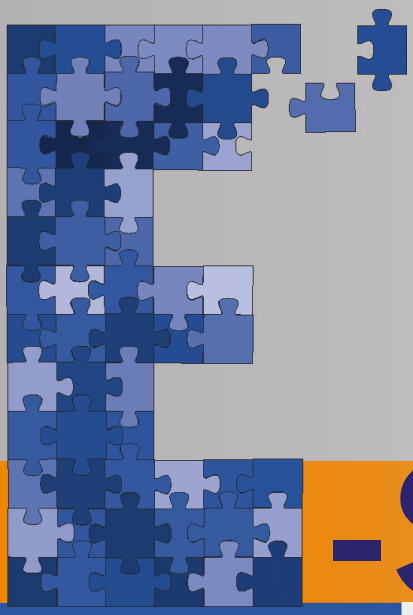
Jonas Kratzke /
Vincent Heuveline (Hrsg.)



E-Science-Tage 2017

Forschungsdaten managen

UNIVERSITÄTS-
BIBLIOTHEK
HEIDELBERG

E-Science-Tage 2017

E-Science-Tage 2017

Forschungsdaten managen

Herausgegeben von

Jonas Kratzke und Vincent Heuveline

UNIVERSITÄTS-
BIBLIOTHEK
HEIDELBERG

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

Publiziert bei heiBOOKS,
Universitätsbibliothek Heidelberg 2017.

Die Online-Version dieser Publikation ist auf heiBOOKS,
der E-Book-Plattform der Universitätsbibliothek Heidelberg,
http://books.ub.uni-heidelberg.de/heibooks, dauerhaft frei verfügbar
(Open Access).
urn: urn:nbn:de:bsz:16-heibooks-book-285-5
doi: https://doi.org/10.11588/heibooks.285.377

ISBN 978-3-946531-75-3 (PDF)

Inhaltsverzeichnis

Grußwort der Ministerin für Wissenschaft, Forschung und Kunst Baden-Württemberg

Theresia Bauer

Die Digitalisierung unserer Lebens- und Arbeitswelt verändert die Art und Weise, wie Wissenschaft und Forschung betrieben werden. Daten, die früher mit aufwändigen Verfahren gesammelt und kodiert wurden, werden heute automatisiert erzeugt. Neue Analysetechniken erlauben die Entwicklung und Überprüfung hochkomplexer Hypothesen und Modelle. Der wissenschaftliche Fortschritt lebt von dem Diskurs und der Zusammenarbeit der Wissenschaftler - sowohl innerhalb einer Disziplin als auch interdisziplinär - und von der Überprüfbarkeit wissenschaftlicher Hypothesen und Modelle. Dafür ist es erforderlich, dass die Daten, auf denen die Hypothesen und Modelle beruhen, dem FAIR-Prinzip (Findable, Accessible, Interoperable, Reusable) entsprechend, auffindbar, zugänglich, interoperabel und wiederverwendbar sind.

Dafür müssen die Hochschulen und Forschungseinrichtungen eine Prozesskette des Datenmanagements schaffen und in ihrem, auch von Fluktuation geprägten Kosmos implementieren. Das Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg hat sich bereits 2014 zur Umsetzung des von Experten erarbeiteten Fachkonzepts E-Science zur Weiterentwicklung der wissenschaftlichen Infrastruktur in Baden-Württemberg bekannt. Seit 2016 fördern wir eine Reihe von Projekten zum Forschungsdatenmanagement und zu Virtuellen Forschungsumgebungen für digitale Forschungsdaten. Hier erarbeiten Vertreter der Fachdisziplinen gemeinsam mit Vertretern der Bibliotheken und Rechenzentren Modelle und Strukturen für das Management, die Archivierung und die digitale Analyse von Forschungsdaten in virtuellen Umgebungen. In diesem Kontext ist die Idee der E-Science-Tage entstanden.

Ich freue mich sehr über die Resonanz, die unsere Konferenz „E-Science-Tage 2017: Forschungsdaten managen“ mit hochrangigen Rednern und mehr als zweihundert Teilnehmern gefunden hat. Ich freue mich besonders über die Bereitschaft der Vortragenden, ihren Tagungsbeitrag nachträglich zu verschriftlichen. Mit der Veröffentlichung ihres Vortrags im Tagungsband tragen sie ihren Teil zur Weiterentwicklung der digitalen Infrastrukturen, Prozesse und Dienste des Forschungsdatenmanagements bei. Angesichts der intensiven und lebhaften Diskussionen, die die E-Science-Tage geprägt haben, wünsche ich dem Tagungsband die ihm gebührende Aufmerksamkeit in der Fachwelt.

Theresia Bauer MdL
Ministerin für Wissenschaft, Forschung und Kunst des Landes Baden-Württemberg

Vorwort der Herausgeber

Jonas Kratzke, Vincent Heuveline

Die E-Science-Initiative stellt sich den Herausforderungen des digitalen Wandels in den Wissenschaften. Ein wesentlicher Bestandteil des digitalisierten Forschens ist das Management teilweise immenser und heterogener Mengen an Forschungsdaten. Die Weiterentwicklung des Forschungsdatenmanagements ist dabei in den vergangenen Jahren und Jahrzehnten selbst zu einem sehr lebhaften Forschungsfeld geworden. Es liegt im gemeinsamen Interesse der Wissenschaftler/innen und Infrastruktureinrichtungen digitale Forschungsmethoden und das nachhaltige Publizieren und Aufbewahren von Daten voranzutreiben.

Die E-Science-Tage 2017 stellten ein zweitägiges Forum für Wissenschaftler/innen aller Disziplinen, Infrastruktureinrichtungen und Vertreter/innen aus der Politik dar, um die aktuellen Fragen rund um das Thema Forschungsdatenmanagement zu diskutieren. Als Gastgeber an der Universität Heidelberg freuten wir uns über die zahlreichen Beiträge in Form von Keynotes, Workshops, einer Podiumsdiskussion, Gesprächstischen, Postern und schließlich von Konferenzvorträgen. Zu Letzteren freuen wir uns ganz besonders, sie in der Form des vorliegenden Tagungsbandes einem breiteren Publikum zugänglich machen zu können. Von der Fortentwicklung eines nachhaltigen Forschungsdatenmanagement profitieren mehr und mehr Fach-Communitys. Dieser Tagungsband zeigt die Fortschritte und die Vielfältigkeit der Ansätze in den verschiedenen Disziplinen. Die tief gehenden Beiträge dieses Bandes decken sowohl die Perspektiven aus den Natur- und Geisteswissenschaften als auch die Entwicklungen seitens der universitären Bibliotheken und Rechenzentren ab.

Nach vorne schauend in das digitale Zeitalter der Forschung, liegen viele Veränderungen und Herausforderungen am Horizont: mehr Digitalisierung in allen Disziplinen sowie kontinuierlich anwachsende Mengen an Rohdaten, die es zu speichern, verarbeiten und archivieren gilt. Ein Schlüssel zum Erfolg liegt in dem Zusammenspiel zwischen Speicherhardware-Providern und fachgerechter Organisation und Aufbewahrung wertvoller Forschungsdaten. Der vorliegende Tagungsband zeugt davon, dass wir auf dem richtigen Weg in die Zukunft sind.

Zahlreiche Unterstützer/innen und Helfer/innen trugen zum Erfolg der E-Science-Tage 2017 bei, welcher schließlich in diesen Tagungsband mündet. Unser ausdrücklicher Dank gilt zunächst dem Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg, welches die Tagung großzügig sowohl in direkter Weise als auch über das Projekt bwFDM-Info finanziell unterstützte. Insbesondere danken wir Frau Dr. Katrin Behaghel für die Begleitung in der Planung und Vorbereitung der Tagung. Außerdem danken wir allen Teilnehmer/innen und Beitragenden sowie den Key-Accountern und Verantwortlichen des Projektes bwFDM-Info für die hervorragende Zusammenarbeit in der Konzeption, dem Review, der Auswahl der einzelnen Beiträge und der

Durchführung der Tagung. Ein besonderer Dank gilt der Geschäftsstelle und dem Servicebereich Future IT – Research & Education (FIRE) des Universitätsrechenzentrums Heidelberg für die umfangreiche Unterstützung vor, während und nach den E-Science-Tagen 2017.

Jonas Kratzke
Prof. Dr. Vincent Heuveline
Universitätsrechenzentrum der Universität Heidelberg

Research Data Management training and support services at both ETH Zurich and EPF Lausanne

Lessons Learned, Best Practices and the Way Forward

Ana Sesartic[1], Aude Dieudé[2]

1 ETH Library, ETH Zurich, ana.sesartic@library.ethz.ch
2 Ecole Polytechnique Fédérale de Lausanne Library, EPFL, aude.dieude@epfl.ch

Abstract. The management of research data throughout its life-cycle ensures its long-term value and preservation, as well as being a key prerequisite for effective data sharing. Many funding bodies mandate the creation of data management plans and open access publication of the research results they funded. In order to concentrate the efforts of different universities, libraries, and IT-services, the project "Research Data Life-Cycle Management: From Pilot Implementations to National Services (Data Life-Cycle Management, DLCM[1])", was launched by eight Swiss Higher Education Institutions on behalf of the former Swiss University Conference (SUC) as part of the programme SUC P-2 "DLCM", which is chiefly designed to improve the handling of scientific information in Switzerland.

With the start of DLCM in 2015 the Libraries of the two Swiss federal institutes of technology, EPF Lausanne and ETH Zurich, rose to the occasion and strengthened their respective initiatives and efforts by creating personalized data management plan (DMP) services and training sessions on research data management. The services and training sessions were set up in close collaboration with the scientific IT departments and the Research Office at EPFL Library and ETH Library, thus enabling the libraries to offer expertise tailored to researchers' needs, and to cover the entirety of the data life-cycle. The closer collaboration with researchers during training sessions and consultations contributed to the establishment of stronger trust relationships between scientists and information professionals at the two Swiss federal institutes of technology. A mutual learning process was sparked, allowing both sides to share best practices and tackle new challenges together.

By building on existing experiences, resources, and tools available within Swiss higher education institutions, the DLCM project focuses on the common goals: to harmonize and strengthen research data management practices across actors (researchers and information professionals), disciplines (sciences, social sciences, and humanities), and institutions (universities, universities of applied sciences, and universities' service providers) in a sustainable way. While universities are obviously competing for funding, students, and scholarly excellence, there is a need to use available resources efficiently and also to share best practices acknowledging the high level of staff mobility between institutions. With this background, we aim to share our insights, experiences, and evolving best practices regarding research data management training, resources, and support services. To this end, our paper focuses on four major aspects: our new data management training, our personalized research data management support services, our lessons learned and how we envision the way forward.

Keywords. Research data management, training, services, best practice, Switzerland

1 https://www.dlcm.ch/

Introduction

The rising production of research data has created new challenges for its management. In order to ensure its continuity, transparency and accountability, the timely and effective management of research data throughout its life-cycle is essential (Louise Corti et al. 2014; Goodman et al. 2014). Proper data management is also a key prerequisite for effective data sharing and publication, which, in turn, increase the visibility of scholarly work and are likely to increase citation rates (H A Piwowar and Vision 2013; Heather A. Piwowar, Day, and Fridsma 2007). Managing research data is furthermore a sine qua non condition for efficient long-term preservation because the latter must rely on sufficient metadata and contextual information being available to make sure that data remains reproducible, reusable and understandable in the long run.

As effective and efficient data management becomes more and more challenging for both researchers and information specialists across institutions, the question arose as to how they can best be reached and supported on a national level. The project "Research Data LifeCycle Management: From Pilot Implementations to National Services (DLCM)", aims to concentrate the efforts of 8 Swiss universities (EPFL, ETH Zurich, Geneva School of Business Administration/University of Applied Sciences and Arts Western Switzerland, University of Basel, University of Geneva, University of Lausanne, University of Zurich), represented by their libraries and IT-services including the existing national service provider for HEI (SWITCH). It was initiated on behalf of the former Swiss University Conference (SUC) as part of the programme SUC P-2 "DLCM", which is chiefly designed to improve the handling of scientific information across the country.

With the start of DLCM in 2015 the EPFL and ETH Libraries reinforced their respective efforts, which had already begun in 2012 with the creation of personalized Data Management Plan (DMP) services and training sessions on research data management, as many funding bodies have mandated the creation of data management plans and open access publication of the research results found. The services and training sessions were set up in close collaboration with the scientific IT departments and Research Office at EPFL Library and ETH Library, thus enabling the libraries to offer expertise tailored to researchers' needs, and to cover the entirety of the data life-cycle. Soon after, the EPFL Library and ETH Library joined their efforts by collaborating on a Data Management Checklist in 2015 and establishing training and consulting services on their own in 2016.

The closer interaction with researchers during tailored training sessions and consultations contributed to the establishment of stronger trust relationships between scientists and information professionals at the two Swiss federal institutes of technology. In the following, we are sharing our insights, experiences, evolving best practices and lessons learned regarding research data management training, resources and support services.

Research data management training

In step with the growing awareness of the value of research data and the risks of losing such data over time, the need for research data management (RDM) has gained increased attention over the last years. RDM combines both the need to manage data over the course of a project, as well as the curation and preservation of data for future work and reference.

To address these questions at the ETH Zurich (see also Sesartic and Töwe 2016 for further information on research data services at the ETH Zurich), the ETH Library Digital Curation Office developed RDM training sessions in the form of a basic 1.5 hour training and an extended half-day workshop. The aims of the trainings are to raise awareness of existing requirements and of benefits to be gained from proper RDM, to introduce some services and tools for RDM as well as to encourage participants to share both their experiences and the methods and tools they use, during the interactive parts of the workshop. Researchers must be empowered to make informed decisions on their data, as they are the experts with the most intimate knowledge. Activating teaching methods, which engaged the participants in group work and discussions, facilitated direct exchange between the peers as well as with the trainers.

The trainings are offered free of charge and are open to everyone, with a focus on members of the ETH. Most participants taking part in the workshop and the short 1.5 hour training about RDM are doctoral students, with some postdocs, senior scientists and technical staff present. They generally showed varying needs and levels of knowledge, but all were aware of the problems surrounding RDM and were happy to learn about possible solutions. The trainings also proved to be excellent marketing instruments, as nearly every training led to invitations from research groups for more tailored trainings, after one or several group members visited our courses.

In order to better cater to the varying groups, the ETH Library in general offers tailored training courses for groups and departments. These can range from 15-minute short mini lectures over coffee or lunch breaks, to full-fledged one-day training workshops. As some departments and institutes already offer similar internal training, communication and coordination with them is key

The ETH Library is planning to establish a dedicated course on RDM and related topics within the ETH curriculum in the future, but to do so will take further time and planning.

To answer the evolving and growing needs of its researchers regarding research data management, the EPFL library set up in January 2015 a steering committee composed of the heads of the IT department, the Research Office and the library to find the best ways to answer them. One month later, on February 2015, the EPFL library received the official green light to start offering in close collaboration with the IT department and the Research Office departments a personalized support service regarding research data management and the preparation of a data management plan. The first six months proved the relevance of such a service, and as a result, the service developed itself and some personalized trainings were also offered to pursue these efforts even further. Therefore in the fall of 2015 two annual trainings on how to optimize research data management were given free of charge by two collaborators of the library, Aude Dieudé[2] and Jan Krause[3], to the entire EPFL community as part of the official staff training service. Due to the success of these two initial workshops in 2015, four training sessions[4] are now offered in both English and French each year and the participants are quite eclectic. They may include principal investigators (PI) and senior researchers, postdoctoral and doctoral students, but also IT specialists, information specialists, project managers, librarians and information specialists. In addition, personalized trainings have been offered on demand to EPFL PIs to train their lab team members in a harmonious way and the feedback we have received has been both encouraging and positive.

Presently in 2017, the EPFL research data management support service team is composed of data librarians, data managers, liaison librarians, and IT project managers. This diversity of back-

2 https://people.epfl.ch/aude.dieude?lang=en
3 https:/people.epfl.ch/jan.krause?lang=en%20
4 http://sfp.epfl.ch/files/content/sites/sfp/files/users/132694/public/Descriptifs/Optimiser_gestion_donnees_recherche.pdf

ground and speciality allows to offer a full picture to the researchers of the different sets of skills relevant for fully considering the data lifecycle management: from its collection and creation, to its description and preservation, and finally to its sharing and reproducibility.

Throughout our training, the team makes sure to fully understand the context, field and specific needs and questions of the participants. Each training is unique and requires dedicated resources and personalized attention; however, here are a series of questions that are regularly asked:

1. What are the best practices for research data management for my lab and my team?
2. How can I ensure that everyone in the lab or all the partnering institutions in this research project is following the same methodologies to save time, energy and money for all of us in the long term?
3. Can you explain to us what a data management plan is, why it is beneficial for us and how to create and maintain one in an efficient, practical and cost-effective way?
4. How much will it cost to store my data during the next ten, fifteen or twenty years?
5. What are the options to safely store, exchange and organize my sensitive data?
6. Do you know of any specific tools that would be best to use in this particular case?
7. Would it be possible to organize a personalized training for my lab in the future?

In providing free expertise and consulting services to researchers, our team is building a solid basis for mutual respect, trust and exchange of best practices. The very fact of having dedicated persons willing to explore with them the best options in their particular scenario while sharing their expertise, networks and knowhow proved to be very beneficial not only for the researchers and their lab, but also for the research institutions.

Knowing how to create a data management plan and how to efficiently manage their data has become a sine qua non condition for receiving research funding from prestigious funding agencies such as the European Commission with its Horizon 2020 program[5] since January 1, 2017, and also from the Swiss National Science Foundation[6] (SNSF) starting in October 2017. These new requirements have the advantage of introducing Swiss researchers gradually but surely to the sensitive questions of reproducibility and open science more generally. In anticipating this transition and responding effectively to these new requirements, the EPFL library paved the way in training a new generation of researchers across disciplines, generations and methodologies to focus on scientific excellence, personalized needs and quality-oriented services to optimize research data management, reproducibility, and open science. In 2016, a specific training for doctoral students was created and developed at EPFL by the library to reach out new generations as early as possible and directly. As a result of all these joined efforts, several institutions across Switzerland and abroad requested and invited the EPFL library to offer personalized training for its participants. Consequently, three trainings focusing on RDM, DMP and data visualization and tools have already been given since 2015 at the Haute Ecole de Gestion (HEG) located in Geneva. In France, two trainings were offered during the Open Access week in 2015 and at the Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques (ENSSIB) in 2016. Recently, the University of Basel invited the EPFL library to give in 2017 a two-day training workshop on how to optimize research data management. The workshop was fully booked within one day and the waiting list is already full, proving how much research data management trainings are both relevant

5 https://ec.europa.eu/digital-single-market/en/news/communication-european-cloud-initiative-building-competitive-data-and-knowledge-economy-europe%20

6 http://www.snf.ch/en/researchinFocus/newsroom/Pages/news-170306-towards-open-research-data.aspx

and valuable for everyone. The use of a questionnaire composed of specific questions ahead of the training has proved to be highly important to fine-tune and personalize the workshop according to the participants' needs. In addition a detailed feedback form has been created to fully assess the quality and value of such a workshop for the participants. Both of these tools gave us the chance to rethink creatively and offer a variety of tailored trainings based on the pool of participants, which has been highly appreciated and valued. In a nutshell, providing research data management training at both EPFL and ETHZ paved the way to the creation of new content to offer innovative and personalized solutions to our institutions. Far from being a fashion trend or another administrative burden, RDM trainings are becoming a sine qua non condition to put into practice best practices and excellence in academic research on a daily basis.

Data management plan checklist

Well-managed data is both part of good scientific practice and a key requirement of many funding organizations. A data management plan is required for all projects participating in the extended Open Research Data pilot of EU's research programme Horizon 2020 (EC - European Commission. European Research Area, n.d.). But even if a funding body does not demand data management, following its principles has numerous advantages: it helps make data findable, accessible, interoperable and reusable, thus adhering to the FAIR principles[7].

To aid the Swiss community and the researchers at EPFL and ETH Zurich create DMPs, the EPFL Library and ETH Library created in close collaboration a Data Management Checklist[8], which is one of first tangible deliverables of the DLCM project. Based on pre-existing national and international policies, the list has been customized for Switzerland and covers both general planning and all the phases throughout the data life-cycle. Special sections cover documentation and metadata, file formats, storage, ethical and intellectual property issues.

The list is currently available through the ETH Library and EPFL Library websites, as well as disseminated via the DLCM portal. Further transformation into an online tool using DMP Online is planned within the DLCM project to make the experience even more interactive. The list can both serve as a starting point for face-to-face discussions of data management issues within research groups and with support staff, as well as a concrete starting point for researchers to individually assess their data management and gather information they need for the creation of a data management plan.

Experience proves that this personalized checklist for Switzerland provides an excellent way to get the conversation started and to demystify the researchers' habits, methodologies, and expectations regarding how they perceive the role of the library offering practical and useful resources. As such it is a perfect and gentle ice breaker to concretely unveil what a data management plan is about and how such questions help to raise delicate and sensitive aspects of the data life cycle management, which can be easily overlooked or underestimated. In a nutshell, the researchers can realize that a data management plan represents only the tip of the iceberg. The heart and soul of this new requirement is to engage the researchers to take some distance from their previous habits to think thoroughly about the best practices, resources, and tools that can be used to optimize research data management for the long-term. Using both this checklist with the DMP template[9] of-

7 http://www.datafairport.org/
8 https://www.dlcm.ch/ressources/dmp-checklist/
9 https://www.dlcm.ch/ressources/dmp-template/

fers a complementary approach, which demonstrates that both RDM and DMP are not rocket science, but rather ways to guide and support researchers, project managers, IT specialists, data librarians and information specialists towards optimized, personalized and long-term oriented research data management.

