Archives of North Rhine-Westphalia in the introduction of e-files. (Friedrich & Schlemmer, 2018) Public archives therefore have extensive experience in the sustainable structuring, storage and documentation of data in a form that allows it to be reused in accordance with the primary purpose of use during the retention period, i.e. over the period of long-term storage. But that’s not all: For more than twenty years, state archives in Germany have been taking over electronic documents for long-term archiving. Most of this data comes from administrative processes. However, some archives, such as the State Archives of Baden-Württemberg Bavaria and Hessen, have already incorporated research data in a narrower sense into their holdings.[2] Archives have also been able to gain a wealth of experience in the design of logical archiving interfaces, i.e. in the evaluation of data and the question of which formats, structures and transport routes these data take to reach the archive and which metadata and documentation material these data must be enriched with in order to keep them as interpretable as possible for the purposes of long-term archiving and to be able to make them usable again even after long periods of time.

4. DATA MANAGEMENT REQUIREMENTS FOR THE LONG-TERM AVAILABILITY OF RESEARCH DATA

If research data is to be kept interpretable for longer periods than just 10 years, this is evidently possible only through format migration according to the common preservation strategy. This in turn requires that the data is ideally migrated to formats that are suitable for long-term archiving before being transferred to the archive. Since a lot of information is lost during migration, the data must be sufficiently explained with metadata and documentation material on the context in which the information was created (Bachmann et al., 2023). In addition, the significant properties of the data must be defined (Puchta et al., 2023). And it must be determined in which structure and with which file and folder names the data should be transferred to the target archive. In numerous public archives in Germany, particularly in the relatively well-staffed state, municipal and church archives, this approach of defining and implementing logical archiving interfaces has been tested and implemented in recent years by a number of producers of data of lasting value from administrative processes. The cultural change towards sustainable data management, in which archiving is already considered when the data is created, is also slowly beginning to become noticeable here. In some cases, such as the cooperation between the Bavarian State Ministry of Food, Agriculture, Forestry and Tourism and the General Directorate of the Bavarian State

[2] These include, among others, the archiving of data from the Württemberg State Museum by the Baden-Württemberg State Archives, biodiversity data from an inventory project at the Berchtesgaden National Park by the General Directorate of the State Archives of Bavaria, and the recording of biotopes in nature reserves in the Hessian State Archives.

Archives, archives are already involved in the conception phase of new IT systems or the conversion of existing systems in order to accompany the integration of appropriate interfaces if they are archival worthy (Holzapfl et al., 2023).

Progress in this direction has also been observed in recent years in born-digital data that matches to a greater extent with the narrower definition of research data, such as the agreements jointly drawn up between data producers and state archives in the federal states on the nationwide archiving of statistical raw data (KLA AG Statistical Data, 2008) since 2008 and of official geodata since 2015 (KLA Geodata, 2015). But public archives are also now involved in advising on the management of data that comes from research processes, with the NFDI offering a suitable framework with sufficient cross-linking opportunities for this. This major initiative not only includes data producers, but also those responsible for research data management as well as a wide variety of data repositories and libraries, of which some also have extensive experience in data curation and digital long-term archiving in the sense of content preservation.

5. COLLABORATION BETWEEN ARCHIVES AND DATA PRODUCERS IN DATA MANAGEMENT

These cooperations, which have so far been rather sporadic, already show the potential of collaboration, but also make clear the challenges of managing data that originates from research processes and may need to be archived permanently or at least for longer than 10 years in order to preserve information. The question of how to evaluate this research data is not easy to answer. The classic archival toolbox of vertical and horizontal comparison or random selection of large quantities of uniform documents quickly reaches its limits here. Formal criteria for identifying outstanding research projects are usually lacking. The archivists responsible for the evaluation are usually not specialists in the respective scientific discipline. In order to identify research projects and data of lasting value, collaboration between long-term archives and data producers or the specialist departments of the respective authorities and institutions is therefore urgently required. This applies not only to the evaluation of the data, but also to its management with the aim of long-term archiving. Only if data producers and target archives work together before the data is created the long-term preservation of the data can be guaranteed and the additional effort required to prepare the data for this purpose kept to a minimum. This is particularly important in the case of data that is often generated in the context of tightly timed research projects with dynamic personnel development. On the basis of close institutional cooperation between the actors, i.e. the data producers, those responsible for research data management at these bodies and the archives, workflows for early evaluation and common principles of data management at the beginning of the life cycle of research data should be developed (Valena, 2024). The Bavarian State Archives are currently exploring options for this within the framework of the NFDI consortium FAIRagro with the research departments of the Bavarian State Ministry of Food, Agriculture, Forestry and Tourism. (StMELF, 2024)