Lessons Learned

- Generally positive feedback from researchers. They are very grateful and thankful for such offers and demand more.
- Generally, the researchers are aware of the problems, but do not know how to solve them and do not even know that solutions exist within the university.
- Mind the cultural differences: teaching RDM in Switzerland isn't the same as teaching RDM in another country, as we learned from colleagues.
- Leading by example is important. Get the professor/head of institute into the boat and the doctoral candidates etc. will follow suit regarding RDM.
- Personalized approach is rewarding: from DMP support service to tailored training to word of mouth reputation and credibility.
- RDM training leads to enhanced collaboration with scientists and better visibility for the libraries, as well as improving their image.
- Constructive collaborations are key (win-win approach): among colleagues from the library, among different sets of services within the institutions (including the library, IT, Research Office, and legal experts to name a few) and between institutions nationally and internationally.
- Centralized and harmonized communication, sensitizing actions and quality support service are key to build momentum and trust while changing the image of the library and of librarians.

The Way Forward

Generally, the Swiss data life-cycle management landscape seems to be on the right track regarding RDM services and trainings. To continue improving and expanding our services, the following course of action has been planned:

- Creation of an interactive Swiss DMP tool inspired by already existing tools such as the DMP OPIDoR (Inist-CNRS): https://dmp.opidor.fr/ which is based on https://dmptool.org/
- RDM training offered earning credit points for students to validate their efforts and engagement at their institution, planned at University of Basel and ETH Zurich. Strong wish to put into practice credited training at EPFL in the near future as well.
- Renewal of financial support and funding for the DLCM project for the 2018-2020 period to continue the efforts and harmonization of research data management across Switzerland and between diverse sets of partners (including to name a few Swiss federal institutes of technology, universities, and foundations).
- Continue to change the mentality of the scientific community in close collaboration and communication with the key stakeholders on a national, European and international level: thanks to the new President of EPFL (Martin Vetterli), swissuniversities (the successor of

the Swiss University Conference), and funding agencies, to name a few, to strengthen research data management, open access, and open science.
– Necessity to increase and strengthen the current team and human resources available in the future to make this change of culture and change of mindset a reality. New skills and a diverse team of experts are necessary in light of new requirements from H2020 and SNSF to continue guiding and training researchers and their team. A robust team is crucial to promote research data management in a personalized way, and to focus on building stronger bridges and connections with researchers based on trust and quality services.
– The library is becoming a one-stop-shop, where open science[10],[11] is one of the core services offered to its institution, putting into light one of the many components and elements following the research data lifecycle management.

Conclusion

In conclusion, we can state the following points and our recommendation are based on our mutual collaboration and national context. Therefore, it goes without saying that they have to be adapted with respect to the respective academic institution, professional environment, national policies in place, and funding resources. Our experience at EPFL and ETH Zurich confirms the relevance of a concrete and need-based approach for researchers. To meet this goal a proactive approach and strategic research data management services are crucial to satisfy new funding requirements. A positive way to emulate a cultural shift in academia is through strong supportive and well-established ambassadors, who can serve as "data champions" within the institution (Higman, Teperek, and Kingsley 2017). Within the academic setting and between institutions, creativity, flexibility and constructive and solid collaborations are key. The ability to think outside of the box and to move beyond academic competitions while building consensus provides a strong basis for a fruitful collaboration among well-known, and yet, usually competitive institutions. In reversing this logic, it becomes possible to see the difference among institutions as complementary assets, which nourish and inspire one another constantly. In this sense, the complementary style and approach between EPFL and ETH Zurich has been particularly valuable for paving the way for further collaborations. As an example, our successful collaboration went from working closely together to create and refine the DMP checklist, to offering personalized RDM training, to participating regularly at international conferences (IDCC[12] and iPRES[13] in 2016, E-Science-Tage[14] in 2017), and to co-write several academic articles (e.g. Burgi, Blumer, and Makhlouf-Shabou 2017). This kind of fruitful collaboration is not the exception to the rule and can be implemented in Switzerland and in other countries worldwide. What it takes is the open mindedness, curiosity and ability to share, engage and collaborate with additional colleagues, teams and projects for the benefit of both parties. It is not the exception, but the norm and will continue to become the case even more in the future in order to save time, energy and money. As an extension of this collaboration, we can for instance think of the recent academic initiative and new strong collaboration

10 http://library.epfl.ch/open-science-workshops
11 http://www.library.ethz.ch/Ueber-uns/Veranstaltungen/Think-check-submit-Making-informed-decisions-in-open-access-publishing
12 http://www.dcc.ac.uk/events/idcc16/programme-presentations
13 http://www.ipres2016.ch/frontend/organizers/media/iPRES2016/_PDF/IPR16.Proceedings_4_Web_Broschuere_Link.pdf
14 https://e-science-tage.de

between EPFL and ETH Zurich with the creation of the newly founded Swiss Data Science Center (SDSC) in 2017: https://datascience.ch/

Acknowledgments

We are very grateful to our respective colleagues, institutions and financial partners, including the DLCM project and swissuniversities, for providing us with the necessary means and support to accomplish these new endeavours in a constructive, innovative and collaborative way. We wish to especially express our thanks to the following persons, without whom we would not have been able to meet these new sets of challenges: Isabelle Kratz, Matthias Töwe, Eliane Blumer, Jan Krause, Gabi Schneider, Pierre-Yves Burgi, René Schneider, Nathalie Lambeng, Lorenza Salvatori, Pascale Bouton, Béatrice Marselli, Guilaine Baud-Vittoz, and the entire DLCM team.

References

Burgi, Pierre-Yves, Eliane Blumer, and Basma Makhlouf-Shabou. 2017. "Research Data Management in Switzerland." *IFLA Journal*, January, 34003521667823. doi:10.1177/0340035216678238.

Corti, Louise, Veerle Van den Eynden, Libby Bishop, and Matthew Woollard. 2014. "*Managing and Sharing Research Data: A Guide to Good Practice"*. London, UK: SAGE Publications Ltd. doi:https://uk.sagepub.com/en-gb/eur/managing-and-sharing-research-data/book240297.

EC - European Commission. European Research Area. n.d. 2016. "Open Research in Horizon 2020." http://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf#view=fit &pagemode=none.

Goodman, Alyssa, Alberto Pepe, Alexander W. Blocker, Christine L. Borgman, Kyle Cranmer, Merce Crosas, and Rosanne Di Stefano. 2014. "Ten Simple Rules for the Care and Feeding of Scientific Data." *PLoS Comput. Biol.* 10 (4). e1003542. doi:10.1371/journal.pcbi.1003542.

Higman, Rosie, Marta Teperek, and Danny Kingsley. 2017. "Creating a Community of Data Champions." *bioRxiv*. http://biorxiv.org/content/early/2017/02/20/104661.abstract.

Piwowar, H A, and T J Vision. 2013. "Data Reuse and the Open Data Citation Advantage." *PerrJ Comput. Sci.* doi:10.7717/peerj.175.

Piwowar, Heather A., Roger S. Day, and Douglas B. Fridsma. 2007. "Sharing Detailed Research Data Is Associated with Increased Citation Rate." *PLoS ONE* 2 (3). doi:10.1371/journal.pone.0000308.

Sesartic, Ana, and Matthias Töwe. 2016. "Research Data Services at ETH-Bibliothek." *IFLA Journal* 42 (4): 284–91. doi:10.1177/0340035216674971.

Service durch Kompetenzbündelung – Das institutionelle Konzept zum Forschungsdatenmanagement der Leibniz Universität Hannover

Anneke Meyer[1], Janna Neumann[2], Volker Soßna[3]

1 Leibniz Universität Hannover
2 Technische Informationsbibliothek, Hannover
3 Leibniz Universität Hannover

Zusammenfassung. Die Leibniz Universität Hannover hat den bedarfsgerechten Auf- und Ausbau des Unterstützungsangebots zum Umgang mit Forschungsdaten als strategisches Ziel definiert, um den eigenen Forschungsstandort zu stärken. Fachpersonal aus dem Dezernat Forschung, den Leibniz Universität IT Services (LUIS) und der Technischen Informationsbibliothek (TIB) haben dazu ein institutionelles Konzept entworfen, das seit Dezember 2016 umgesetzt wird. Ausgangspunkt des Konzepts bildete eine Umfrage zum Umgang mit Forschungsdaten an der Leibniz Universität Hannover, die durch qualitative Interviews ergänzt wurde.
Das institutionelle Konzept umfasst folgende Elemente:

– Etablierung einer Policy zum Umgang mit Forschungsdaten für die gesamte Universität
– Beratung und Schulung für Wissenschaftlerinnen und Wissenschaftler und die Service-Einrichtungen
– Auf- und Ausbau eines institutionellen Datenrepositoriums und Entwicklung von Schnittstellen zum Forschungsinformationssystem und zum Volltextrepositorium
– Universitätsübergreifende Kooperation & Vernetzung

Die vier Elemente befinden sich in einem unterschiedlichen Umsetzungsstand. Bereits seit 2014 führen die beteiligten Institutionen gemeinsam Beratungen und Schulungen durch und nutzen dafür zur Qualitätssicherung und gegenseitigen Information gemeinsame Dokumentationssysteme. In diesem Bereich konnten in den letzten zwei Jahre Erfahrungen gesammelt werden und Prozesse entsprechend optimiert werden.

Die Herausforderung des Ansatzes an der Leibniz Universität besteht darin, ein einrichtungsübergreifendes Service-Angebot vorzuhalten und kollaborativ weiter zu entwickeln. Dadurch ist gewährleistet, dass Kompetenzen effektiv gebündelt werden und sich keine Parallelstrukturen an einzelnen Einrichtungen bilden. Durch die gemeinsam entwickelten Services werden Wissenschaftlerinnen und Wissenschaftler mit einer Stimme und auf mehreren Ebenen zum aktiven und bewussten Umgang mit Forschungsdaten angeregt.

In diesem Artikel werden die ersten Erfahrungen in der Umsetzung der einzelnen Elemente des institutionellen Konzepts sowie in der Zusammenarbeit beleuchtet. Außerdem wird ein Ausblick auf die zukünftig angestrebte Entwicklung gegeben.

Schlagwörter. Forschungsdatenmanagement, Beratung, Schulungen, institutionelles Konzept

Forschungsdatenmanagement als strategisches Ziel

Der freie Zugang zu Daten wird forschungspolitisch immer stärker eingefordert (vgl. RfII 2016). Daher werden auch die Drittmittelgeber diese Forderung nach und nach konsequenter in ihren Anforderungskatalog für eine Förderung aufnehmen.

War es bisher ausreichend, sich an die Regeln der guten wissenschaftlichen Praxis (vgl. DFG 2013) zu halten und Daten mindestens 10 Jahre aufzubewahren, wird jetzt zunehmend gefordert, dass Daten nicht nur konserviert, sondern auch öffentlich zugänglich gemacht werden. So hat die Deutsche Forschungsgemeinschaft beispielsweise im Jahr 2015 Leitlinien zum Umgang mit Forschungsdaten (vgl. DFG 2015) veröffentlicht. Das europäische Förderprogramm Horizon 2020 (European Commission 2016) fordert zudem die Teilnahme geförderter Projekte am Open Research Data Pilot mit der Möglichkeit, jederzeit unter Angabe einer Begründung auszusteigen (European Commission 2016).

Die Umsetzung dieser Forderungen oder Möglichkeiten ist aber nicht trivial, sondern wirft für die einzelne Forscherin oder den einzelnen Forscher unter anderem diese Fragen auf: Welche meiner Daten sind veröffentlichungswürdig? Wer kommt für die Zusatzkosten auf? Welchen Vorteil habe ich von einer Veröffentlichung meiner Daten?

Um die Spannung zwischen Forderungen der Politik beziehungsweise der Drittmittelgeber und den berechtigten Fragen der Forschenden aufzulösen, bedarf es eines institutionellen Rahmens, der auch individuellen Bedürfnissen gerecht wird. Dieser Rahmen bestehend aus Richtlinien und Unterstützungsangeboten muss aktiv an die Forschenden kommuniziert werden.

Die Leibniz Universität Hannover hat sich das strategische Ziel gesetzt, ihre Forschenden in allen Belangen des Datenmanagements zu unterstützen. Sie ist sich daher der Bedeutung eines institutionellen Rahmens für gutes Forschungsdatenmanagement bewusst.
Konkret sollen diese Ziele verfolgt werden:

- Institute verabschieden fach- und institutsspezifischer Richtlinien zum Umgang mit Forschungsdaten,
- das Kompetenzniveau des wissenschaftlichen Personals beim Umgang mit Forschungsdaten steigt und das Bewusstsein für die Vorteile eines guten Forschungsdatenmanagements wird aktiv gefördert,
- die Anzahl der Open Access-Veröffentlichungen von Forschungsdaten steigt,
- durch überzeugende Forschungsdatenmanagement-Konzepte in Drittmittelanträgen wird die Erfolgsquote bei der Einwerbung von Drittmitteln erhöht.

Diese Ziele sollen durch eine universitätsweite Policy, Beratung & Schulung und ein Repositorium für die Datenpublikation erreicht werden. Außerdem müssen sich auch die beteiligten Service-Einheiten vernetzen und miteinander kooperieren, um eine qualitativ hochwertige Weiterentwicklung des gesamten Angebots garantieren zu können.

Das Projekt an der Leibniz Universität Hannover

Das Projekt zur Konzepterstellung eines institutionellen Forschungsdatenmanagement an der Leibniz Universität Hannover ist im Jahr 2014 mit einer Umfrage und anschließenden qualitativen Interviews mit ausgewählten Teilnehmern gestartet. Die Auswertung (Hauck et al. 2016) hat ergeben, dass folgende Unterstützungsangebote zum Forschungsdatenmanagement vom wissenschaftlichen Personal der Leibniz Universität benötigt werden:

- Regelungen und Handlungsempfehlungen
- Beratung zu allgemein Fragen sowie speziell zu rechtlichen und technischen Aspekten
- Schulungen zum Umgang mit Forschungsdaten
- IT-Infrastruktur zum Austausch sowie zur Archivierung und Publikation von Daten

Entsprechende Angebote standen bisher entweder noch nicht in ausreichendem Umfang zur Verfügung oder wurden kaum in Anspruch genommen, da sie zu wenig bekannt waren. Es entstanden Reibungsverluste im wissenschaftlichen Alltagsbetrieb, zum Beispiel weil Daten unzureichend gesichert und dokumentiert wurden und deshalb verlorengingen oder unbrauchbar wurden. Gleichzeitig wandten die Forschenden viel Zeit auf, um sich eigenständig Know-how anzueignen und in kleinem Maßstab eigene Speicherstrukturen zu betreiben. Durch einen Ausbau des Beratungs- und Schulungsangebots sowie die Erweiterung der IT-Infrastruktur um ein Datenrepositorium sollen solche Reibungsverluste zukünftig minimiert werden.

Ein weiteres Ergebnis war, dass die verbesserten Unterstützungsangebote proaktiv an die Forscherinnen und Forscher kommuniziert werden müssen, damit sie Teil der Routine-Abläufe des Wissenschaftsbetriebs werden können. Daher richtet die Leibniz Universität Hannover das Augenmerk gezielt auf Forschungsprojekte in der Antrags- oder Abschlussphase und darauf, dass Schulungen zum Umgang mit Forschungsdaten Teil der Ausbildung des wissenschaftlichen Nachwuchses werden.

Im Folgenden werden kurz die einzelnen Elemente des Konzepts sowie der Umsetzungsstand vorgestellt.

Policy

Eine universitätsweit gültige Policy zum Umgang mit Forschungsdaten wird voraussichtlich im Mai 2017 verabschiedet. Sie wird zukünftig den grundlegenden Handlungs- und Orientierungsrahmen für alle Einrichtungen und Forschenden der Leibniz Universität bilden. Der Text orientiert sich an bereits bestehenden Richtlinien anderer Universitäten[1] und umfasst folgende Punkte:

- Definition des Begriffs „Forschungsdaten“
- Empfehlung zum offenen Umgang mit Forschungsdaten über die Grundsätze der guten wissenschaftlichen Praxis hinaus (unter Berücksichtigung rechtlicher Rahmenbedingungen)
- Empfehlung zur Erstellung von instituts- oder projektspezifischen Richtlinien zum Umgang mit Forschungsdaten. Darin können zum Beispiel Rollen und Verantwortlichkeiten, Workflows und Konventionen für die Dateiablage festgelegt werden.
- Angebote zur Unterstützung des wissenschaftlichen Personals beim Forschungsdatenmanagement

Für spezifischere Handlungsempfehlungen werden noch ergänzende Leitfäden und Merkblätter ausgearbeitet sowie vertiefende Informationen auf den Webseiten zum Forschungsdatenmanagement[2] der Leibniz Universität angeboten.

Das Service-Team hat es sich zum Ziel gesetzt, auf Grundlage der allgemeinen Policy möglichst viele Institute und größere Verbundprojekte proaktiv auf die Vorteile einer detaillierten internen Policy anzusprechen. Es ist geplant, die zukünftigen Sprecherinnen und Sprechern von Sonderforschungsbereichen und Graduiertenkollegs bereits in der Antragsphase davon zu überzeugen, das Verfassen einer eigenen Policy in den Antrag mit aufzunehmen.

1 Ein Verzeichnis von Forschungsdatenpolicies deutscher Forschungsinstitutionen ist unter http://www.forschungsdaten.org/index.php/Data_Policies#Institutionelle_Policies zu finden.

2 www.fdm.uni-hannover.de

Beratung und Schulung

Beratung und Schulung hilft den Forschenden, konkrete Herausforderungen im Umgang mit Daten zu bewältigen. Gleichzeitig wird das Bewusstsein gestärkt, dass Datenmanagement und die Publikation von Daten die Arbeit der Forscherinnen und Forscher effizienter und sichtbarer macht. An der Leibniz Universität werden daher die Beratungskompetenzen im Forschungsdezernat, in der Technischen Informationsbibliothek (TIB) und bei den Leibniz Universität IT Services (LUIS) in einer virtuellen Beratungseinheit gebündelt (Abb. 1). Die einzelnen Mitglieder dieses Beratungsteams haben unterschiedliche Kompetenzschwerpunkte, die sich gegenseitig ergänzen. Sie treffen sich regelmäßig zu Arbeits- und Abstimmungstreffen, führen zusammen Schulungen durch und dokumentieren Beratungen in einem gemeinsamen, internen Wiki.

Der Datenreferent im Dezernat Forschung ist verantwortlich für die Koordination der Service-Einheit. Durch die verteilte Beratungsstruktur werden auch die Wissenschaftlerinnen und Wissenschaftler erreicht, die außerhalb der drittmittelfinanzierten Projekte Forschung betreiben.

Abbildung 1. An der Leibniz Universität Hannover werden Beratungskompetenzen zum Umgang mit Forschungsdaten einrichtungsübergreifend gebündelt. (Grafik: Volker Soßna).

Das Fachpersonal für Forschungsdatenmanagement an den beteiligten Einrichtungen kann sowohl individuell als auch über die gemeinsam genutzte Funktions-E-Mail-Adresse[3] kontaktiert werden. Durch die Dokumentation von Beratungen im gemeinsamen Wiki und durch regelmäßige Besprechungen informieren sich die Mitglieder des FDM-Teams gegenseitig über eingegangene Beratungsanfragen. Detailfragen werden jeweils von den Personen beantwortet, die entsprechende Spezialkenntnisse haben. Auf diese Weise wird sichergestellt, dass alle Beraterinnen und Berater

3 forschungsdaten@uni-hannover.de

in den unterschiedlichen Einrichtungen über die vorherigen Beratungsinhalte informiert sind. Auskunft Suchende müssen ihr Anliegen nicht erneut schildern, wenn sie an eine weitere Fachberaterin oder Fachberater verwiesen werden.

Beratungsschwerpunkte im Dezernat Forschung

In den Workflow der bereits bestehenden Antragsberatung im Forschungsdezernat wird die Beratung zum Forschungsdatenmanagement systematisch eingebunden, so dass sichergestellt ist, dass Forschende sowohl in den Arbeitsplänen als auch in der Kostenplanung die Ansätze für das Datenmanagement in die Anträge integrieren. Im Mittelpunkt stehen dabei die Graduiertenkollegs und Sonderforschungsbereiche der Deutschen Forschungsgemeinschaft (DFG) sowie größere EU- oder Bundesprojekte.

Beratungsschwerpunkte in der Technischen Informationsbibliothek (TIB)

Kompetenzen zur Publikationsberatung existieren bereits im Bereich der Publikationsdienste der TIB. Dieser Bereich wird um die Beratung zum Forschungsdatenmanagement und zur Datenpublikation erweitert. Hierunter fällt unter anderem die Beratung zu Fragestellungen bzgl. der Schutzfähigkeit von Daten (Urheberrecht, Eigentumsverhältnisse) und der Wahl geeigneter (Nachnutzungs-) Lizenzen.

Beratungsschwerpunkte in den Leibniz Universität IT-Services (LUIS)

Die IT-Services erbringen Beratungsleistungen zu technischen Aspekten des Forschungsdatenmanagement. Fragen beziehen sich hier häufig auf die Einbindung von Diensten der LUIS in Forschungsvorhaben und in die Workflow von Arbeitsgruppe. Beim Thema Langzeitarchivierung treten Fragen zu geeigneten Dateiformaten, zur Dokumentation von Forschungstätigkeiten sowie zu Anforderungen an die Informationssicherheit (Vertraulichkeit, Verfügbarkeit und Integrität von Daten) auf.

Schulungen

Schulungen werden bislang vor allem über das Weiterbildungsprogramm und über die Graduiertenakademie der Leibniz Universität als halbtägige Workshops angeboten. Auf diese Weise werden sowohl allgemein alle Forschenden als auch speziell die Promovierenden angesprochen. Die Themen erstrecken sich von allgemeinen Einführungskursen bis hin zu speziellen Themen wie Dateibenennung und Dateiablage, Datenmanagementpläne oder auch Metadaten und Datenpublikation.