6. CONCLUSION

Although there is still no generally accepted definition of the term research data, it can be assumed that many public archives already archive research data and will continue to be responsible for archiving research data in the future. With their experience in advising data producers and designing archiving interfaces for data that originate from administrative processes, public archives have skills that are also important for the management and sustainable FAIRification of research data. Similar to digital administrative documents, research data also requires close cooperation between data producers, research data management and the responsible archives as early as possible in the data life cycle in order to enable the data to be reused for relevant periods of time. However, it has also been rightly pointed out that the challenges and in particular the question of resources stand in the way of a systematic strategy for the archiving of research data in the narrower sense by public archives (Naumann, 2024). In view of scarce resources, a wide range of tasks and the parallel analog and digital sorting of administrative documents that will have to be managed over the next few decades, the archiving of research data does not appear to be a priority for many public archives. However, if these data come from data producers within the area of responsibility of the public archives and have lasting value for scientific or other purposes, the archiving of these data also falls under the responsibility of the respective archives. The extent to which research data in the narrower sense, which comes from other data producers, should also be archived in public archives, for example in an extension of the archival collection mandate, still needs to be discussed within the archive community (Naumann, 2024). Last but not least there are further and legal questions that need to be answered here (Hodenberg et al., 2023). Ultimately, initiatives such as the NFDI, which not only function as a service portfolio and data infrastructure but also form an exchange platform and a competence network, offer the opportunity to develop work-efficient solutions and standards for these tasks together with data management experts from a wide range of memory institutions and scientific fields (Grau et al., 2023) in the spirit of open archival science.

REFERENCES

Bachmann, C., Beider Wieden, B., Eichler, V., Graf, S., Hering, R., Keitel, C., Kleon, M., Kluttig, T. and Wettmann, A. (2008). *KLA-AG Bewertung von Statistikunterlagen. Abschlussbericht*. Retrieved at https://www.bundesarchiv.de/assets/bundesarchiv/de/Downloads/Berichte/abschlussbericht-statistikunterlagen-kla.pdf (accessed on 11.09.2024).

Bachmann, C., Herget, R., Stehr, M., Unger, M., Puchta, M. and Schmalzl, M. (eds.). (2023). *Richtlinien zur Verzeichnung von Archivgut der Staatlichen Archive Bayerns*. München: Generaldirektion der Staatlichen Archive Bayerns. Retrieved at https://www.gda.bayern.de/fileadmin/user_upload/Medien_fuer_Unterseiten/Verzeichnungsrichtlinien_Stand_10-2023.pdf (accessed on 11.09.2024).

Bayerisches Staatsministeriums für Ernährung, Landwirtschaft, Forsten und Tourismus (StMELF). (2024). *Workshop zu Brennpunkten und Lösungsansätzen im Forschungsdatenmanagement am 21.—22.10.2024*. Retrieved at https://www.stmelf.bayern.de/ministerium/forschung/forschungsdatenmanagement/index.html (accessed on 11.09.2024).

Bundesministerium für Bildung und Forschung (BMBF). (2024). *Eckpunkte BMBF Forschungsdatengesetz 2024*. Retrieved at https://www.bmbf.de/SharedDocs/Downloads/de/2024/240306_eckpunktepapier-forschungsdaten.pdf?__blob=publicationFile&v=3 (accessed on 11.09.2024).

Deutsche Forschungsgemeinschaft (DFG). (2019). *Leitlinien zur Sicherung guter wissenschaftlicher Praxis. Kodex*. Retrieved at https://www.dfg.de/resource/blob/173732/4166759430af8dc2256f0fa54e009f03/kodex-gwp-data.pdf (accessed on 11.09.2024).

Friederich, C. and Schlemmer, M. (2018). Das E-Government-Gesetz NRW und die Praxis der Behördenberatung. Ein Werkstattbericht aus dem Landesarchiv NRW. *Informationswissenschaft: Theorie, Methode Und Praxis*, *5*(1), 95—104. Retrieved at https://doi.org/10.18755/iw.2018.10 (accessed on 11.09.2024).

Gesetz für die Nutzung von Daten des öffentlichen Sektors (Datennutzungsgesetz - DNG). (2021). Retrieved at https://www.gesetze-im-internet.de/dng/ (accessed on 11.09.2024)

Grau B., Schmalzl M. and Unger M. (2023). Die Beteiligung der staatlichen Archive Bayerns in der NFDI. *Archiv. Theorie und Praxis*, *76*(1), 40—42. Retrieved at https://www.archive.nrw.de/sites/default/files/media/files/Archivar20231InternetmitAnzeigen.pdf (accessed on 11.09.2024).

Hodenberg C., Naumann K. and Siegers P. (2023). Wie Archive und Forschungsdaten zueinander finden: ein gegenseitiger Lernprozess. *Archiv. Theorie und Praxis, 76*(3), 187—194. Retrieved at https://www.archive.nrw.de/sites/default/files/media/files/Archiv.theorie-praxisHeft3-2023Internet_0.pdf (accessed on 11.09.2024).