Für die Workshops haben sich Gruppengrößen von 10-15 Personen als ideal erwiesen, um sowohl angemessen auf Fragen eingehen zu können als auch Arbeit in Kleingruppen zu gewährleisten. Seit 2014 werden regelmäßig mindestens einmal im Jahr einführende allgemeine Workshops im Weiterbildungsprogramm der Universität angeboten. Hinzu kommt seit 2016 ein weite-

rer Kurs im Qualifizierungsprogramm der Graduiertenakademie. Im März 2017 fand erstmals ein vertiefender Workshop zu einem speziellen Themenbereich. Solche Vertiefungsworkshops finden zukünftig mindestens einmal pro Jahr statt. Die Resonanz zu den Einführungsworkshops ist in den letzten zwei Jahren deutlich angestiegen (von sechs Teilnehmenden in 2014 zu 14 Teilnehmenden in 2016). Bei den vertiefenden Workshops ist das Interesse und damit die Teilnehmerzahl abhängig vom jeweiligen Themenschwerpunkt. Des Weiteren wurden maßgeschneiderte Informationsveranstaltungen für einzelne Institute durchgeführt. Für 2018 ist es geplant, auch einen Kurs über die Hochschulübergreifende Weiterbildung Niedersachsen anzubieten, um den Wirkungskreis zu erweitern.

Nach den Schulungen wird regelmäßig über einen kurzen Fragebogen ein Feedback der Teilnehmenden eingeholt. So kann überprüft werden, ob die Unterstützungsangebote zum Forschungsdatenmanagement den Bedürfnissen der Forschenden gerecht werden. Durch sehr heterogene Gruppenzusammenstellungen und damit zusammenhängende unterschiedliche Vorkenntnisse und Erwartungen konnten zwar nicht immer alle Bedürfnisse gleich gut abgedeckt werden, so dass einige Themen dem einen zu lang, dem anderen Teilnehmer jedoch zu kurz behandelt wurden. Die Rückmeldungen zu den bisherigen Schulungen sind dennoch überwiegend positiv ausgefallen. Grundsätzlich werden die Schulungen so konzipiert, dass auch angemessen auf individuelle Fragen eingegangen werden kann.

Auf der FDM-Webseite der Leibniz Universität werden dazu ergänzende Informationsangebote bereitgestellt wie beispielsweise Leitfäden und Merkblätter, Links zu externen Ressourcen, ein Glossar und FAQs zum Forschungsdatenmanagement sowie Schulungsunterlagen und Informationen zu den angebotenen Kursen.

IT-Infrastruktur

Die vorhandenen IT-Dienste und -Infrastrukturen für die Speicherung, Übertragung, Verarbeitung und Archivierung von Forschungsdaten werden derzeit erweitert, verbessert und aufeinander abgestimmt. Zusätzlich wird ein institutionelles Datenrepositorium (auf Basis von CKAN[4]) aufgebaut, in dem spätestens ab 2018 Daten im Open Access werden können.[5] Dieses Angebot richtet sich insbesondere an Forschende derjenigen Disziplinen, für die noch keine etablierten Fachrepositorien existieren.

Das Datenrepositorium soll über geeignete Schnittstellen auch mit anderen IT-Diensten der Leibniz Universität kommunizieren können. So wird aktuell ein Forschungsinformationssystem (PURE[6]) eingeführt, in dem grundlegende Informationen über alle Forschungsaktivitäten und -ergebnisse und damit auch die Datenpublikationen der Leibniz Universität erfasst und miteinander verknüpft werden.[7] Seit Ende 2015 verfügt die Leibniz Universität zudem über ein institutionelles Volltext-Repositorium (auf Basis von DSpace[8]), das die Veröffentlichungen von Textpublikationen und Zweitveröffentlichungen von Mitarbeiterinnen und Mitarbeitern der Universität im Open

4 https://ckan.org/
5 Die Leibniz Universität unterstützt den freien Zugang zu Wissen. Sie hat eine eigene Resolution zu Open Access verabschiedet, siehe Leibniz Universität Hannover (2011)
6 https://www.elsevier.com/solutions/pure
7 Informationen zum Aufbau eines Forschungsinformationssystems an der Leibniz Universität Hannover: https://www.dezernat4.uni-hannover.de/fis.html
8 http://www.dspace.org/

Access ermöglicht.[9] Zwischen Forschungsinformationssystem, Text- und Datenrepositorium soll künftig der automatisierte Austausch von Metadaten möglich sein, damit Forschende diese nicht doppelt eintragen müssen. Das übergreifende Ziel ist eine Verknüpfung von Personen, Projekten, Text- und Datenpublikationen. Im Frühjahr 2017 beginnt die Testphase für das Datenrepositorium, das spätestens 2018 vollumfänglich genutzt werden kann.

Kooperation & Vernetzung

Die drei beteiligten Einrichtungen arbeiten eng zusammen und stimmen sich bei den Arbeitsschritten untereinander ab. Auf diese Weise wird gemeinsames Wissen aufgebaut und gemeinsam Erfahrungen gesammelt, die auf die weitere Entwicklung des institutionellen Konzepts Einfluss haben.

Darüber hinaus orientiert sich das FDM-Team an der Leibniz Universität an erfolgreichen nationalen und internationalen Vorbildern. Um einen regelmäßigen Austausch von Know-how und Erfahrung zu gewährleisten, engagieren sich seine Mitglieder in verschiedenen Netzwerken, Gremien und Beiräten, um aktuelle Entwicklungen schnell aufgreifen zu können.[10] Durch hochschulübergreifende Kooperationen kann zudem ein Synergie-Effekt bei der Entwicklung von Tools und Methoden erzielt werden. So können Entwicklungen, wie beispielsweise existierende Softwaretools für den Entwurf von Datenmanagementplänen, in der eigenen Institution nachgenutzt und weiterentwickelt werden.

Auf europäischer Ebene wird ein Austausch mit Institutionen angestrebt, die bereits über ausgearbeitete Dienste zum Umgang mit Forschungsdaten verfügen und mehrjährige Erfahrung mit deren Betrieb gesammelt haben. So kann sich die Leibniz Universität Hannover nicht nur an erfolgreichen Vorbildern orientieren sondern sich gleichzeitig auch international weiter vernetzen. Es bestehen bereits Kontakte zu den britischen Universitäten in Lancaster und St Andrews. Durch die Teilnahme an Fortbildungen des Digital Curation Centers in Edinburgh konnten wichtige Kenntnisse und Anregungen für das Service-Angebot zum Forschungsdatenmanagement an der Leibniz Universität Hannover gewonnen und neue Kontakte geknüpft werden.

Kommunikationsstrategien

Viele Wissenschaftlerinnen und Wissenschaftler sehen die Aufbereitung, Dokumentation und Publikation von Daten als zeitraubende Zusatzarbeit an, die sie möglichst auf ein Minimum beschränken wollen. Die Leibniz Universität Hannover strebt daher einen Kulturwandel hin zu einem umfassenden und professionellen Forschungsdatenmanagement an. Eine Herausforderung

9 http://www.repo.uni-hannover.de/

10 Dazu gehören unter anderem:

- die gemeinsame AG Forschungsdaten der Deutschen Initiative Netzwerkinformation e.V. (DINI) und des Kompetenzzentrums für Langzeitarchivierung nestor
- das Netzwerks der deutschen Forschungs- und Technologiereferentinnen und –referenten
- der Landesarbeitskreis Niedersachsen für Informationstechnik (LANIT)
- die Schwerpunktinitiative „Digitale Information“ der Allianz der deutschen Wissenschaftsorganisationen.
- AG Informationsinfrastruktur der Landeshochschulkonferenz Niedersachsen

besteht daher darin, den Forschenden den Nutzen des Forschungsdatenmanagements für ihre wissenschaftliche Arbeit deutlich zu machen. Vorgaben seitens der Förderer und der Wissenschaftsorganisationen haben in Deutschland häufig nur empfehlenden Charakter. Sie erzeugen daher derzeit bei den Forschenden nur wenig Handlungsdruck. Die Anforderung an das Beratungs- und Schulungsteam ist es demnach, zu vermitteln, dass ein gutes Forschungsdatenmanagement langfristig viel Zeit und Ressourcen sparen und zudem die Qualität der Auswertung verbessern kann.

Gleichzeitig müssen die vorhandenen Unterstützungsangebote potentiellen Nutzerinnen und Nutzern bekannt gemacht werden. Eine Webseite, Flyer, Leitfäden und andere vom Team erarbeitete Informationsmaterialien ermöglichen ein selbständiges Informieren. Auf diese Weise können jedoch erfahrungsgemäß nicht alle Interessierten erreicht werden. Daher werden Graduiertenkollegs und Sonderforschungsbereiche zusätzlich gezielt proaktiv angesprochen, um ihnen konkrete Vorschläge zum Forschungsdatenmanagement zu unterbreiten. Diese Projekte können langfristig als Multiplikatoren wirken und damit einen hohen Einfluss auf die Doktorandenausbildung und auf die Forschungstätigkeit allgemein ausüben. Dem Forschungsservice kommt eine wichtige Rolle zu, da an dieser Einrichtung besonders enger Kontakt zu zahlreichen Projekten besteht.

Fazit und Ausblick

Die vier Elemente des institutionellen Konzepts sollen kontinuierlich weiter entwickelt werden. Erste Erfolge zeigen sich in gestiegenen Teilnehmerzahlen bei den Schulungen, in regelmäßigen Anfragen von Instituten und Projekten nach spezifischen Workshops und Beratung sowie in der systematischen Einbindung der Forschungsdatenmanagement-Beratung in den Beratungsablauf im Forschungsdezernat. Die ersten Pilotkunden für das Repositorium sind bereits gefunden. Die Policy soll in Kürze vom Senat der Leibniz Universität dem Präsidium zur Verabschiedung empfohlen werden.

Parallel erarbeitet das Projektteam auf Wunsch des Präsidiums der Leibniz Universität ein Geschäftsmodell für die Beratung und Schulung sowie für das Repositorium. Dieses Modell ist Voraussetzung dafür, dass die Services zum Forschungsdatenmanagement der Leibniz Universität zukünftig auch von Angehörigen kleinerer niedersächsischer Universitäten genutzt werden können, für die der Aufbau eigener Strukturen aufgrund ihrer Größe unwirtschaftlich wäre. Eine Herausforderung des Konzepts besteht darin, die Kommunikationsstrategien zu verbessern und noch stärker daraufhin zu arbeiten, einrichtungsübergreifend gemeinsame Strukturen aufzubauen. Dafür bedarf es des fortgesetzten Engagements der beteiligten Institutionen an der Leibniz Universität, um gemeinsam Arbeitsergebnisse zu erzielen und keine Parallelstrukturen entstehen zu lassen.

Darüber hinaus wird ein Kulturwandel in der Universität insgesamt angestrebt. Forschungsdatenmanagement soll nicht länger als Spezialgebiet von Einzelpersonen angesehen werden. Vielmehr sollen vertiefte Kenntnisse auf diesem Gebiet zunehmend als selbstverständliche Schlüsselkompetenz im Wissenschaftsbetrieb betrachtet werden, und zwar sowohl für die Forschenden als auch für das unterstützende Personal in Technik und Verwaltung. Perspektivisch ist es daher wichtig, dass nicht nur die Forscherinnen und Forscher zum Umgang mit Forschungsdaten geschult werden, sondern auch das weitere Personal in den beteiligten Einrichtungen der Leibniz Universität.

Literaturangaben

RfII - Rat für Informationsinfrastrukturen. 2016. „*Leistung aus Vielfalt*“. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland. Göttingen. Online verfügbar unter urn:nbn:de:101:1-201606229098, zuletzt geprüft am 13.03.2017.

Deutsche Forschungsgemeinschaft. 2013. *Sicherung Guter Wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft“*, Wiley-VCH, ergänzte Auflage 2013. Online verfügbar unter http://doi.org/10.1002/9783527679188.oth1 .

Deutsche Forschungsgemeinschaft. 2015. *Leitlinie zum Umgang mit Forschungsdaten*. Online verfügbar unter http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf, zuletzt geprüft am 13.03.2017.

European Commission. 2016. *“Horizon 2020 Programme, Guidelines on FAIR Data Management in Horizon 2020”*. Online verfügbar unter http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf, zuletzt geprüft am 13.03.2017.

European Commission. 2016. *“Horizon 2020 Programme, Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020”*. Online verfügbar unter http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf, zuletzt geprüft am 13.03.2017.

Hauck, Reingis, Reiko Kaps, Hans Georg Krojanski Anneke Meyer, Janna Neumann und Volker Soßna. 2016. *Der Umgang mit Forschungsdaten an der Leibniz Universität Hannover: Auswertung einer Umfrage und ergänzender Interviews 2015/16*. (Hannover: Institutionelles Repositorium der Leibniz Universität Hannover). Online verfügbar unter https://doi.org/10.15488/265, zuletzt geprüft am 13.03.2017.

Leibniz Universität Hannover. 2011. *„Open-Access-Resoution“*. Online verfügbar unter https://www.uni-hannover.de/fileadmin/luh/content/webredaktion/universitaet/ziele/open-access-resolution.pdf. zuletzt geprüft am 13.3.2017

RADAR: A Research Data Management Repository for Long Tail Data

Ena Brophy[1], Matthias Razum[2]

1,2 FIZ Karlsruhe – Leibniz Institute for Information Infrastructure

Abstract. The transparency and reproducibility of scientific results are increasingly based on digital data. In compliance with good scientific practice, data needs to be published, accessible, and re-usable. RADAR is a generic infrastructure providing archival and publication services for research data. RADAR focuses on the "long tail of science", which often lacks sufficient research data infrastructure. It offers two service levels: data archival ("dark archive" with variable retention periods and user-defined access rights) and data archival with publication (guaranteed retention period of 25+ years, DOI assignment, and flexible licensing options).

Users can upload, edit, structure and describe (collaborative) data in an organisational workspace. Administrators and curators can manage access and editorial rights before the data enters the preservation and optional publication level. Data consumers may search, access, download and retrieve usage statistics on the data via the RADAR portal. For data consumers, findability of research data is of utmost importance. The metadata of published datasets can be harvested via a local OAI provider or the DataCite Metadata Store. Additionally, RADAR provides an application programming interface (API) for easy integration of RADAR functionality with existing systems and workflows at the user's side.

RADAR relies on two academic data centres under German jurisdiction to provide its services. The novel two-stage service and business model provides one-off payments and institutional subscription services. RADAR is intended to become an integral part of the international information infrastructure which also allows the integration of third-party services. RADAR was designed by a research consortium of academic institutions for the academic community. Through cooperation with researchers, data centres, scientific societies as well as publishers, RADAR ensures that the resulting infrastructure is designed to meet the requirements of the academic community.

Keywords: RADAR, research data repository; repository; preservation; information infrastructure; research data management; data archiving; data publication.

Introduction: Data Management for the Long Tail

The collection and organisation of data is a fundamental element of the research process. In compliance with good scientific practice, data needs to be published, accessible, and re-usable. Digital data offer the potential for greater return on investment, provided that data is properly managed and shared among researchers (Berman, et al. 2010, Buckland 2011). The academic community is becoming more interested in collecting and providing access to datasets produced at their institution for reuse. Driving this is the transparency and reproducibility of scientific results which is recognised as a primary research output based on digital data (Treloar und Harboe-Ree 2008, Klump 2009, Neuroth, et al. 2012). While the focus has been on the accessibility of 'big data', i.e. disciplines whose output produces large volumes of data, many research studies produce smaller

datasets. This poses a challenge to the academic community who needs to manage and sustain access to research data that does not necessarily fall within the scope of discipline-based solutions. A survey conducted by the journal Science in 2011 stated that 48.3% of respondents were working with datasets that were less than 1GB in size, and over half of respondents reported that they stored their data only in their laboratories (Science Editorial 2011). Solutions may differ from discipline to discipline in size, scale, project duration etc. (C. L. Borgman 2015, Borgman, et al. 2015). This is true in particular for long tail data which often lacks sufficient research data infrastructure. Best practice for the data management of long tail data is often dependent on the community.

What has emerged in the last number of years for both big and long-tail data is the need for being open to be searched, cited and downloaded for potential reuse. Funders such as the National Science Foundation require researchers to include a data management plan as part of their proposal for funding (National Science Foundation 2011). In Germany, the German Research Foundation published guidelines for "Safeguarding Good Scientific Practice" to ensure that data produced as part of scientific studies are recognised as primary research output (DFG 2013). In 2016, stakeholders from academia, industry, publishers and funding agencies published a concise and measurable set of principles called the FAIR Data Principles (Findable, Accessible, Interoperable and Re-usable). These principles place specific emphasis on enhancing the ability of machines to automatically find and reuse data (Wilkinson 2016, FORCE11 2016). To highlight the importance of keeping data FAIR, the European Commission adopted the FAIR Data Principles and released new Guidelines on FAIR Data Management in Horizon 2020 (European Commission 2016). The EC guidelines include several important changes that aim to improve the quality of project results, achieve greater efficiency, and achieve progress and growth of a transparent scientific process. Consequently, research institutions, universities and libraries are becoming more interested in collecting and providing access to datasets produced at their institution that do not fall within the scope of big data or discipline-based repositories. In addition, researchers themselves start to look for data services. This paper presents a multidisciplinary solution - the RADAR (Research Data Repository) service[1], a generic research data repository for data preservation and publication in research to include the social sciences and humanities.

The RADAR Service

RADAR was developed as a cooperation project of five research institutes from the fields of natural and information sciences[2]. The technical RADAR infrastructure is provided by the FIZ Karlsruhe – Leibniz Institute for Information Infrastructure and the Steinbuch Centre for Computing (SCC), Karlsruhe Institute of Technology (KIT). The sustainable management and publication of research data with DOI-assignment was provided by the German National Library of Science and Technology (TIB). The Ludwig-Maximilians-Universität Munich (LMU), Faculty for Chemistry and Pharmacy and the Leibniz Institute of Plant Biochemistry (IPB) provided the scientific knowledge and specifications and ensure that RADAR services can be implemented to become part of the scientific workflow of academic institutions and universities.

1 RADAR (Research Data Repository): https://www.radar-service.eu
2 RADAR Project website: https://www.radar-projekt.org/display/RD/Home

The heterogeneity of research data is an important issue for the research community. Therefore the primary goal of RADAR was to establish an interdisciplinary research data repository, which is sustained by research communities and supported by a stable business model. RADAR has approached this problem by focusing on real scientific workflows and established best practice throughout the testing phase of product development. During the project phase, a number of public workshops were held to gather requirements and discuss technical, organisational and legal matters. Furthermore, a scientific advisory board was established. A wealth of requirements, feedback, and advice was collected via both channels. After the first two years of the project a test system was launched enabling the academic community to test RADAR using datasets from different subject areas. This approach provided the project team with significant insight, which allowed them to design the service with the academic community at its centre to ensure the service will be effective.

The upload of scientific data into repository collections is a continuing challenge for researchers and their affiliated research institutions. To facilitate this RADAR provides a generic infrastructure, which delivers archival and publication services for research data. RADAR offers a suite of services to ensure that the requirements of funding agencies and good scientific practice are met.

Basic service: Data Preservation

For data providers, RADAR offers format-independent preservation services to store data in compliance with specific institutional or funder requirements periods (e.g., 10 years according to DFG recommendations). This includes secure preservation of up to 15 years with the data remaining unpublished, and the requirement of a minimum set of metadata. By default, the data and associated metadata will not be published, unless specified otherwise by the data provider. RADAR offers a flexible data and metadata access management so that data providers are able to share preserved datasets with other RADAR users if desired and manage the external visibility of the associated metadata.

Extended Service: Data Publication

For making data citable, traceable and reusable, RADAR offers a combined service of research data archival and publication. Datasets published in RADAR are identified by DOI. Using the DOI, datasets can be referenced persistently and unambiguously. The service also includes an optional embargo period for the publication of submitted data that may be subsequently prolonged if necessary. The metadata describing the dataset is published already during the embargo and datasets are allocated a DOI. This ensures that datasets can be found and cited already when they are deposited, while downloads will only be possible once the embargo period has expired. Within the publication service, a peer review option may be used: In this case, the respective dataset is "frozen" for the duration of the peer review process and receives a secure "review-URL" provided by RADAR which may be forwarded to an editor or reviewer responsible for a corresponding manuscript submission. As such, manuscript and data may be inspected simultaneously during a review process.

Architecture

The system architecture is based on an expendable API structure, referred to as 'Computing Centre API' in Fig. 1. This structure allows an integration of multiple computing centres that use various storage systems (e.g. TSM, SamQFS, DMS, HPSS). To reach a uniform archiving interface, the API hides these various storage systems and technologies. The storage is managed by using a repository software which consists of two parts. A back end addresses general tasks such as storage access and bitstream preservation, whereas the front end implements RADAR-specific workflows. Front-end workflows include various data services: Metadata management, access control, data ingest processes, as well as the licensing for reuse and publishing of research data with DOI. Archival Information Packages and Dissemination Information Packages are provided in a BagIt-structure in ZIP format. The RADAR API enables users to integrate the archival backend into their own systems and workflows. RADAR stores the data in two academic data centres under German jurisdiction. The Steinbuch Centre for Computing (SCC) at Karlsruhe Institute of Technology (KIT) acts as the primary data centre, holding two copies of the data at two different locations. A third copy is replicated to the Zentrum für Informationsdienste und Hochleistungsrechnen at TU Dresden. The metadata catalogue and the software is hosted by SCC. The two data centres employ different hardware and software systems as well as differing administrative procedures. This approach adds an additional level of security and helps avoiding systematic errors that may put corrupt large chunks of the archived data.

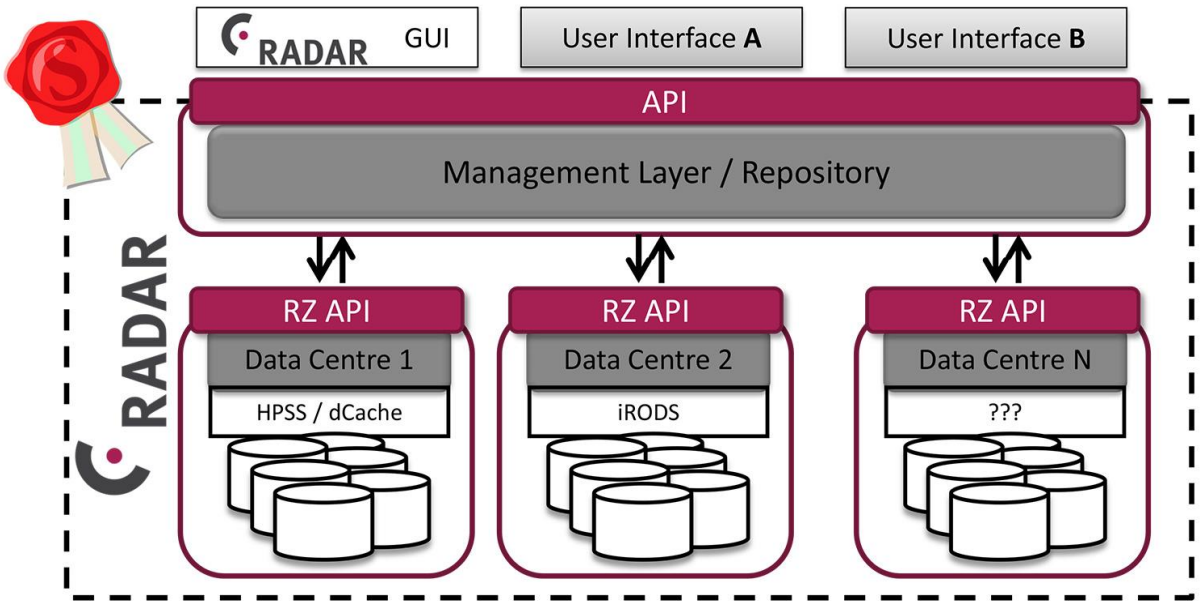


Figure 1. RADAR architecture with data ingest and API (Brophy & Razum 2017)

TIB Hannover provides as German DataCite agency the necessary systems and interfaces for registering DOIs assigned to published data sets. All communication between RADAR and DataCite is handled via REST calls.

Metadata Schema

Metadata are essential to the traceability, access and effective use of scientific data. In RADAR, submitted data must be accompanied by a set of basic descriptive metadata elements that document and describe a particular resource. The following scheme aims to enhance the traceability and usability of research data by maintaining a discipline agnostic character and simultaneously allowing a description of discipline specific data.

The RADAR Metadata Schema (Table 1.) includes ten mandatory fields, which represent the general core of the schema[3]. These fields contain the main requirements for DOI registration, in accordance with DataCite Metadata Schema 4.0. (DataCite 2016) and must be supplied when submitting metadata to RADAR. Additionally, 13 optional metadata parameters serve the purpose of describing discipline specific data.

Table 1. RADAR Descriptive Metadata Schema. The Schema contains a mandatory set of generic parameters to allow for the accurate and consistent identification of a resource for citation and retrieval purposes. Subsequently, optional parameters can be described, to meet the requirements of discipline-specific datasets.

Descriptive Metadata: 10 standard parameters	Mandatory parameters for basic information
Identifier (RADAR ID/DOI)	A unique string, which identifies a resource. Handle for data preservation/DOI for Data publication service)
Creator	Persons involved in producing the data
Title	Study/Data title
Publisher	Corporate/Institutional or personal name
Production Year or time span	Year, in which data was created or refers to
Publication Year	Year, in which the resource was published
Subject Area	Scientific fields appropriate for the resource
Resource	Resources content (e.g. dataset, model, software)
Rights	Rights management statement (e.g. CC BY)
Rights holder	Institution/Person holding rights.
Descriptive Metadata: 13 optional parameters	**Parameters for discipline specific data description**
Additional title	Additional title type (e.g. translated title)
Description	Further information (e.g. abstract)
Keyword	Keywords describing the subject focus
Contributor	Associated institution/person
Language	Primary language used or relevant to resource
Alternative Identifier	Unique string within its domain of issue (e.g. local identifier)
Related Identifier	Identifiers of related resources
Geo Location	Region/Place where resource originated/refers to
Data Source	Data origin (e.g. instrument, observation, trial)
Software type	Software used for data production and processing
Data Processing	Specifies further processing (e.g. statistics)
Related Information	Further information (e.g. database number)
Funder Information	Funder information

The parameters were implemented with a combination of controlled vocabularies and free-text entries, thereby covering heterogeneous data produced by a multiple of disciplines. The controlled

3 RADAR Schema documentation: https://www.radar-service.eu/en/radar-schema

vocabulary entries were defined in accordance with established regulations in mind (e.g. ISO standards). RADAR clients, who wish to enhance the prospects of their metadata being found, cited and linked to original research are strongly encouraged to submit the optional parameters in addition to the mandatory set of properties. The metadata of datasets that are published in RADAR will be available under the Creative Commons Zero licence (Creative Commons 2014) RADAR will actively disseminate all published metadata to DataCite. Metadata of datasets that are only archived (not published) in RADAR are only available to the data provider, unless otherwise specified. Moreover, a support service for data harvesting of published metadata via OAI-PMH interface is provided.

Business Model

The business model, including the services presented in the previous section, ensures a sustainable operation environment for the data archive as well as for institutional users. From the start, RADAR focuses on publicly funded research institutions and universities in Germany. This limitation is mainly driven by contractual and legal issues. RADAR strives to loosen some of the limitations in the near future to broaden the potential user base and expand to neighbouring European countries.

The ongoing operation of RADAR is not based on project funding. Operational costs include personnel, marketing and travel expenses and fees for the basic IT infrastructure. Half of these costs are taken over by FIZ Karlsruhe, which understands RADAR as an important building block of the information infrastructure and an excellent fit for its mission. The other half of the operational costs and all variable costs (which are mainly the costs for maintaining three copies of the data in two data centres) are factored into the pricing of the service. Being charged for such a service might turn away researchers, but at the same time, it might be a healthy exercise to re-evaluate the data produced in the course of a project and decide which data needs to be published, which can be archived and which might even be deleted.

RADAR offers two different pricing models for the two service levels: for archived data, the amount of stored data defines the price per year[4]. Institutions may end contracts and move the data to other service providers any time. For published data, the message from the academic community was very clear that there needs to be a guarantee that the data is available independent from the contractual situation. Thus, RADAR offers a one-time payment model for published data with a guaranteed retention period of at least 25 years. One-time payments also work well with the research system which relies in most cases on project funding with no option for ongoing payments after the end of a project. Due to the corporation with outstanding partners, RADAR can offer very competitive pricing for its service[5].

Conclusion & Outlook

As universities and research institutions are increasingly interested in collecting and providing access to datasets produced at their institution, not all of this data will fall within the scope of big

4 RADAR Pricing Information: https://www.radar-service.eu/en/pricing
5 RADAR Pricing Structure: https://www.radar-service.eu/en/pricing

data or discipline based repositories. The researchers from long tail of science will start to look to libraries to provide support and data services for their datasets.

With RADAR, we present a solution that has been designed by a research consortium of academic institutions for the academic community. This interdisciplinary approach, competitive pricing, option to integrate RADAR with existing services and workflows, and compliance with German and European legislation makes RADAR a viable option for research data archival and publication. The novel two-stage service and business model combined with a trustworthy repository for institutions and their researchers provides a contribution to ensure a better availability, sustainable preservation and publishability of research data for present and future academic communities.

Acknowledgements

We gratefully acknowledge the tremendous efforts of everyone involved in providing support throughout the project. In particular, we wish to thank the RADAR project team for their ongoing input to the RADAR service. We thank the scientific advisory board of RADAR for their contributions to our discussions on data management, research data services and infrastructures. We also thank the academic community for evaluating the test system. Their constructive comments have helped to improve the RADAR service before we moved to production.

Funding

RADAR was developed as part of a three-year project funded by the German Research Foundation (DFG) from 2013 to 2016 (http://www.radar-projekt.org) and is placed within the program "Scientific Library Services and Information Systems (LIS)" on restructuring the national information services in Germany.

References

Berman, Francince, Brian Lavoie, Paul Ayris, G. Sayeed Choudhurry, Elizabeth Cohen, Pual Courrant, Lee Dirks, Amy Friedlander, Vijay Gurbaxani, Anita Jones, Ann Kerr, Clifford Lynch, Daniel Rubinfeld, Chris Rusbridge, Roger Schonfeld, Abby Smith Rumsey, and Anne Van Camp. 2010. *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information: Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access.*

Borgman, Christine L., Peter T. Darch, Ashley E. Sands, Irene V. Pasquetto, Milena S. Golshan, Jilian C. Wallis, and Sahron Traweek. 2015. "Knowledge infrastructures in science: Data, diversity, and digital libraries." *International Journal on Digital Libraries* 16, 3: 207–227.

Borgman, Christine L. 2015. *Big data, little data, no data: Scholarship in the networked world*. Cambridge, MA: MIT Press.

Buckland, M. 2011. “Data management as bibliography.” *Bulletin of the American Society for Information Science and Technology* 37: 34–37.

Creative Commons. 2014. Creative Commons Licenses. https://creativecommons.org/licenses/.

DataCite. 2016. DataCite Metadata Schema 4.0. https://schema.datacite.org/.

DFG, Deutsche Forschungsgemeinschaft. 2013. Safeguarding Good Scientific Practice. Recommendations of the Commission on Professional Self Regulation in Science.: WileyVCH.

European Commission. 2016. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

FORCE11. 2016. https://www.force11.org/group/fairgroup/fairprinciples.

Klump, J. 2009. “Managing the Data Continuum.” http://oa.helmholtz.de/fileadmin/user_upload/Data_Continuum/klump.pdf.

National Science Foundation. 2011. “Proposal Preparation Instructions.” Grant Proposal Guide.

Neuroth, H, S Strathmann, A Oßwald, R Scheffel, J Klump, and J Ludwig. 2012. “Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme.”
Science Editorial, 2011. “Challenges and Opportunties.” Science 331. 6018: 692-693.

Treloar, Andrew, and Cathrine Harboe-Ree. 2008. “Data management and the curation continuum: how the Monash experience is informing repository relationships.”

Wilkinson, Mark D. et al. 2016. “The FAIR Guiding Principles for scientific data management and stewardship.” Nature, March.

Open Data in den Sozial- und Wirtschaftswissenschaften: Das Forschungsdatenrepositorium SowiDataNet

Patrick J. Droß[1], Mathis Fräßdorf[2], Paul Kubaty[3], Julian Naujoks[4]

1,2,3,4 Wissenschaftszentrum Berlin für Sozialforschung

Zusammenfassung. Diverse Empfehlungen und Richtlinien von Forschungsförder- und Wissenschaftsorganisationen verdeutlichen die Bedeutung der Nachnutzung von Forschungsdaten. Die Forderung nach Open Data – also der Archivierung, Veröffentlichung und Nachnutzung von Forschungsdaten – wird immer lauter. Stellvertretend für die Sozial- und Wirtschaftswissenschaften rekonstruiert der Beitrag am Beispiel des Repositoriums SowiDataNet zentrale Anforderungen für den Aufbau einer neuen Forschungsdateninfrastruktur mit dem Ziel des Data Sharings. Drei Aspekte müssen besonders berücksichtigt werden: Fachspezifische Besonderheiten, institutionelle Bedürfnissen und nicht zuletzt die subjektiven Bedenken aus dem Forschungsalltag der Wissenschaftler/innen.

Schlagwörter. Forschungsdaten, Open Data, Repositorium, Sozialwissenschaften, Wirtschaftswissenschaften

Open Data: Ein neues Arbeitsgebiet für Wissenschaft und Infrastruktur

Forschungsdaten nehmen in den empirischen Wissenschaften, insbesondere in den Sozial- und Wirtschaftswissenschaften, einen immer größeren Raum ein. Sie sind nicht nur ein wichtiger Ausgangspunkt wissenschaftlicher Arbeit, sondern werden auch zunehmend als Teil des wissenschaftlichen Outputs gesehen. Wo es früher in Veröffentlichungen nur Beschreibungen von Datenerhebungen gab, wird es heute immer häufiger als gute Praxis angesehen, auch die den Veröffentlichungen zugrundeliegenden Daten bereit zu stellen.

Diese Entwicklung hat unterschiedliche Ursachen: Zunächst wurden im Sinne der guten wissenschaftlichen Praxis Forderungen nach mehr Transparenz und besseren Möglichkeiten der Nachvollziehbarkeit wissenschaftlicher Ergebnisse erhoben. Bereits 1998 formulierte die DFG entsprechende Grundsätze in ihren Empfehlungen (vgl. DFG 1998). Zu diesem Zeitpunkt war an eine entsprechende Umsetzung – insbesondere an die Verfügbarkeit von Daten – allerdings noch nicht zu denken. Die technischen Voraussetzungen waren schlichtweg nicht gegeben. In der Tat rückte erst die Open-Access-Bewegung den Zusammenhang zwischen den Potenzialen des Internets und den Möglichkeiten der Zugänglichmachung wissenschaftlicher Ergebnisse in den Fokus der Diskussion.

Insbesondere durch die Berliner Erklärung aus dem Jahr 2003 wurden diese Möglichkeiten mit einem explizit politischen Anspruch verbunden: Es ging um den Abbau von Schranken zum Wissen, ganz gleich, ob sie technischer, sozialer oder wirtschaftlicher Natur sind. Wissenschaftlicher Output wurde als Teil eines digitalen Gemeinguts betrachtet (vgl. Blasetti et al. 2017). Die nachvollziehbare Logik dahinter: Resultate aus öffentlich finanzierter Forschung sollten auch der

Allgemeinheit zur Verfügung stehen. Das umfasst sämtliche im Forschungsprozess erzielten Ergebnisse, also nicht nur den fertigen Journal-Artikel, sondern zugleich die Forschungsdaten, möglicherweise auch Syntaxfiles und Fragebögen. Auch vor dem Hintergrund einer effizienten Mittelvergabe wird diese Forderung verständlich: Wenn Daten verfügbar gemacht werden, können andere diese Daten möglicherweise nachnutzen und müssen keine eigenen Erhebungen durchführen. In der Praxis gibt es mittlerweile zahlreiche Forschungsförderer, die entlang dieser Argumentation eine Veröffentlichung von Forschungsdaten verbindlich einfordern: Bspw. erwartet die EU in ihrem Rahmenprogramm „Horizon 2020“ explizit die Bereitstellung von Daten aus Forschungsprojekten, die über ihre Förderlinie erstellt werden.

Unabhängig von den eher (wissenschafts-)politischen Argumenten gibt es aber auch Anreize für Wissenschaftler/innen, die Forschungsdaten produzieren, diese auch zu veröffentlichen. Zunächst ist hier eine erhöhte Sichtbarkeit der eigenen Arbeit hervorzuheben. Man wird nicht nur als Autor einer Textpublikation wahrgenommen, sondern auch als Datenproduzent, der anderen eine Nachnutzung ermöglicht und sich außerdem nicht scheut, dass die eigene Arbeit überprüfbar wird. Überdies gibt es Untersuchungen, die nahelegen, dass durch die Verfügbarkeit von Forschungsdaten auch die Zitation der entsprechenden Textpublikationen steigt und sich so die Bewertung der Autorin über die gängigen Indices ebenfalls erhöht (vgl. Piwowar et al. 2017). Langfristig ist zudem damit zu rechnen, dass die Produktion von Forschungsdaten in die Beurteilung der wissenschaftlichen Leistung einfließen wird.

SowiDataNet – Kurzportrait

Im Rahmen des SowiDataNet-Projektverbundes arbeiten GESIS – Leibniz-Institut für Sozialwissenschaften, die Deutsche Zentralbibliothek für Wirtschaftswissenschaften (ZBW) das Wissenschaftszentrum Berlin für Sozialforschung (WZB) und das Deutsche Institut für Wirtschaftsforschung (DIW Berlin) am Aufbau einer neuen Forschungsdateninfrastruktur für die Sozial- und Wirtschaftswissenschaften. Das Projekt ist gefördert durch die Leibniz-Gemeinschaft. Zentraler Baustein ist die Entwicklung eines web-basierten Forschungsdatenrepositoriums, welches es Wissenschaftler/innen in einem institutionellen Arbeitsumfeld erlaubt, ihre Daten nachhaltig zu dokumentieren, sicher und dauerhaft zu archivieren, zu veröffentlichen und damit anderen Forscher/innen zur Nachnutzung zur Verfügung zu stellen. Einen besonderen Fokus legt das Projekt auf die speziellen Bedarfe der beiden Fachcommunities sowie auf eine möglichst flexible und praxisnahe Einbindung des Repositoriums in die Workflows des institutionellen Forschungsdatenmanagements.

Den Instituten soll die Möglichkeit gegeben werden, für die Veröffentlichung ihrer Daten, die in der Regel in mittleren und kleinen Forschungs- oder auch Promotionsprojekten produziert werden, einen zentralen Infrastrukturservice zu nutzen, ohne in eine eigene Infrastruktur zur Datenpublikation investieren zu müssen. Dies hat zudem den Vorteil, dass Forschungsdaten der Sozial- und Wirtschaftswissenschaften in Deutschland nach einheitlichen Standards archiviert und dokumentiert, zitiert und zugänglich gemacht werden können. Die Zusammenführung von Forschungsdaten aus unterschiedlichen Instituten führt zu innovativen Recherche-Möglichkeiten, bspw. nach thematischen Schlagworten oder nach unterschiedlichen Erhebungsmethoden. Dabei verbessert eine zentrale Anbindung an das GESIS-Datenarchiv substanziell die Möglichkeiten einer effektiven Langzeitarchivierung von Forschungsdaten und stellt trotz der Fluktuation der

Forscher/innen über Institutsgrenzen hinweg sicher, dass Forschungsdaten zentral aufbewahrt und in einer Form dokumentiert werden, die es erlaubt, sie später auch ohne Unterstützung durch die Datenproduzent/innen nachvollziehen und nachnutzen zu können.

In SowiDataNet werden Forschende und Forschungsdatenmanager/innen daher die Möglichkeit erhalten, ihre Datensätze sowie relevante Begleitdokumente (etwa Fragebögen, Codebooks, Skripte oder Technical Reports) eigenständig in das Repositorium einzupflegen und mit detaillierten Metadaten zu beschreiben. Forscher/innen können hierbei anhand von frei wählbaren Zugangsklassen selbst definieren, unter welchen Bedingungen – z.B. freie Verfügbarkeit bis hin zu Embargofristen – sie den Zugriff auf ihre Daten erlauben. Zur sicheren und persistenten Identifizierung von Forschungsdaten nutzt SowiDataNet den Service der Registrierungsagentur für Sozial- und Wirtschaftsdaten da|ra zur DOI-Vergabe.

Die Generierung von Metadaten ist ein wesentlicher Bestandteil im Dokumentationsprozess von Forschungsdaten. Metadaten sind unerlässlich für die Auffindbarkeit sowie die erneute Verwendung der archivierten Daten. Das Metadatenschema von SowiDataNet enthält daher die obligatorischen Kernelemente, die zur Beschreibung der Forschungsdaten notwendig sind, erlaubt aber darüber hinaus die Angabe einer Vielzahl fachspezifischer Metadaten-Elemente. Um die Auffindbarkeit und Nachnutzung der Daten zu vereinfachen, ist zudem eine inhaltliche Erschließung von Forschungsdaten mittels fachspezifischer Thesauri vorgesehen.

Im Folgenden werden einige zentrale Punkte der im Rahmen des Projektes durchgeführten Anforderungsanalyse beleuchtet und anschließend die konkreten Umsetzungsschritte dargestellt.

Anforderungen an das Data Sharing

Viele Forschungsförder- und Wissenschaftsorganisationen empfehlen und fordern die Sicherung und Veröffentlichung von Forschungsdaten und begründen dies mit dem Argument, dass dadurch Anknüpfungspunkte für weitere darauf aufbauende Forschung geschaffen werden (vgl. DFG 2015, Allianz der deutschen Wissenschaftsorganisationen 2010). Aus dieser Perspektive ist eine ausschließliche Archivierung bzw. geschlossene Aufbewahrung von Forschungsdaten nicht erstrebenswert. Das Ziel liegt vielmehr in der Nachnutzung der Daten – doch wie kann das „Data Sharing“ in der Praxis konkret aussehen?

Der Rat für Sozial- und Wirtschaftsdaten (RatSWD) hat in seinem jüngst verfassten Ratgeber „Forschungsdatenmanagement in den Sozial-, Verhaltens- und Wirtschaftswissenschaften“ darauf hingewiesen, dass die „Dokumentation […] so ausgestaltet sein [sollte], dass eine Nachnutzung der Daten möglich ist“ (RatSWD 2016, 10). Zwar könne die reine Sicherung von Daten bereits mit einer „Minimaldokumentation“ (ebd.) erfolgen, die Möglichkeit der Nachnutzung sei allerdings erst dann gegeben, wenn die Forschungsdaten nutzerfreundlich dokumentiert werden. Dies deckt sich mit Ergebnissen weiterer Untersuchungen, nach denen die Qualität der Datendokumentation ausschlaggebend für eine effiziente Nachnutzung ist (vgl. Fecher/Puschmann 2015, Fecher et al. 2015). Forschungsdaten sollten also mit umfangreichen Metadaten beschrieben und qualitätsgesichert kuratiert werden. Da sich sowohl die Wege der Datengenerierung als auch die Formen der Datennutzung je nach wissenschaftlicher Disziplin stark unterscheiden, sind fachspezifische Standards der Datendokumentation unerlässlich: Bereits 2010 hat die Allianz der deutschen Wissenschaftsorganisationen angeführt, dass die Möglichkeiten der Nachnutzung von Forschungsdaten von der fachspezifischen Nachvollziehbarkeit „der Art und Weise der Datenerhe-

bung, des Umfangs und der Vernetzbarkeit des Datenmaterials sowie der praktischen Brauchbarkeit der Daten“ (Allianz der deutschen Wissenschaftsorganisationen 2010) abhängig sind. Ähnlich lautet der Aufruf der DFG, „angemessene Regularien zur disziplinspezifischen Nutzung und ggf. offenen Bereitstellung von Forschungsdaten zu entwickeln“ (DFG 2015, 2).