Holzapfl J., Nestl A., Puchta M., Schmalzl M. and Unger M. (2023). Quick Wins und dicke Bretter. Übernahme und Archivierung von Fachverfahren. *Archivar, 76*(1), 15—24. Retrieved at https://www.archive.nrw.de/sites/default/files/media/files/Archivar20231InternetmitAnzeigen.pdf (accessed on 11.09.2024).

Kindling, M. und Schirmbacher, P. (2013). "Die digitale Forschungswelt" als Gegenstand der Forschung. *Information – Wissenschaft – Praxis, 64*, 127—136. Retrieved at https://doi.org/10.1515/iwp-2013-0017 (accessed on 11.09.2024).

KLA/AdV-AG. (1. 6. 2015). *Guidelines for the nationwide uniform archiving of geographic reference data.* Retrieved at https://www.bundesarchiv.de/assets/bundesarchiv/de/Downloads/Erklaerungen/guidelines-geoarchiving-kla.pdf (accessed on 11.09.2024).

Markus, K., Naumann, K., Schmalzl, M., Watson, J. and Triebel D. (10. 5. 2024). Long Term Archiving in the NFDI. *Zenodo.* Retrieved at https://zenodo.org/records/11109480 (accessed on 11.09.2024).

Nationale Forschungsdaten Infrastruktur (NFDI). (2023). *Stellungnahme zum Forschungsdatengesetz*. Retrieved at https://www.nfdi.de/wp-content/uploads/2023/05/NFDI-Stellungnahme-zum-Forschungsdatengesetz.pdf (accessed on 11.09.2024).

Naumann, K., Raphael, L. and Siegers, P. (2024). Abstimmungsbedarf zur Überlieferungsbildung und Archivierung von Forschungsdaten. In: Becker, I. C., Haffer, D., Lehrmann, F. and Meier, R (eds.), *Archivists meet Historians. Transferring Source criticism to the digital age. Beiträge zum 27. Archivwissenschaftlichen Kolloquium der Archivschule Marburg* (pgs. 217–239). Marburg: Arhivschule Marburg.

Paul-Stüve, T., Schürmann, T., Valena, P. (2023). Survey covering status of long-term preservation and long-term archiving in ESS in Germany (NFDI4Earth Deliverable D2.4.1). *Zenodo.* Retrieved at https://doi.org/10.5281/zenodo.11200016 (accessed on 11.09.2024).

Puchta, M., Ksoll-Marcon, M., Grau, B., Kirstein, M., Unger, M., Schmalzl, M. and Nestl, A. (eds.). (2023). *Fachkonzept für das Digitale Archiv der Staatlichen Archive Bayerns*. München: Generaldirektion der Staatlichen Archive Bayerns. Retrieved at https://doi.org/10.5281/zenodo.7743888 (accessed on 11.09.2024).

Rat für Informationsinfrastrukturen (RfII). (2016). *Leistung Aus Vielfalt. Empfehlungen Zu*

Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in

Deutschland. Göttingen: Rat für Informationsinfrastrukturen. Retrieved at https://d-nb.info/1104292440/34 (accessed on 11.09.2024).

Rat für Informationsinfrastrukturen (RfII). (2019). *Herausforderung Datenqualität – Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel*. Göttingen: Rat für Informationsinfrastrukturen. Retrieved at https://rfii.de/download/herausforderung-datenqualitaet-november-2019/# (accessed on 11.09.2024).

Richtlinie (EU) 2019/1024 des Europäischen Parlaments und des Rates vom 20. Juni 2019 über offene Daten und die Weiterverwendung von Informationen des öffentlichen Sektors (PSI-Richtlinie). (2019). *Amtsblatt der Europäischen Union,* (L 172/56). Retrieved at https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32019L1024 (accessed on 11.09.2024).

Valena, P. (2024). Tagungsbericht zum Workshop zur Entwicklung von Kriterien für die Bewertung der Archivwürdigkeit von Daten der Bio-/Geowissenschaften am 14.12.2023 in Potsdam. *Zenodo.* Retrieved at https://doi.org/10.5281/zenodo.13330901 (accessed on 11.09.2024).

Summary

For several years, various initiatives in Europe have been trying to improve the findability, accessibility, interoperability and reusability of data for science and research. At the same time, the needs of a digitalized administration require electronic data to be stored in a fully interpretable manner, sometimes over several decades. This also requires consistent and sustainable records or data management. The lecture examines the question of what contribution public archives can make to overcoming these challenges and argues that data producers and archives must cooperate closely even before the beginning of the data life cycle.

Typology: 1.01 Original Scientific Article