Im Projektaufbau von SowiDataNet wurden diese Aspekte von Beginn an integriert: Dementsprechend galt bei der Konzeption die Prämisse, dass die Entwicklung eines Forschungsdatenrepositoriums im Zusammenspiel von Infrastruktur und Forschung erfolgen muss. Dies war ausschlaggebend dafür, dass mit dem GESIS – Leibniz-Institut für Sozialwissenschaften und der Deutschen Zentralbibliothek für Wirtschaftswissenschaften (ZBW) bzw. dem Wissenschaftszentrum Berlin für Sozialforschung (WZB) und dem Deutschen Institut für Wirtschaftsforschung (DIW Berlin) zwei etablierte Einrichtungen der Leibniz-Gemeinschaft jeweils aus der Infrastruktur und aus der sozial- und wirtschaftswissenschaftlichen Forschung beteiligt waren.

Die disziplinäre Begrenzung auf die Sozial- und Wirtschaftswissenschaften folgte dem Anspruch, die notwendigen detaillierten fachspezifischen Standards für die Datendokumentation aufzustellen und damit das Repositorium an den Bedürfnissen der Fachcommunities auszurichten. Im Zeitraum von 2014 bis 2015 wurden deshalb im Rahmen einer Anforderungsanalyse Projektworkshops durchgeführt und leitfadengestützte Experteninterviews mit empirisch arbeitenden Forscher/innen aus den Sozial- und Wirtschaftswissenschaften geführt. Diese Interviews sollten Einblicke in die jeweiligen Arbeitsweisen und Bedürfnisse der Wissenschaftler/innen geben, um diese in die aufzubauende Forschungsdateninfrastruktur einfließen zu lassen. Der Gedanke war hierbei, dass dabei letztlich auch die Bereitschaft der Forscher/innen steigen dürfte, das Repositorium zu nutzen. In den folgenden Abschnitten werden zentrale Erkenntnisse aus dieser Anforderungsanalyse zusammengefasst: Auf der fachspezifischen Ebene müssen zunächst disziplinäre Eigenheiten berücksichtigt werden. Institutionell kommen konkrete Anforderungen der Forschungseinrichtungen zum Ausdruck und auf der forschungspraktischen Ebene ist es von Bedeutung, den (Arbeits-)Alltag und die subjektiven Erwartungen der Forscher/innen miteinzubeziehen.

Fachspezifische Anforderungen

Die fachliche Zielgruppe des SowiDataNet-Projektverbunds ist die sozial- und wirtschaftswissenschaftliche Forschungsgemeinschaft in Deutschland. Die beiden Disziplinen eint nicht nur, dass sie im beachtlichen Maße empirisch arbeiten, sie kennzeichnen sich vor allem durch eine erhebliche Vielfalt in den Wegen und Methoden der Datengenerierung zur Beantwortung ihrer Forschungsfragen. So werden in klassischen quantitativen Surveys unter Verwendung standardisierter Erhebungsinstrumente Mikrodaten erhoben. In experimentellen Laborsettings wird der Einfluss gezielter Stimuli auf Probanden untersucht oder in Feldexperimenten das Verhalten ganzer Untersuchungsgruppen in den Blick genommen. Vielfach werden jedoch auch prozessproduzierte bzw. Sekundärdaten aufbereitet, weiterverarbeitet und mit neuen Informationen angereichert. Schließlich umfasst die qualitative Forschung ein ganzes Spektrum unterschiedlicher Erhebungsmethoden: Dies reicht von klassischen Experteninterviews über ethnografische Forschungsansätze (teilnehmende Beobachtung, Feldnotizen, Fotografie), der Codierung von Sekundärtexten (Zeitungsartikel, Parteiprogramme) bis hin zu Videoaufzeichnungen. Diese methodischen Zugänge werden immer häufiger im Rahmen von Mixed-Methods-Studien miteinander verwoben, um die Vorteile der einzelnen Ansätze zu kombinieren.

Das Feld der empirischen Forschungspraxis in den Sozial- und Wirtschaftswissenschaften erweist sich somit als äußerst heterogen. Die unterschiedlichen Forschungsansätze bringen teils unterschiedliche Prioritäten im Hinblick auf Datenmanagement und Datenarchivierung mit sich. Geringe Fallzahlen (z.B. in Experimenten) oder nicht standardisierbare Methoden (z.B. qualitative Interviews) führen bspw. zu höheren Anforderungen an den Datenschutz, als dies bei anonymisierten Umfragen der Fall ist. Bei der Verwendung von prozessproduzierten bzw. Sekundärdaten muss hingegen die Frage der Nutzungsrechte hinreichend geklärt sein, was wiederum bei Erhebungen durch Forschende selbst meist keine große Rolle spielt. Aus diesen fachspezifischen Besonderheiten lassen sich infolgedessen zwei wesentliche Anforderungen an ein Repositorium für sozial- und wirtschaftswissenschaftliche Forschungsdaten ableiten: Erstens ist es für die potenzielle Nachnutzung der Daten unerlässlich, durch die Erfassung fachspezifischer Metadaten den Entstehungskontext der Daten nachvollziehbar zu machen. Zweitens ist es für die Nachvollziehbarkeit der Forschungsdaten notwendig, dass sich die teils komplexen methodischen Designs im Repositorium abbilden lassen und auch für einzelne Datensätze Angaben zur Methodik erfassen lassen.

Institutionelle Anforderungen

Die Empfehlungen der Förder- und Wissenschaftsorganisationen für einen geregelten Umgang mit Forschungsdaten haben nicht nur Auswirkungen auf die Forscher/innen, sondern nehmen selbstverständlich auch deren Arbeitgeber, Universitäten und außeruniversitäre Forschungsinstitute, in die Pflicht. Denn als einzelne/r Forscher/in sind die Anforderungen nur mit großem Aufwand leistbar, was wiederum Effizienzverluste und damit negative Auswirkungen für den Arbeitgeber hätte. Entsprechend stehen wissenschaftliche Institutionen vor der Herausforderung, Lösungen zu präsentieren, die es den Mitarbeiter/innen erlauben, die Kriterien der guten wissenschaftlichen Praxis zu erfüllen.

Gleichzeitig wird die Abfrage institutioneller Forschungsleistungen mittelfristig zunehmen – z.B. im Rahmen von Evaluierungen – und dabei Forschungsdaten nicht ausklammern. Ähnlich wie bei Text-Publikationen haben Institutionen ein Interesse, dass Informationen über die Forschungsdatenproduktion zentral verfügbar sind und Forscher/innen nicht in Eigeninitiative ihre Forschungsdaten z.B. in sozialen Netzwerken, bei kommerziellen Anbietern, auf Projektwebseiten oder auf privaten Homepages zur Verfügung stellen.

Auch die Anerkennung von Forschungsdaten als wissenschaftlicher Output liegt im Interesse der Institutionen, deren Mitarbeiter/innen in ihrer Forschung Daten produzieren. Hierzu gilt es, zunächst infrastrukturelle Voraussetzungen zu schaffen, die es überhaupt erst ermöglichen, dass diese eigenständige Forschungsleistung als solche wahrgenommen werden kann.

Forschungspraktische Anforderungen

Um Perspektiven aus dem praktischen Forschungsalltag in die Entwicklung von SowiDataNet miteinzubeziehen, wurden im Rahmen der Anforderungsanalyse zehn Forscher/innen zu ihren Erwartungen an ein Forschungsdatenrepositorium mit dem Ziel der Nachnutzung befragt. In der Auswertung ließen sich unterschiedliche Bedenken – vor allem nicht-technischer Natur – heraus-

stellen, die in der folgenden Abbildung aufgelistet sind. Mehrfachnennungen in einem einzelnen Interview wurden einfach gewertet.

Abbildung 1. Bedenken der Forschenden sortiert nach Anzahl der Interviews mit entsprechender Nennung, N=10 (Patrick J. Droß)

Fast alle interviewten Personen teilten die Befürchtung, dass aus der Archivierung und der damit verbundenen Beschreibung und Aufbereitung der Daten ein zusätzlicher Arbeitsaufwand resultiert. Die folgende Aussage kann hierfür als exemplarisch angesehen werden: „Die Frage ist natürlich, mit welchem Aufwand das verbunden ist. Also ich wäre bereit, die Daten alle ins Netz zu stellen, hätte allerdings keine Lust, daran einen Monat zu arbeiten, um die einzugeben, dann sage ich nee, also keinen Bock drauf“. Häufig wiesen die befragten Personen auf ihre geringen Spielräume in Bezug auf die finanziellen und zeitlichen Ressourcen hin. Verschiedentlich wurde die knappe Kalkulation von Forschungsprojekten angeführt, aus der nicht selten eine Aus- und Überlastung abzulesen waren – ein grundlegendes Hemmnis, was die Bereitschaft angeht, für eine Dokumentation von Forschungsdaten zusätzliche Mehrarbeit in Kauf zu nehmen.

Die Mehrzahl der Interviewten gab zudem an, dass die bisherige Datendokumentation häufig nur für den internen Gebrauch konzipiert ist, um beispielsweise Informationen für die eigene Weiterverwendung oder für Kollegen/innen festzuhalten. Eine strukturierte Datendokumentation mit dem Ziel der Veröffentlichung müsste ganz andere Arbeitsweisen innerhalb von Forschungsgruppen nach sich ziehen: „Ich weiß nicht, wie es woanders ist, aber bei uns gibt’s das im eigentlichen Sinne nicht, so eine strukturierte Dokumentation, jeder macht das irgendwie auf seine Art. […] Wenn man es sozusagen mal für die Allgemeinheit vernünftig aufbereiten wollte, würde man das wahrscheinlich ganz anders konzipieren“. Meist sind die Daten demnach nicht in solch einem Ausmaß dokumentiert, dass sie ohne weiteres der wissenschaftlichen Gemeinschaft zur Nachnutzung bereitgestellt werden könnten.

Eine weitere Befürchtung betrifft den Besitz- bzw. Erstnutzungsanspruch. Sechs von zehn Personen haben den Wunsch geäußert, ihre erhobenen Daten zunächst selbst verwenden zu wollen, bevor sie sie zur Nachnutzung bereitstellen: „Also speziell die aufbereiteten Sekundärdaten, da haben wir dann manchmal Anfragen, da sind wir aber sehr zurückhaltend, einfach, weil es wahnsinnig viel Arbeit war und weil man jetzt erstmal nichts davon hat, wenn man den Datensatz rausgibt. Also eigentlich müssten wir selber noch erstmal mehr davon veröffentlicht haben, wenn

ich das mal so frank und frei sagen kann“. Die Sorge besteht darin, dass eine zu frühe Publikation der Forschungsdaten dazu führen könnte, dass andere Forschende die Ergebnisse schneller veröffentlichen. In diesem Verständnis wird den Forschungsdaten eine gewisse Exklusivität eingeräumt. Als intellektuelles Kapital können sie zum Beispiel für neue Projektanträge verwendet werden, wie es eine befragte Person klar zum Ausdruck bringt: „Also wir haben uns auch nicht so sehr bemüht [die Daten zugänglich zu machen], ich würde auch immer sehen, dass es ja auch ein bisschen gefährlich ist, gerade solche originären Daten, die wenig vorhanden sind, breit zu streuen, weil das natürlich in unserer Hand auch ein Pfund war. Wenn ich das gleich auf den Markt schmeiße, dann ist das natürlich weg, und wir wussten ja auch nicht, inwieweit wir selber noch weitere Analysen damit durchführen“.

SowiDataNet: dokumentieren – veröffentlichen – nachnutzen

Ohne den Prozess der Datenkuratierung und -veröffentlichung mit SowiDataNet hier in seiner gesamten Breite darstellen zu können, soll im Folgenden zumindest auf drei Spezifika des Repositoriums eingegangen werden, auf deren Entwicklung in Folge der Anforderungsanalyse ein besonderer Schwerpunkt lag: Die Unterscheidung von Projekt- und Objektebene bei der Datenbeschreibung, die spezifischen Funktionen für institutionelle Datenkuratoren sowie die Darstellung der institutionellen Sammlung in einer „Vitrine“.

Datenbeschreibung auf Projekt- und Objektebene

Wie die Anforderungsanalyse im Projekt SowiDataNet zeigte, stehen Forscher/innen bei der Veröffentlichung ihrer Daten vor der Frage, wie sich die teils komplexen Forschungsdesigns in einer Datenpublikation abbilden lassen. Unsicherheiten bestehen daher vor allem in Bezug auf die möglichen Granularitätsstufen einer Datenpublikation: Sollen bspw. zehn experimentelle Untersuchungsreihen besser gemeinsam oder in zehn getrennten Datenprojekten veröffentlicht werden? Wie können die methodischen Unterschiede einer Mixed-Methods-Studie beschrieben werden, ohne dass der für das Verständnis der Daten unerlässliche Gesamtkontext des Studiendesigns verloren geht? Auf diese Fragen kann es aus informationstechnischer Sicht kaum pauschale Antworten geben. Es bedarf vielmehr flexibler und pragmatischer Lösungen im Einzelfall sowie der Unterstützung durch erfahrene Datenkurator/innen.

Abbildung 2. Die Projekt-Objekt-Ebene in SowiDataNet (Paul Kubaty)

Um flexible Lösungen auf unterschiedlichen Granularitätsstufen zu ermöglichen, unterscheidet das SowiDataNet-Repositorium bei der Dokumentation der Forschungsdaten zwischen einer Projekt- und einer Objektebene. Unter der Projektebene wird das zu veröffentlichende Datenprojekt als Ganzes verstanden, während sich die Objektebene auf die einzelnen Dateien (Datensätze und Begleitdokumente) bezieht. Wird im Repositorium ein neues Datenprojekt angelegt, können zunächst Basisangaben gemacht sowie inhaltliche Beschreibungen auf der Projektebene vorgenommen werden (s. Abb. 2). Auch Angaben zur Methodik lassen sich auf dieser übergeordneten Ebene erfassen. Beim Upload einzelner Datensätze lässt es das Repositorium jedoch ebenfalls zu, dass zusätzliche Metadaten auf der Ebene einzelner Objekte erfasst werden. Dies führt in der Konsequenz dazu, dass unterschiedliche Datenobjekte in einer Datenpublikation zusammengefasst werden können, ohne dass auf die differenzierte methodische Beschreibung einzelner Objekte verzichtet werden muss.

Spezielle Funktionen für institutionelle Datenkuratoren

Forscher/innen aus den Sozial- und Wirtschaftswissenschaften äußern vielfach Bedenken hinsichtlich des Arbeitsaufwands, der mit der Veröffentlichung ihrer Forschungsdaten verbunden ist. Institutionellen Beratungs- und Unterstützungsangeboten kommen daher bei der weiteren Verbreitung des Open-Data-Gedankens eine Schlüsselrolle zu. Datenkurator/innen können frühzeitig über die formalen Anforderungen an die Datenaufbereitung, sinnvolle Schritte der Datendokumentation sowie über mögliche Embargofristen und Lizenzen informieren. In den oftmals arbeitsintensiven Abschlussphasen der Forschungsprojekte können die Forscher/innen jedoch auch von einer

aktiven Unterstützung bei der Datenkuratierung profitieren. Neue Infrastrukturangebote sollten daher idealerweise die Arbeit des institutionellen Forschungsdatenmanagements unterstützen und sich flexibel in dessen Workflows integrieren lassen.

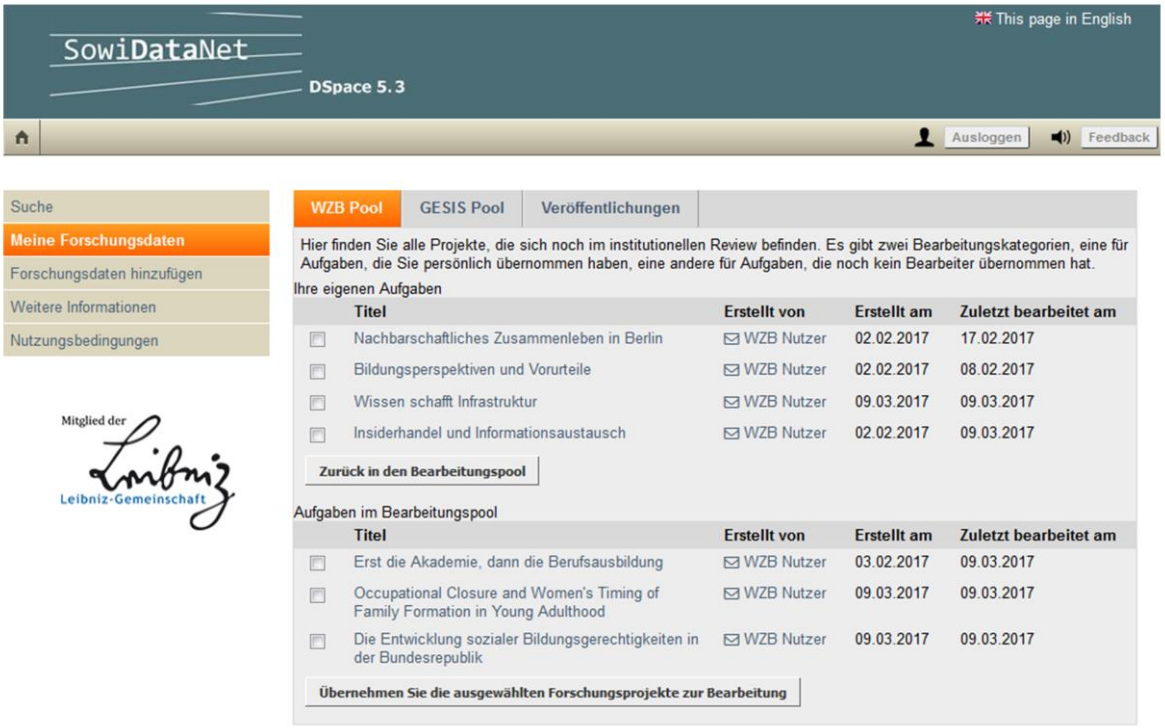


Abbildung 3. Der institutseigene Pool in SowiDataNet (Gesis)

SowiDataNet implementiert daher Funktionen, die sich speziell an die Datenkurator/innen des jeweiligen Forschungsinstitutes richten (s. Abb. 3). So können Forscher/innen in einem ersten Schritt neue Datenprojekte selbst anlegen, Forschungsdaten hochladen und mittels standardisierter Metadaten beschreiben. Bereits während der Bearbeitung können sie Kommentarfunktionen nutzen, um offene Fragen festzuhalten. Ist die Eingabe seitens der Forscher/innen beendet, wird das Datenprojekt in einen institutionellen Projektpool übergeben. Auf diesen kann in einem zweiten Schritt der/ die Institutskurator/in zugreifen und für ein ausgewähltes Projekt einen inhaltlichen Reviewprozess starten. Dabei prüft er die Daten, Metadaten und Begleitdokumente nach formalen Kriterien, auf Lesbarkeit, Vollständigkeit und korrekte Beschreibung. Er kann auf Fragen der Forscher/innen eingehen und bei Bedarf in Abstimmung mit den Forscher/innen Informationen ergänzen bzw. direkt selbst Korrekturen vornehmen. Wenn erforderlich, kann ein Datenprojekt auch an die Forscher/innen zurückgeben werden. Als ein Angebot zur Standardisierung und als Arbeitshilfe für den/ die Kurator/in wird von SowiDataNet systemseitig eine Checkliste bereitgestellt, entlang derer die eingereichten Datenprojekte überprüft werden können. Diese Checkliste soll sich künftig an die jeweiligen Bedarfe des Instituts anpassen lassen und auch nach dem offiziellen Projektstart in Zusammenarbeit mit den Nutzern weiterentwickelt werden. Ist der institutionelle Review abgeschlossen, übermittelt der/ die Kurator/in das Datenprojekt in einem dritten Schritt an GESIS, den technischen Betreiber des Repositoriums. Hier erfolgen letzte technische Kontrollen, bevor das Projekt über den Registrierungsservice da|ra mit der Vergabe einer DOI veröffentlicht wird.

Darstellung der institutionellen Sammlung

Wie eingangs beschrieben, wird es zunehmend als gute wissenschaftliche Praxis angesehen, nicht nur die Auswertungsergebnisse, sondern auch die den Auswertungen zugrunde liegenden Forschungsdaten zu veröffentlichen. Auch wenn sich die Institute dazu entscheiden, einen zentralen Service zur Datenpublikation zu nutzen, wird dennoch das Interesse bestehen, die eigenen Forschungsdaten lokal sichtbar zu machen und als institutionelle Forschungsleistungen zu präsentieren. SowiDataNet bietet daher in einem Zusatzmodul die Möglichkeit, die institutionelle Datensammlung innerhalb des eigenen Webauftritts darzustellen und durchsuchbar zu machen. Diese institutionelle Vitrine kann an das jeweilige Corporate Design des Instituts angepasst werden (s. Abb. 4).

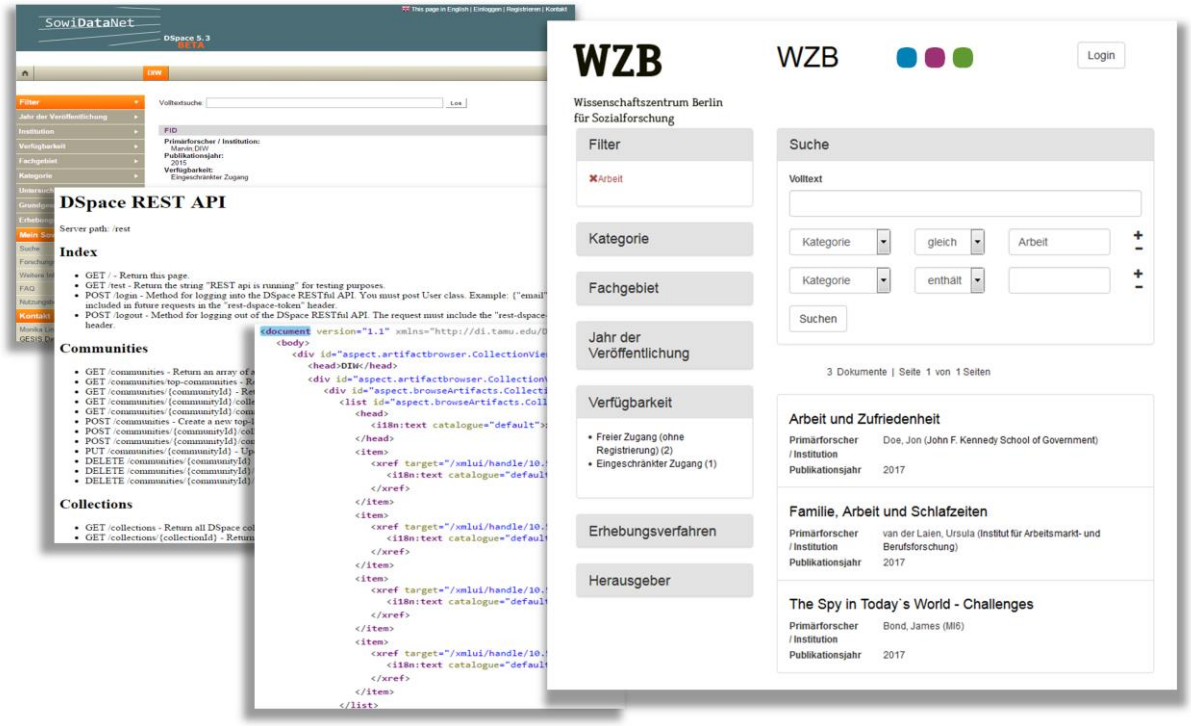


Abbildung 4. Die institutionelle Ansicht in SowiDataNet (Timo Borst, Patrick Droß)

Mit der institutionellen Vitrine stellt SowiDataNet ein eigenes Frontend für die Institute bereit. Die Suchfunktionen entsprechen dabei eins zu eins der Suche auf der SowiDataNet-Startseite (Freitextsuche sowie Filtern über Facetten). Lediglich der Suchraum wird auf die jeweilige institutionelle Datensammlung begrenzt. Der Informationsaustausch zwischen Vitrine und SowiDataNet erfolgt über eine REST API. Daher sind sowohl die Metadaten als auch die Datensätze selbst direkt über die Vitrine verfügbar.

Zur Nutzung der Vitrine besteht zum einen die Möglichkeit diese durch SowiDataNet als zentralen Webservice hosten zu lassen. In diesem Fall kann die Vitrine entweder über einen iFrame in die Institutswebseiten integriert oder als eigenständige Unterseite der Instituts-Homepage aufgerufen werden. Für beide Varianten kann die Ansicht durch Anpassung einer institutsspezifischen CSS-Datei individualisiert werden, z.B. durch die Einbindung eines Logos oder die Anpassung von Farbe und Schriftart. Alternativ besteht die Möglichkeit eine lokale Installation auf den Servern des Instituts vorzunehmen und den Quellcode an die eigenen Bedarfe anzupassen.

Zusammenfassung und Schluss

Am Beispiel von SowiDataNet wurden im vorliegenden Beitrag zentrale Herausforderungen und mögliche Lösungsansätze auf dem Weg zur verstärkten Nachnutzung von Forschungsdaten in den Sozial- und Wirtschaftswissenschaften aufgezeigt. Zentraler Aspekt war dabei die Durchführung einer Anforderungsanalyse, um die Bedarfe der Forscher/innen zu erfragen und diese in die Entwicklung einfließen zu lassen. Dabei wurde auf fachspezifische, institutionelle und forschungspraktische Anforderungen eingegangen. SowiDataNet soll künftig als zentrale Anlaufstelle dienen, über die Forschungsdaten aus den Sozial- und Wirtschaftswissenschaften gesucht und nachgenutzt werden können. Durch die explizite Beteiligung der Fachcommunities und den Fokus auf die praktischen Aufgaben der Datenkuratierung wird zugleich die Rolle des institutionellen Forschungsdatenmanagements gestärkt und der Kulturwandel hin zum Data Sharing befördert.

Literaturangaben

Allianz der deutschen Wissenschaftsorganisationen. 2010. „Grundsätze zum Umgang mit Forschungsdaten.“ Online verfügbar unter http://www.allianzinitiative.de/de/handlungsfelder/forschungsdaten/grundsaetze.html. zuletzt geprüft am 13.03.2017.

Blasetti, Alessandro, Mathis Fräßdorf, Patrick J. Droß und Julian Naujoks. 2017. „Digital ist teilbar: Potenziale und Erfolgsbedingungen von Open Access und Open Data.“ *WZB-Mitteilungen* (155): 34-37. Online verfügbar unter https://wzb.eu/sites/default/files/publikationen/wzb_mitteilungen/34-37winwm155web.pdf. zuletzt geprüft am 13.03.2017.

Deutsche Forschungsgemeinschaft (DFG). 1998. „Sicherung guter wissenschaftlicher Praxis – Denkschrift.“ Erste Auflage. Weinheim: Wiley-VCH. Online verfügbar unter http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf. zuletzt geprüft am 13.03.2017.

Deutsche Forschungsgemeinschaft (DFG). 2015. „Leitlinien zum Umgang mit Forschungsdaten: Verabschiedet durch den Senat der DFG am 30. September 2015.“ Online verfügbar unter http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf. zuletzt geprüft am 13.03.2017.

Droß, Patrick J. und Monika Linne. 2016. „Sicheres und einfaches Data Sharing mit SowiDataNet: Dokumentieren – veröffentlichen – nachnutzen.“ *Bibliotheksdienst* 50 (7): 649–60. doi:10.1515/bd-2016-0079.

Fecher, Benedikt, Sascha Friesike, Marcel Hebing, Stephanie Linek und Armin Sauermann. 2015. „A Reputation Economy: Results from an Empirical Survey on Academic Data Sharing.“ *DIW Discussion Papers* 1454. Online verfügbar unter http://hdl.handle.net/10419/107687. zuletzt geprüft am 13.03.2017.

Fecher, Benedikt und Cornelius Puschmann. 2015. „Über die Grenzen der Offenheit in der Wissenschaft – Anspruch und Wirklichkeit bei der Bereitstellung und Nachnutzung von Forschungsdaten.“ *Information - Wissenschaft & Praxis* 66 (2-3): 146–50. doi:10.1515/iwp-2015-0026.

Piwowar, Heather A., Roger S. Day und Douglas B. Fridsma. 2007. „Sharing detailed research data is associated with increased citation rate.“ *PloS one* 2 (3): e308. doi:10.1371/journal.pone.0000308.

Rat für Sozial- und Wirtschaftsdaten (RatSWD). 2016. „Forschungsdatenmanagement in den Sozial-, Verhaltens- und Wirtschaftswissenschaften: Orientierungshilfen für die Beantragung und Begutachtung datengenerierender und datennutzender Forschungsprojekte.“ RatSWD Output Series 3. Online verfügbar unter http://www.ratswd.de/dl/RatSWD_Output3_Forschungsdatenmanagement.pdf. zuletzt geprüft am 13.03.2017.

Integration von Forschungsdaten in Open-Access-Publikations- und Suchsysteme

Birte Lindstädt[1]

1 ZB MED Informationszentrum Lebenswissenschaften, Köln

Zusammenfassung. Forschungsdaten sollen dem Wissenschaftssystem Open Access zur Verfügung gestellt werden, um Transparenz und Nachnutzbarkeit zu ermöglichen. Eine Vielzahl dezentraler, teilweise fachspezifischer Infrastrukturen zu Speicherung, Archivierung, Nachweis und Zugriff auf Forschungsdaten existieren bereits bzw. sind im Aufbau begriffen.
Es geht jedoch nicht nur um die isolierte Betrachtung von Forschungsdaten, sondern um ihre Integration in vorhandene bzw. aufzubauende Publikations- und Suchsysteme. Als zitierfähiger wissenschaftlicher Output sollten Forschungsdaten in ihre Zusammenhänge mit Textpublikationen wie Journalartikeln oder mit anderen Forschungsdatensätzen gestellt werden. Diese Zusammenhänge sind wiederum in den Publikationssystemen abzubilden. Nach der Publikation ist auch dafür zu sorgen, dass der Nachweis der Forschungsdaten in relevante Suchportale integriert wird.
Als Beispiel eines solchen integrierten Ansatzes wird das Konzept von ZB MED - Informationszentrum Lebenswissenschaften dargestellt. PUBLISSO als Open-Access-Publikationsportal und LIVIVO als lebenswissenschaftliches Suchportal bilden die Grundlagen dieses Ansatzes. In beiden künftig miteinander verzahnten Systemen werden die Spezifika lebenswissenschaftlicher Forschungsdaten berücksichtigt, sofern sie bei der Publikation und Suche eine Rolle spielen.

Schlagwörter. Open Access, Forschungsdaten, Publikationsportal, Suchportal, Lebenswissenschaften

Besonderheiten lebenswissenschaftlicher Forschungsdaten

Bevor auf das Publikationsportal eingegangen wird, werden kurz die Spezifika lebenswissenschaftlicher Forschungsdaten dargestellt. Nach Definition von ZB MED umfassen die Lebenswissenschaften die Fächer Medizin, Gesundheitswesen, Umwelt-, Ernährungs- und Agrarwissenschaften. Eine Erweiterung dieser Kerndisziplinen stellen beispielsweise die Biologie oder die Psychologie dar, für die ebenfalls Publikationsdienstleistungen erbracht werden können.

Die Forschungsdaten der relevanten Fächer unterscheiden sich inhaltlich in hohem Maße: auf der einen Seite stehen medizinische, meist personenbezogene Daten wie Blutprobenergebnisse, Röntgenbilder oder Ergebnisse von Ernährungs- oder Klinischen Studien. Auf der anderen Seite Bodenmesswerte, Emissionswerte in Tierställen oder Wetterdaten.

Von der Art der Daten finden sich jedoch durchaus Gemeinsamkeiten für den Umgang bei der Verarbeitung, Speicherung und Publikation: Bilder (z.B. MRT, Satellitenaufnahmen), Videos (z.B. Operationsfilme, Interviews), statistische Daten, sog. Big Data (z.B. genomische Sequenzen, Daten von Landmaschinen) oder geräteabhängige Daten (z.B. Röntgengerät, Emissionsmessgerät).

Eine wesentliche Besonderheit in der Medizin, aber auch teils in den Ernährungswissenschaften, sind personenbezogene Daten, die entsprechenden Datenschutzbedingungen unterliegen. Unter anderem führen diese zu folgenden Rahmenbedingungen, die im projektbezogenen Forschungsdatenmanagement berücksichtigt werden müssen:

- Schutz persönlicher Interessen (Personenbezug von Phänotypdaten, *omics-Daten, Biomaterial) / Pflicht zur Anonymisierung,
- gesetzliche Aufbewahrungsfristen (minimal und maximal),
- rechtlicher Rahmen: Geflecht aus MBO-Ä (ärztliche Schweigepflicht), Bundes-/ Landesdatenschutzgesetz, Gesetz zur wirtschaftlichen Sicherung der Krankenhäuser und zur Regelung der Krankenhauspflegesätze,
- ethische Aspekte bzgl. Erhebung und Nutzung,
- komplexes Regelwerk bzgl. klinischer Studien („Gute klinische Praxis“),
- Schutz kommerzieller Interessen (Innovationsschutz),
- proprietäre Formate.

Daher spielen für medizinische Forschungsprojekte Datenschutzkonzepte, die Patienteninformationen, Einwilligungsverfahren, Pseudonymisierungs- und Anonymisierungsverfahren berücksichtigen, eine große Rolle. Diese sind in der Regel Voraussetzung, um überhaupt eine Datenpublikation anstreben zu können.

Aufgrund dieser Rahmenbedingungen stehen beim Teilen von Daten in der Medizin auch nicht unbedingt “offene” Daten im Sinne des Open Access im Vordergrund, sondern es spielen unterschiedliche Zugangsweisen zu Forschungsdaten eine Rolle:

- Teilzugang nach Anfrage,
- Modell „Transferstelle“ (z.B. Transferstelle für Daten und Biomaterialienmanagement, Universitätsmedizin Greifswald),
- Zugang zu faktisch anonymisierten Daten (DeStatis, Statistisches Bundesamt: Gesundheit)
- Zugang “on Screen”,
- Zugang „remote“: „Anfrage einschicken“ (z.B. Informationssystem Versorgungsdaten (Datentransparenz) Daten, DaTraV / DIMDI),
- Zugang „on Site“: Auswertung vor Ort (z.B. DaTraV Forscherplatz).

Dies gilt selbstverständlich auch für sensible Daten anderer lebenswissenschaftlicher Disziplinen wie beispielsweise ökologische Daten für bedrohte Spezies.

In der Regel müssen die genannten Besonderheiten bereits vor einer Datenpublikation berücksichtigt und entsprechende Anonymisierungsverfahren etc. durchgeführt worden sein. Sie spielen für das projektbezogene Forschungsdatenmanagement folglich eine große Rolle. Bei der eigentlichen Publikation von und der Recherche nach lebenswissenschaftlichen Forschungsdaten kommen u.a. folgende fachbezogene Aspekte zum Tragen:

- Berücksichtigung von Fach-Thesauri bei der Verschlagwortung bzw. bei der Suche (Medical Subject Headings MeSH, Multilingual Agricultural Thesauraus AGROVOC, Umweltthesaurus UMTHES),
- fachbezogene Metadaten, z.B. geographischer Ort, Datenerhebungsform
- Auswahl relevanter Fachgruppen aus fachlichen Klassifikationen, z.B. Dewey Decimal Classification DDC,
- Auswahl einer geeigneten “offenen” Lizenz für die Publikation
- Vergabe eines DOI.

Das ZB MED Publikationsportal PUBLISSO

PUBLISSO umfasst verschiedene Publikationsplattformen: Zum einen Plattformen für die html-basierte Open-Access-Erstpublikationen von Journalartikeln, Kongressbeiträgen und Büchern bzw. Buchkapiteln (Publikationsplattform Lebenswissenschaften „gold“) und zum anderen das Fachrepositorium Lebenswissenschaften für Zweitveröffentlichungen als PDF und die Publikation unterschiedlicher Dateien, z.B. Forschungsdaten. Es besteht jeweils die Möglichkeit, zu Publikationen gehörende Forschungsdaten parallel zu veröffentlichen und auf diese zu verweisen, wobei die Datenpublikation im Fachrepositorium Lebenswissenschaften oder in anderen Repositorien erfolgen kann (z.B. durch die Kooperation von ZB MED mit Dryad, einem englischsprachigen Forschungsdaten-Repositorium aus den USA). Die Forschungsdaten stellen eine eigenständige Publikation dar und sind mit den zugehörigen Publikationen verknüpft. Durch die gegenseitige Verlinkung ist es beispielsweise möglich, Forschungsdaten, die zu einem Journalartikel gehören, beim Lesen des Volltexts direkt aufzurufen. Der Volltext kann zudem auch über die publizierten Forschungsdaten gefunden werden.

Die folgende Graphik gibt einen Überblick über die Struktur des Publikationsportals (vgl. Abb. 1).

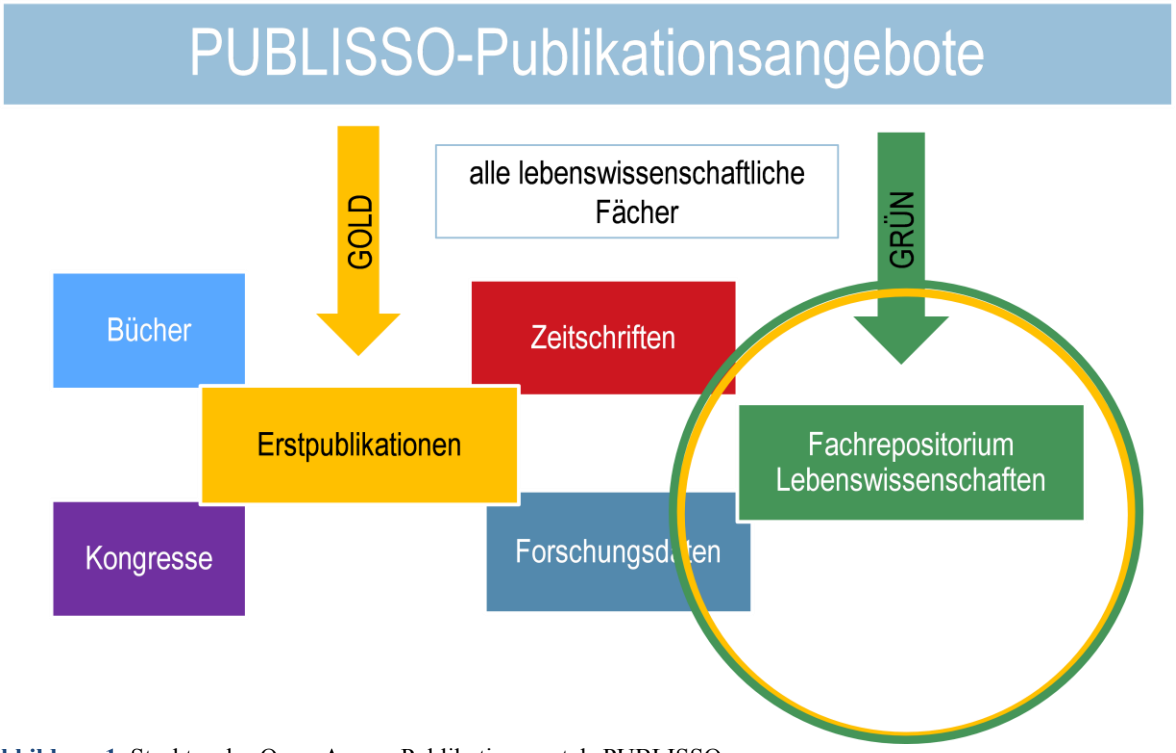


Abbildung 1. Struktur des Open-Access Publikationsportals PUBLISSO

Die Kernangebote der Publikationsplattform sind derzeit die Produkte German Medical Science (GMS -medizinische Open-Access-Zeitschriften, -Kongressabstracts und -Forschungsberichte „gold“) und Living Handbooks (Monografien „gold“). Es ist geplant, die beiden Angebote in den nächsten Jahren zusammenzuführen und auf alle Fächer der Lebenswissenschaften auszuweiten.

Die „Living Handbooks“ erlauben es, Forschungsergebnisse kapitelweise und zeitnah zu publizieren und regelmäßig zu aktualisieren, ohne von einem langwierigen Print-Publikationsprozess und der gleichzeitigen Fertigstellung aller eingereichten Kapitel abhängig zu sein. Dadurch sind

die Bücher weniger statisch, sondern „leben“. Ihre Attraktivität erhalten sie auch durch die Möglichkeit der Einbindung von multimedialen Inhalten, Forschungsdaten etc. Dafür wurde das Content-Management-System Drupal (Open Source) an die Anforderungen des wissenschaftlichen Publizierens angepasst. Mit „Living Textbooks of Hand Surgery“ wurde bereits der Prototyp eines Open-Access-Handbuchs (Living Handbooks) entwickelt.

Künftig wird die Publikationsplattform „gold“, so ausgebaut, dass darauf Zeitschriften, Kongresse und Bücher sowie - mittelfristig --auch zugehörige Forschungsdaten aus einer Hand publiziert werden können, um jedem der Fächer im Zuständigkeitsbereich von ZB MED die gleichen Publikationsmöglichkeiten zu bieten. Die Sichtbarkeit der Publikationen wird zusätzlich gestärkt und Querverweise, auch interdisziplinär, ermöglicht.

Fachrepositorium Lebenswissenschaften

Das Fachrepositorium Lebenswissenschaften wird gemeinsam mit dem technischen Kooperationspartner Hochschulbibliothekszentrum Nordrhein-Westfalen (hbz) aus- und aufgebaut und basiert auf der technischen Grundlage Fedora bzw. Drupal für die Ansicht. Neu entwickelte Erfassungsmasken erlauben es unter anderem, neben Monographien auch weitere Publikationsformate wie selbstständige und unselbstständige Literatur aufzunehmen und zweitzuveröffentlichen. Auch Video-, Bild- und Audiodateien können aufgenommen und über einen integrierten Viewer direkt abgespielt oder dargestellt werden. Die technische Grundlage erlaubt es zudem, weitere Forschungsdaten (singulär sowie in Verbindung mit einer Publikation) zu publizieren. Sofern noch nicht vorhanden, erhalten alle aufgenommenen Publikationen einen persistenten Identifikator (DOI-Digital Object Identifier).

Ziele und Strategien für die Publikation von Forschungsdaten im Fachrepositorium Lebenswissenschaften

Die Publikationsmöglichkeiten für Forschungsdaten im Rahmen von PUBLISSO bauen auf dem strategischen Ziel auf, bereits vorhandene Infrastrukturen zur Datenpublikation in den Lebenswissenschaften aufzuzeigen und an den Stellen eigene Angebote aufzubauen, wo Lücken identifiziert werden. Dies bezieht sich beispielsweise auf den sog. long tail der Forschungsdaten, also Daten, die ein geringes Datenvolumen aufweisen, in verschiedenen Datenformaten vorliegen und somit nur schwer standardisierbar sind, aber auch auf lebenswissenschaftliche Teildisziplinen, in denen Möglichkeiten zur Datenarchivierung und -publikation weitgehend fehlen.

Metadatenschema für die Publikation von Forschungsdaten

Als Orientierungshilfe für die Entwicklung eines Metadatenschema zur Erfassung von Forschungsdaten wurden zunächst Metadatenschemata existierender (Daten-) Repositorien analysiert und auf eine Übertragbarkeit für die vorliegende Aufgabenstellung geprüft.

Eine der wichtigsten Quellen stellte das DataCite-Metadatenschema dar, das aktuell in der Version 4.0 vom September 2016 vorliegt. DataCite ist eine internationale Organisation zur Vergabe von DOIs für Forschungsdaten, bei der ZB MED Mitglied ist. Das bei DataCite verwen-

dete Metadatenschema hat keinen fachlichen Fokus, da die Referenzierung verschiedenster Objekte aus allen Disziplinen angestrebt wird. Es versucht jedoch auch fachspezifische Aspekte einzubinden, z.B. durch das Feld GeoLocation. Außerdem setzt es die Hürde zur Registrierung und damit Publikation von Forschungsdaten durch lediglich sechs verpflichtende Metadaten recht niedrig an und unterscheidet darüber hinaus in „empfohlene“ und „optionale“ Felder.

Um den fachspezifischen Anforderungen lebenswissenschaftlicher Forschungsdaten gerecht zu werden, wurde in der Entwicklung des Metadatenschemas für das Fachrepositorium eine Reihe fachspezifischer Metadaten und Standards berücksichtigt wie z.B. die fachliche Zuordnung, geographische Angaben oder der Thesaurus AGROVOC für die Verschlagwortung. Das Ziel der Nachnutzbarkeit wird durch Kriterien wie „Beschreibung“ oder „Hinweise zur Nutzung“ als verpflichtende Metadatenangaben erreicht, unabhängig davon, ob eine Textpublikation existiert, die ebenfalls Beschreibungen liefert.

Bei einer Forschungsdatenpublikation im Fachrepositorium Lebenswissenschaften kann darüber hinaus ein komplexes Beziehungsnetzwerk abgebildet werden: In den Metadaten ist verzeichnet, ob die Forschungsdaten zu einer Textpublikation gehören, ob sie Teil eines größeren Datensatzes sind oder ob es mehrere Versionen gibt. Bei Forschungsdatensätzen, die aus mehreren Teilen bzw. Dateien oder zugehörigen Dateien wie Beschreibungen bestehen, ist es möglich, alle Bestandteile unter einen Metadateneintrag zu stellen. Alle Verknüpfungen werden möglich, da die Forschungsdaten mit eigenem DOI im Fachrepositorium Lebenswissenschaften abgelegt und auch die zugehörigen Publikationen nach Möglichkeit mit einem persistenten Identifier in die Metadaten aufgenommen werden. Die DOI-Registrierung erfolgt über DataCite.

Auf der Grundlage der genannten Aspekte und insbesondere auf dem DataCite-Schema, aber auch auf anderen Beispielen (wie u.a. PANGAEA Data Publisher for Earth & Environmental Science) sowie den Anforderungen der Forschungsdaten in den lebenswissenschaftlichen Disziplinen fußt das für das Fachrepositorium Lebenswissenschaften entwickelte Metadatenschema:

Tabelle 1. Metadatenschema für Forschungsdaten im Fachrepositorium Lebenswissenschaften (Pflichtfelder).

Metadatum (übergeordneter Begriff)	Feldname	Feldname (untergeordnet)
Titel	Titel	
Urheberschaft	Autor*in (Linked Data: Gemeinsame Normdatei GND)	Nachname
		Vorname
		ORCID (optional)
		Affiliation (optional)
	Körperschaft (wenn kein Autor vorhanden)	
Dateiupload	Hochzuladende Datei	
	Format (xls, jpeg, etc.)	
	Medientyp (Bild, Video, Software, etc.)	
	Größe	
	Zugriffsrechte (open access, Embargo)	Embargofristende
	Copyrightjahr	
	Lizenz	Empfehlung: Open Data Commons Open Database License (ODbL)
	DOI	Neu
		Vorhanden
	Zuletzt hochgeladen	

Erschließung	Abstract	Sprache
	Fachgruppenzuordnung (Medizin, Umwelt, Agrar, Ernährung, interdisziplinär)	
	DDC-Klassifikation (Auswahl lebenswissenschaftlicher Fachgruppen)	
	Sprache (dt.,engl., frz., span., ital.)	

Tabelle 2. Metadatenschema für Forschungsdaten im Fachrepositorium Lebenswissenschaften (optionale Felder).

Metadatum (übergeordneter Begriff)	Feldname	Feldname (untergeordnet)
Beteiligte	Beteiligte Personen (Linked Data: Gemeinsame Normdatei GND)	ORCID
		Affiliation
	Förderer	Förder-ID
Erfassung	Schlagworte (Linked Data: AGROVOC)	Sprache
	Datenerhebungsform (z.B. Probe, Interview, Genomsequenzierung)	
	Erhebungszeit	Zeitpunkt
		Zeitraum
	Erfassungsort	Koordinaten (Point)
		Koordinaten (Box)
Externe Referenzen	Verwendete Publikationen	
	Zugehörige Publikationen	
	Versionen	Vorgänger
		Nachfolger

Dieses Schema stellt die aktuelle Basis für die Publikation von lebenswissenschaftlichen Forschungsdaten dar und soll in Kooperation mit den fachlichen Communities weiterentwickelt werden, um spezifischer auf die disziplinabhängigen Bedarfe eingehen zu können.

Dazu beteiligt sich ZB MED an Forschungsprojekten, wie beispielsweise an dem “Verbundvorhaben Emissionsminderung Nutztierhaltung - Einzelmaßnahmen”, in dem das Bundeslandwirtschaftsministerium technische Maßnahmen in Tierställen zur Emissionsminderung erproben lassen möchte. Hierbei soll ZB MED die Erstellung eines Datenmanagementplans sowie die Publikation der Forschungsdaten im Fachrepositorium Lebenswissenschaften, in der Regel Messergebnisse, übernehmen.

Zur Verbesserung der Sichtbarkeit der Publikationen werden alle Inhalte des Fachrepositoriums Lebenswissenschaften in das ZB MED Suchportal LIVIVO übernommen. Eine OAI-Schnittstelle erlaubt darüber hinaus das harvesten durch andere Systeme wie beispielsweise Bielefeld Academic Search Engine BASE.

Das ZB MED Suchportal LIVIVO

LIVIVO - das ZB MED-Suchportal für Lebenswissenschaften (https://www.livivo.de) bietet umfassende und kostenfreie Recherchewerkzeuge für die interdisziplinäre Literaturversorgung in den Fächern der Lebenswissenschaften.

Die Datengrundlage von LIVIVO bilden qualitätsgeprüfte und kuratierte Datenquellen. Sie umfassen ein großes Spektrum der Literatur in den Lebenswissenschaften. Wichtige Datenquellen sind die ZB MED-Kataloge und Artikeldaten (CCMED, CCGREEN), German Medical Science,

Medline (PubMed) , AGRIS, AGRICOLA, Fachkataloge, Verlagsdaten, der Dissonline-Dienst der Deutschen Nationalbibliothek und ein fachspezifischer Auszug von BASE, der Bielefeld Academic Search Engine.

Zur effektiven Verarbeitung und Anreicherung dieser Daten wurde 2016 das ZB MED-Knowledge Environment (ZB MED-KE) eingeführt, ein universelles Instrument zum Import von Metadaten und Volltexten sowie deren Verarbeitung und Nachnutzung. Es stellt somit die umfassende Datenbasis für LIVIVO dar und dient gleichzeitig als unverzichtbarer Bestandteil der Informationsdienste und der Forschungsaktivitäten an ZB MED (s. Abb. 2).

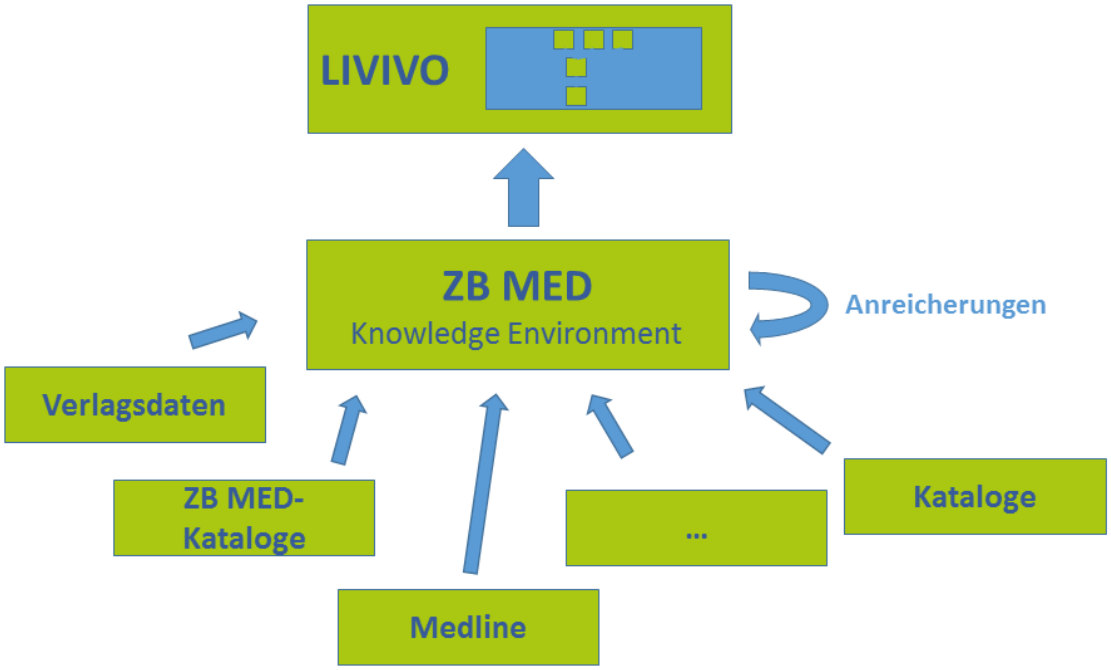


Abbildung 2. Das ZB MED-Suchportal LIVIVO

Die auf den PUBLISSO-Plattformen publizierten oder eingebundenen Forschungsdaten werden über Schnittstellen bzw. Konverter in das ZB MED-KE eingespeist und im Suchportal LIVIVO nachgewiesen. Sie sind somit breit auffindbar. LIVIVO stellt aktuell Such- bzw. Filtermöglichkeiten für mit Textpublikationen verknüpfte Forschungsdaten zur Verfügung.

Zurzeit sind bereits die Datensätze des Repositoriums DRYAD in Verknüpfung mit den korrespondierenden Volltexten eingebunden. Künftig sollen auch singuläre Forschungsdaten suchbar werden.

Das Angebot recherchierbarer Forschungsdaten soll deutlich erhöht werden, indem die Daten weiterer DataCite-Datenzentren sowie anderer qualitätsgeprüfter Forschungsdatenrepositorien (z.B. auf der Grundlage des Meta-Portal für Forschungsdatenrepositorien re3data.org) eingebunden werden. Der über DataCite vergebene DOI und die somit vorhandenen Metadaten sind ein wichtiges Kriterium, Forschungsdaten in LIVIVO nachweisen zu können. Daher stellt das Vorhandensein eines persistenten Identifikators ein wichtiges Qualitätskriterium dar.

Durch die in LIVIVO integrierte semantische Erschließung ist außerdem eine Kontextualisierung der Forschungsdaten möglich. “Recherchen werden durch linguistische Verfahren aufbereitet und semantisch mit sprachunabhängigen Konzepten annotiert. Als Fachthesauri werden für die Fächer Medizin und Gesundheit die Medical Subject Headings, für die Ernährungs-, Umwelt- und

Agrarwissenschaften AGROVOC und UMTHES verwendet. Durch das Abbilden der Fachbegriffe in unterschiedlichen Sprachen auf ihre linguistischen Repräsentationen können sprachübergreifend Suchergebnisse gefunden werden. Gleichzeitig wird die Suche nach Wortvarianten und Synonymen ermöglicht."

Abschlussbemerkung

Deutlich wird, dass die Verfügbarkeit und Publikation von lebenswissenschaftlichen Forschungsdaten in Informationsinfrastrukturen eingebettet sein muss, die sowohl die bibliothekarischen als auch die fachspezifischen Anforderungen an die Daten berücksichtigen. Insofern gilt es künftig die Zusammenarbeit von ZB MED und der lebenswissenschaftlichen Forschung noch enger zu verzahnen und die Publikations- und Suchportale gemeinsam mit den Forschenden weiterzuentwickeln.

Literaturangaben

Arning, Ursula, Birte Lindstädt, und Jasmin Schmitz. 2016. „PUBLISSO: „Das Open-Access-Publikationsportal für die Lebenswissenschaften", *GMS Medizin - Bibliothek - Information* 16 (3). doi: 10.3205/mbi000370.

Bielefeld Academic Search Engine. Online verfügbar unter https://www.base-search.net. zuletzt geprüft am 06.03.2017.

Data Cite: „Meta Data Scheme 4.0". Online verfügbar unter https://schema.datacite.org/meta/kernel-4.0/doc/DataCite-MetadataKernel_v4.0.pdf. zuletzt geprüft am 03.03.2017.

Deutsche Nationalbibliothek: „Katalog der DNB". Online verfügbar unter http://search.dissonline.de/. zuletzt geprüft am 06.03.2017.

Dryad Data Repository. Online verfügbar unter http://datadryad.org/. zuletzt geprüft am 06.03.2017.

Food and Agriculture Organisation of the United Nation: „agris". Online verfügbar unter; http://agris.fao.org/. zuletzt geprüft am 06.03.2017.

German Medical Science. Online verfügbar unter http://www.egms.de/. zuletzt geprüft am 06.03.2017.

German Medical Science: „GMS Books". Online verfügbar unter http://www.gms-books.de. zuletzt geprüft am 06.03.2017.

National Agricultural Library: „NAL Catalogue - agricola“. Online verfügbar unter http://agricola.nal.usda.gov. zuletzt geprüft am 06.03.2017.

NCBI: „PubMed.gov“. Online verfügbar unter https://www.ncbi.nlm.nih.gov/pubmed. zuletzt geprüft am 06.03.2017.

Open Data Commons: “Legal Tools for Open Data”. Online verfügbar unter 06.03.2017, http://opendatacommons.org/licenses/odbl/. zuletzt geprüft am 06.03.2017.

Christioph Poley. 2016. ”LIVIVO - “Neue Herausforderungen an das ZB MED-Suchportal für Lebenswissenschaften”. *GMS Medizin - Bibliothek - Information* 16(3), doi: 10.3205/mbi000376

Reifegradmodelle für ein integriertes Forschungsdatenmanagement in multidisziplinären Forschungsorganisationen

Jonas Oppenländer[1], Falko Glöckler[2], Jana Hoffmann[3], Claudia Müller-Birn[4]

1,4 Institut für Informatik, FU Berlin
2,3 Museum für Naturkunde, Berlin

Zusammenfassung. Forschungsdatenmanagement (FDM) ist ein relativ junges Forschungsgebiet in Deutschland. Der Fokus der FDM-Bestrebungen liegt oft auf dem Ausbau der Infrastruktur und der effizienten Speicherung der Daten einzelner Wissenschaftsdisziplinen oder Communities. Ein integriertes Forschungsdatenmanagement sollte jedoch an den Forschungsprozessen ausgerichtet sein und die Bedeutung der Daten über disziplinäre Grenzen hinweg sicherstellen.

Multidisziplinäre Forschungsorganisationen, wie beispielsweise Forschungsmuseen, zeichnen sich durch sehr heterogene Forschungsdaten aus, die oft in stark voneinander abweichenden Forschungspraktiken und Analysemethoden erzeugt werden. Damit wird der Aufbau eines nachhaltigen Forschungsdatenmanagements erschwert. Es fehlt eine ganzheitliche Betrachtung und Modellierung des FDMs, welches als Grundlage für strategische Entscheidungen genutzt werden kann. Eine solche Betrachtung sollte konkrete Handlungsmöglichkeiten in Abhängigkeit vom Entwicklungsstand des Forschungsdatenmanagements der jeweiligen Forschungsbereiche und Forschungsprojekte aufzeigen. Reifegradmodelle, wie z.B. das Capability Maturity Model (CMM), können diese Lücke schließen. Sie ermöglichen es (1) bestehende Praktiken in Organisationen zu erheben und zu bewerten; (2) neue Handlungsoptionen und Entwicklungsmöglichkeiten anhand ihres Fähigkeitsgrades abzuleiten; sowie (3) notwendige fachliche Ressourcen und weitere Mittel zu identifizieren.

In dem Forschungsvorhaben int.FDM soll ein Reifegradmodell für integriertes Forschungsdatenmanagement entwickelt werden, welches in multidisziplinären Forschungsorganisationen eingesetzt werden kann. Als paradigmatische Struktur für multidisziplinäre Forschungsorganisationen sollen in diesem Vorhaben Forschungsprozesse am Museum für Naturkunde (MfN), dem Leibniz-Institut für Evolutions- und Biodiversitätsforschung, dienen. Dieser Beitrag wird eine Übersicht über bestehende Reifegradmodelle im FDM geben und das geplante Forschungsvorhaben näher beschreiben.

Schlagwörter. Forschungsdatenmanagement, Forschungsdaten, multidisziplinäre Forschungsorganisationen, heterogene Daten, Reifegradmodell

Einleitung

In den letzten Jahrzehnten wurde die Forschungspraxis innerhalb aller Disziplinen zunehmend datenintensiv. Diese methodische Änderung wird oft als das „vierte Paradigma“ der wissenschaftlichen Forschung bezeichnet (Hey 2009). Daten sind die Grundlage für neues Wissen. Sie können als Synchronisationspunkt angesehen werden, an dem sich unterschiedliche Fachdisziplinen mit ihren teilweise sehr unterschiedlichen Fragestellungen treffen. Wie kann aber das Wissen über die Daten und ihre Bedeutung über disziplinäre Grenzen hinweg sichergestellt werden? Ein integrier-

tes Forschungsdatenmanagement (FDM) ist eine zentrale Basis dafür. In dessen Zentrum steht die gesamte Praxis des Umgangs mit Forschungsdaten (Kindling und Schirmbacher 2013). FDM ist in Deutschland im Vergleich zum angloamerikanischen Raum ein relativ junges Forschungs- bzw. Themengebiet. Der Fokus der FDM-Bestrebungen liegt oft auf der Infrastruktur und Speicherung der Daten sowie der Bereitstellung und Veröffentlichung für spezielle Wissenschaftsdisziplinen und Communities.

Überblick über bestehende Ansätze des FDMs

Beim Aufbau eines nachhaltigen Forschungsdatenmanagements sehen sich Forschungsorganisationen mit einer Vielzahl von nationalen und internationalen Ansätzen konfrontiert, die es innerhalb der eigenen Strategie zu berücksichtigen gilt. Ein Überblick findet sich in Burger et al. (2013) und wird in Tabelle 1 anhand von Beispielen genauer erläutert.

Tabelle 1. Übersicht der Handlungsempfehlungen und Betrachtungsebenen.

Schwerpunkt	Ansatz	Beispiel
International	Handlungs-empfehlungen	European Open Science Agenda (RTD 2016)
	Richtlinien & Policies	Open Science Policy (Open Science Policy Platform 2016)
		Guidelines on FAIR Data Management in Horizon 2020 (EC 2016)
	Prinzipien & Grundsätze	Fair Data Principles (FORCE11 o.J.)
National	Handlungs-empfehlungen	Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland (RfII 2016)
		Sicherung guter wissenschaftlicher Praxis. Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft“ (DFG 2013)
		Empfehlungen der Hochschulrektorenkonferenz (HRK 2015)
	Richtlinien & Policies	Leitlinien zum Umgang mit Forschungsdaten (DFG 2015)
		Richtlinien für Zuwendungsanträge auf Ausgabenbasis (BMBF o. J.)
	Prinzipien & Grundsätze	Grundsätze zum Umgang mit Forschungsdaten (Allianzinitiative 2010)
	Gesetze	Bundesdatenschutzgesetz (Recht 2016)
Organisations-bezogen	Handlungs-empfehlungen	Handlungsempfehlungen in Ergänzung zu den Grundsätzen zum Umgang mit Forschungsdaten (HU Berlin 2014b)
	Richtlinien & Policies	Leitfaden zum Forschungsdaten-Management (Ludwig und Enke 2013)
		Leitlinien zum Umgang mit digitalen Forschungsdaten an der TU Darmstadt (TU Darmstadt 2015)
	Prinzipien & Grundsätze	Grundsätze zum Umgang mit Forschungsdaten an der Humboldt-Universität zu Berlin (Forschungsdaten-Policy) (HU Berlin 2014a)
	Checklisten	Checkliste zum Forschungsdatenmanagement (Enke et al. 2011)
	Lebenszyklus-modelle	DCC Curation Lifecycle Model (Higgins 2008)
	Datenmanage-mentpläne	Anleitung zur Erstellung eines Datenmanagementplans (HU Berlin 2014c)

Auf nationaler Ebene regeln Fördereinrichtungen wie die DFG den Umgang mit Forschungsdaten im Rahmen ihrer Förderrichtlinien[1]. Die Herausforderung besteht darin, die Empfehlungen und Richtlinien im organisationsbezogenen FDM zu verankern, um sowohl die Anforderungen der Fördergeber, als auch die der eigenen Forschungsorganisation adäquat aufeinander abzustimmen. Die zahlreichen existierenden Leitlinien und Empfehlungen helfen zwar bei dem Aufbau eines strategischen FDM-Ansatzes, sind aber als Werkzeug zur kontinuierlichen Erfassung des Zustands des FDMs in einer multidisziplinären Forschungsorganisation unzureichend (Kindling et al. 2013). Datenmanagementpläne (DMP) hingegen sind eher vorhabenspezifisch und oft sehr stark auf die konkreten Vorgaben der Förderorganisationen ausgerichtet. Bestehende institutionelle Besonderheiten können daher nur schwer oder gar nicht in den Datenmanagementplänen berücksichtigt werden.

Herausforderung des FDMs in multidisziplinären Forschungsorganisationen

An multidisziplinären Organisationen, wie Hochschulen und Forschungsmuseen, stellt das Management von Forschungsdaten eine besondere Herausforderung dar. Es stellt sich die Frage, wie man den Zustand des Forschungsdatenmanagements an multidisziplinären Einrichtungen erfassen und bewerten kann. Die Herausforderungen bestehen in (1) den komplexen Anforderungen der unterschiedlichen Stakeholder (interner und externer Akteure) und Zielgruppen, (2) der Datenheterogenität im Hinblick auf Volumen, Komplexität, Langlebigkeit und Strukturierungsgrad, (3) den großen Unterschieden in der Datengenerierung, den Analyseverfahren und Forschungspraktiken sowie (4) den unterschiedlichen Standards. Es fehlt eine ganzheitliche Betrachtung und Modellierung des FDMs unter Berücksichtigung aller Stakeholder, Prozessarten und Entwicklungsstufen des FDMs als Grundlage für strategische Entscheidungen.

Reifegradmodelle im Forschungsdatenmanagement

Reifegradmodelle können als ein Lösungsansatz für eine umfassende und kontinuierlich nachgeführte Modellierung von Organisationsstrukturen des FDMs in einer Forschungsorganisation genutzt werden. Ausgangspunkt der Reifegradmodelle ist die stichtagsbezogene oder regemäßige Bewertung der Prozesse einer Forschungsorganisation. Reifegradmodelle sind auch ein Werkzeug für das Management kultureller Veränderungsprozesse. Das bekannteste Reifegradmodell wird im folgenden Kapitel kurz beschrieben. Anschließend werden speziell an das FDM angepasste Reifegradmodelle vorgestellt.

Grundlegende Begriffe

Die Grundlage aller später entwickelten Reifegrademodelle bildet das Capability Maturity Model (CMM). Das CMM wurde am Software Engineering Institute der Carnegie Mellon Universität in Pittsburgh im Jahre 1984 entwickelt (Paulk et al. 1993). Das ursprüngliche Ziel des Modells war die Bewertung der Qualität der Softwareentwicklungsprozesse von Organisationen. Das CMM

1 Vgl. https://www.cms.hu-berlin.de/de/dl/dataman/arbeiten/dmp_erstellen/foerderer.

kann zur statischen Messung des Reifegrads (staged model) oder zur kontinuierlichen Verbesserung (continuous model) der Organisation eingesetzt werden. Die statische Messung der Reifegrade (maturity levels) findet vor allem bei der Betrachtung der gesamten Organisation Anwendung, während sich die kontinuierliche Verbesserung des Reifegrads auf die Ermittlung der Fähigkeitsgrade (capability levels) in den einzelnen Prozessbereichen fokussiert. Im CMM wurden fünf aufeinander aufbauende Reifegrade bzw. Fähigkeitsgrade definiert (initial, repeatable, defined, managed, optimizing). Jedem Reifegrad sind Prozesse zugeordnet, welche wiederum in verschiedene Prozessbereiche (process areas) gruppiert sind. Den Prozessen sind generische und spezifische Ziele (generic goals and specific goals) zugeordnet. Die Ziele werden jeweils durch mehrere Praktiken (practices) beschrieben. Generische Ziele (und deren zugehörigen Praktiken) beziehen sich auf mehrere Prozessbereiche. Der Reifegrad einer Organisation in den definierten Prozessbereichen kann durch standardisierte Erhebungsmethoden, z.B. SCAMPI (SEI 2011), bestimmt werden. Die Messung des Reifegrads erfolgt von den Praktiken über die Prozesse hin zum Prozessbereich. Die spezifischen Ziele müssen im CMM erfüllt sein, um den Reifegrad in dem jeweiligen Prozessbereich nachzuweisen.

Das CMM wurde im Jahre 2003 in das Capability Maturity Model Integration (CMMI) erweitert, um die bessere Übertragbarkeit auf Organisationen in der (Software-)Entwicklung, Beschaffung und dem Service zu gewährleisten. Es wurde in zehn Sprachen übersetzt und in Organisationen in über 100 Ländern angewandt[2].

Vergleich bestehender Reifegradmodelle

Das CMM wurde seit seiner Entwicklung auf unterschiedliche Anwendungsbereiche übertragen bzw. angepasst. Beispiele sind das ITIL[3]-Maturity-Model für IT-Service-Organisationen (ITIL 2013), der Bereich des Projektmanagements (Crawford 2007), das Supply Management (Mettler 2010) und andere Businessprozesse (OMG 2008). Speziell für das Datenmanagement entwickelte Reifegradmodelle wurden im angloamerikanischen Raum durch Organisationen wie Microsoft und Fannie Mae implementiert[4]. Die Nutzung von Reifegradmodellen im Forschungsdatenmanagement befindet sich in Deutschland jedoch noch am Anfang.

Es wurden insgesamt neun Reifegradmodelle im Forschungsdatenmanagement durch eine bestehende Übersicht (Becker et al. 2009) und weitere Recherchen (Online-Suchmaschine) identifiziert (Tabelle 2). Diese Modelle wurden in unterschiedlichen Kontexten entwickelt und von privaten Instituten (DMM), Unternehmen (DGCMM, MIKE2.0-IMM), aber auch für Organisationen (CMM-RDM, DSMM) und auf nationaler Ebene (RDMCMG) herausgegeben.

2 Vgl. http://cmmiinstitute.com.
3 Information Technology Infrastructure Library
4 Vgl. http://cmmiinstitute.com/resource-tag/case-study.

Tabelle 2. Übersicht bestehender Reifegradmodelle für das FDM.

Reifegradmodell	Akronym	Herausgeber	Quelle
Data Management Maturity	DMM	CMMI Institute	CMMI Institute 2014
Data Management Capability Assessment Model	DCAM	Enterprise Data Management Council (EDM)	edmcouncil.org
CMM for Scientific Data Management	CMM-RDM	Syracuse University	Qin et al. 2014
IBM Data Governance Council Maturity Model	DGCMM	IBM	IBM 2007
Method for an Integrated Knowledge Environment (MIKE2.0) Information Maturity Model	MIKE2.0-IMM	BearingPoint	openmethodology.org
Data Management Practice Maturity	DMPM	Data Blueprint & Virginia Commonwealth University	Aiken et al. 2007
Scientific Data Stewardship Maturity Matrix	DSMM	North Carolina State University	Peng et al. 2015
Research Data Management Capability Maturity Model	RDMCMG	Australian National Data Service (ANDS)	ANDS 2017
Managing Research Data Project Maturity Model	JISCMRD	Joint Information Systems Committee (JISC)	Fowler 2012

Die Prozessbereiche der Modelle wurden aufgrund ihres Wortlauts analysiert und verglichen (Tabelle 3). Die Kategorien wurden dabei iterativ durch offenes Kodieren entwickelt.

Die Reifegradmodelle sind stark auf Infrastruktur (Technologie, Architektur) fokussiert. Strategie, Governance und Datenqualität nehmen ebenfalls einen hohen Stellenwert in den Modellen ein. Die interne Organisation wird von vier der neun Modelle mit mindestens einem eigenen Prozessbereich berücksichtigt. Ein eigenständiger Prozessbereich, der die Stakeholder berücksichtigt, findet sich jedoch nur in zwei der untersuchten Reifegradmodelle. Gerade externe Organisationen, deren Einfluss durch Regeln, Empfehlungen und Förderrichtlinien wie oben beschrieben nicht zu vernachlässigen ist, werden in den Reifegradmodellen nur ungenügend berücksichtigt. Auch die Datensicherheit nimmt nur in zwei Modellen einen eigenen Prozessbereich ein. Die Kommunikation, Nachverwendung und Visualisierung der Daten haben im Vergleich zu den Prozessbereichen einen geringen Stellenwert. Das Training nimmt nur im JISC-Modell einen eigenen Prozessbereich ein.

Tabelle 3. Vergleich der Prozessbereiche der Reifegradmodelle.

	DMM	DCAM	CMM-RDM	DGCMM	Mike2.0-IMM	DMPM	DSMM	RDMCMG	JISCMRD
Data Strategy (Strategie und strategische Planung)	✓	✓		✓	✓	✓		✓	✓
Data Governance (Governance, Management, Stewardship, Arbeitsanweisungen, Anforderungsmanagement, Integrität)	✓	✓	✓	✓	✓	✓	✓	✓	
Data Infrastructure (Technologie, Architektur, Infrastructur, Speicherung & Preservation)	✓	✓	✓	✓	✓		✓	✓	✓
Data Quality Management (Analyse, Monitoring & Control, Qualitätskontrolle, Audits)	✓	✓		✓	✓	✓	✓		
Data Risk Management (Risikomanagement & Konformität)	✓			✓	✓				✓
Data Curation (Kuration, Akquisition & Integration)	✓		✓	✓		✓			
Data Communication (Daten-Repräsentation, Usability, Verbreitung & Publikation, Interne und Externe Kommunikation)			✓				✓		
Data Security (Sicherheit, Datenschutz & Zugang)				✓			✓		
Data Stakeholder Management (Stakeholdermanagement und -pflege, Finanzierung der Datenverwaltung)	✓								✓
Organisation (Struktur der internen Organisation, Unterstützungsprozesse und -Services)				✓	✓	✓		✓	
Training (Mitarbeitertraining, Workshops, etc.)									✓

Reifegradmodell für integriertes FDM an multidisziplinären Forschungsorganisationen

Ziel des hier beschriebenen Forschungsvorhabens ist es, ein angepasstes Reifegradmodell für integriertes Forschungsdatenmanagement (int.FDM) zu entwickeln. Dieses Reifegradmodell soll in multidisziplinären Forschungsorganisationen eingesetzt werden, um in Abhängigkeit vom Ent-

wicklungsstand des FDMs Handlungsmöglichkeiten aufzuzeigen. Die gewonnenen Erkenntnisse sollen als Grundlage für die Entwicklung einer nachhaltigen FDM-Strategie verwendet werden. Darüber hinaus sollen konkrete Problembereiche sichtbar werden. Das langfristige Ziel ist eine auf der Ausgangsanalyse aufbauende kontinuierliche Erfassung des Entwicklungsstands zu etablieren und so zur kontinuierlichen Verbesserung und modellbasierten Steuerung der FDM-Strategieimplementierung beizutragen. Als paradigmatische Struktur für multidisziplinäre Forschungsorganisationen dienen Forschungsprozesse am Museum für Naturkunde (MfN).

Das Museum für Naturkunde, Berlin

Das Museum für Naturkunde (MfN) ist ein integriertes Forschungsmuseum der Leibniz Gemeinschaft. Es stellt eine bedeutende Forschungsinfrastruktur in der internationalen und nationalen Forschungslandschaft dar. Die Forschung am MfN ist in vier Forschungsbereichen organisiert, die sich schwerpunktmäßig spezifischen Forschungsleitthemen widmen, wie z.B. der Genomik der Artbildung, der Evolutionären Morphologie, der Biodiversitätsdynamik und der Impakt- und Meteoritenforschung, aber auch der wissenschaftlichen Sammlung als globale Forschungsinfrastruktur, der Integrativen Biodiversitätsentdeckung, der Dynamischen Informations- und Wissensintegration sowie dem Wissens- und Technologietransfer[5]. Der Forschungsbereich Digitale Welt und Informationswissenschaft unterstützt den Prozess des FDMs im MfN auf allen Ebenen. Die wissenschaftlichen Sammlungen mit mehr als 30 Millionen biologischen, paläontologischen und mineralogischen Objekten werden durch die umfangreichen Bestände der Bibliothek, des Tierstimmenarchivs und der historischen Arbeitsstelle ergänzt. Die digitale Erschließung dieser Bestände, sowie deren Bereitstellung und Verfügbarmachung in Forschungsdatenbanken, z.B. Europeana[6], GeoCASe[7], Global Biodiversity Information Facility (GBIF)[8] und GFBio[9], ist eine zentrale Aufgabe des MfN. Darüber hinaus betreibt das MfN eine Vielzahl hochmoderner Laboratorien und IT-Forschungsinfrastrukturen für die Beforschung geo- und biowissenschaftlicher Fragestellungen, z.B. das MicroCT-Labor[10] und das Rechenlabor für Impaktmodellierung und Genomik.

Forschungsdaten entstehen am MfN bei der Forschung an Objekten. Daher können zwei Formen von Daten unterschieden werden: sammlungsbezogene Forschungsdaten und genuine Forschungsdaten, z.B. Messdaten, Beobachtungsdaten und Modellierungsdaten. Sammlungsbezogene Forschungsdaten sind „Metadaten“ der physischen Objekte, aber auch digitale Medien und digitale Sammlungen (z.B. Volumendaten aus der Computertomographie). Diese sammlungsbezogenen Forschungsdaten werden kontinuierlich erweitert und unterliegen einer strengen Qualitätskontrolle. Genuine Forschungsdaten entstehen ebenfalls bei der (sammlungsbezogenen) wissenschaftlichen Arbeit und können durchaus auch ohne direkten Bezug zur Infrastruktur der Objektsammlung betrachtet werden. Das Management dieser Daten unterliegt in der Regel den jeweiligen Vorgaben und Bestimmungen der Projektträger sowie der guten wissenschaftlichen Praxis im Umgang mit Forschungsdaten. Der Übergang zwischen Sammlungs- und Forschungsdaten ist fließend.

5 Vgl. https://www.naturkundemuseum.berlin/de/einblicke/forschung.
6 Vgl. http://www.europeana.eu/portal/de.
7 Vgl. http://www.geocase.eu/.
8 Vgl. http://www.gbif.org/.
9 Vgl. https://www.gfbio.org/.
10 Vgl. https://muellerlaboratory.wordpress.com/ct-facility/.

Int.FDM

Das int.FDM Reifegradmodell soll (1) eine strukturelle Analyse der Entstehung und Verwertung der Daten im Forschungsprozess erlauben, (2) ein flexibles, anpassbares und verbindliches Werkzeug zur Vermittlung der modellierten Strukturen aller am Forschungsprozess Beteiligten darstellen, (3) die Ableitung neuer Handlungsoptionen und Entwicklungsmöglichkeiten anhand ihres Fähigkeitsgrades erlauben und (4) die Identifizierung notwendiger fachlicher Ressourcen und Sachmittel ermöglichen. Dabei sollen bestehende FDM-Praktiken in multidisziplinären Forschungsorganisationen abhängig von der Governance-Ebene, z.B. Förderer, Projekt, Forscher, und unter Berücksichtigung des Forschungsdatenlebenslaufes erhoben und bewertet werden. Eine organisationsweite Anwendung soll durch die angestrebte Integration mit bestehenden FDM-Werkzeugen, z.B. RDMO (Research Data Management Organizer)[11], ermöglicht werden.

Ausblick

Die nächsten Schritte sind eine Bewertung der bestehenden Reifegradmodelle bezüglich ihrer Fähigkeit Governance- und Stakeholderebenen abzubilden, eine Eignungsanalyse der bestehenden Modelle zur Bestimmung des Reifegrads des FDM in einer multidisziplinären Organisation und das Testen des entwickelten Modells an konkreten Anwendungsfällen.

Dazu soll wie im Folgenden beschrieben vorgegangen werden. Das int.FDM Reifegradmodell wird im Vergleich zu bestehenden Reifegradmodellen um Governance-Ebenen für die Prozessbereiche und Leistungsindikatoren in Abhängigkeit von den Prozessbereichen und Fähigkeitsgraden erweitert. Durch die Entwicklung von Ansätzen zur Modellkalibrierung wird sichergestellt, dass es in unterschiedlichen Kontexten angewendet werden kann. Eine nutzerzentrierte Modellvalidierung soll gewährleisten, dass im theoretischen abgeleiteten Modell die Bedürfnisse der Fachexperten und der Anwender berücksichtigt werden. Dazu werden vorab Kriterien entwickelt, mit Hilfe derer die Erkenntnisse aus der Experten- und Anwenderperspektive bewertet werden können. Neben der Validierung ist die Verifizierung ein zentraler Prozessschritt bei der nachhaltigen Entwicklung eines Reifegradmodells für integriertes Forschungsdatenmanagement. Dafür werden Pilotprojekte am MfN ausgewählt, die als repräsentativ für multidisziplinäre Forschungsorganisationen gelten können. In einem ersten Schritt werden die für die Modellkalibrierung erforderlichen Stakeholder in den Projekten identifiziert und eine Kalibrierung durchgeführt. Anschließend erfolgt die Anwendung der jeweiligen Reifegradmodelle in Pilotprojekten. Der Vergleich der Ergebnisse der Modellanwendung werden genutzt, um die strategischen Handlungsfelder zu bestimmen, Managementoptionen abzuleiten und das Modell letztendlich weiter zu entwickeln.

11 Vgl. https://rdmorganiser.github.io/.

Literaturangaben

Aiken, P. H., M. D.Allen, B. Parker, und A Mattia. 2007. „Measuring Data Management Practice Maturity: A Community's Self-Assessment". *IEEE Computer* 40: 4, 42-50.

Allianziniative. 2010. „Grundsätze zum Umgang mit Forschungsdaten". Allianz der deutschen Wissenschaftsorganisationen. Online verfügbar unter: http://www.allianzinitiative.de/de/ zuletzt geprüft am 10.08.2017.

ANDS - Australian National Data Service. 2017. "Creating a data management framework". Online verfügbar unter http://www.ands.org.au/__data/assets/pdf_file/0005/737276/Creating-a-data-management-framework.pdf. zuletzt geprüft am 10.08.2017

Becker, J., R. Knackstedt, und J. Pöppelbuß 2009. „Dokumentationsqualität von Reifegradmodellentwicklungen". Arbeitsbericht Nr. 123, Westfälische Wilhelms-Universität Münster, Institut für Wirtschaftsinformatik, Münster.

BMBF (o.J.): Richtlinien für Zuwendungsanträge auf Ausgabenbasis - AZA. Vordruck Nr. 0027. Bundesministerium für Bildung und Forschung (Hrsg.).

Burger, M., M. Kindling , L. Liebenau, C. Lienhard, S. Lilienthal, P. Plewka, S. Pohlkamp, K. Reinhardt, M. Rügenhagen, K. Schulz, E. Simukovic, K. Sticht und M. Walther. 2013. *Forschungsdatenmanagement an Hochschulen: Internationaler Überblick und Aspekte eines Konzepts für die Humboldt-Universität zu Berlin*. Humboldt-Universität zu Berlin. urn:nbn:de:kobv:11-100210226.

CMMI Institute. 2014. *Data Management Maturity Model*. Version 1.0. CMMI Institute.

Crawford, J. K. 2007. *Project Management Maturity Model*. 2. Edition. Boca Raton, FL: Auerbach/CRC Press.

DFG. 2013. Sicherung guter wissenschaftlicher Praxis. Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft". Deutsche Forschungsgemeinschaft, Wiley, 1. Auflage.

DFG. 2015. Leitlinien zum Umgang mit Forschungsdaten. Deutsche Forschungsgemeinschaft, Deutsche Forschungsgemeinschaft (Hrsg.). Online verfügbar unter http://www.dfg.de/foerderung/antragstellung_begutachtung_entscheidung/antragstellende/antragstellung/nachnutzung_forschungsdaten/

EC. 2016. H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020. Europäische Commission, Directorate-General for Research & Innovation (Hrsg.). Version 3.0, 26 July 2016.

Enke, H., N. Fiedle, T. Fischer, T. Gnadt, E. Ketzan, J. Ludwig,, T. Rathmann, G. Stöckle. 2011. Checkliste zum Forschungsdaten-Management. Version 0.6. Entwurfsversion zur öffentli-

chen Kommentierung. WissGrid (Hrsg.), Online verfügbar unter: http://www.wissgrid.de/publikationen/deliverables/wp3/WissGrid-oeffentlicher-Entwurf-Checkliste-Forschungsdaten-Management.pdf.

FORCE11 (Hrsg.) (o.J.): The Fair Data Principles - For Comment. FORCE11-Webseite. Online verfügbar unter: https://www.force11.org/group/fairgroup/fairprinciples.

Fowler, S. 2012. JISC Managing Research Data Project Maturity Model. Online verfügbar unter http://www2.uwe.ac.uk/services/library/using_the_library/Services%20for%20researchers/maturity-model-v.1.1.pdf.

Hey, T., S. Tansle, K. M. Tolle et al..2009. *The fourth paradigm: data-intensive scientific discovery*. 1. Auflage. Redmond, WA, Vereinigte Staaten: Microsoft Research.

Higgins, S.. 2008. *The DCC Curation Lifecycle Model. Int. Journal of Digital Curation* 3: 1, 134-140.

HRK. 2015. *Wie Hochschulleitungen die Entwicklung des Forschungsdatenmanagements steuern können: Orientierungspfade, Handlungsoptionen, Szenarien*. 19. Mitgliederversammlung der Hochschulrektorenkonferenz. 10. November 2015, Kiel.

HU Berlin. 2014a. Grundsätze zum Umgang mit Forschungsdaten an der Humboldt-Universität zu Berlin (Forschungsdaten-Policy). Humboldt-Universität zu Berlin (Hrsg.).

HU Berlin. 2014.b. Handlungsempfehlungen in Ergänzung zu den Grundsätzen zum Umgang mit Forschungsdaten an der Humboldt-Universität zu Berlin. Humboldt-Universität zu Berlin (Hrsg.).

HU Berlin. 2014c. Anleitung zur Erstellung eines Datenmanagementplans (DMP) in Horizon 2020. Version 0.6. Humboldt-Universität zu Berlin (Hrsg.).

IBM. 2007. The IBM Data Governance Council Maturity Model: Building a roadmap for effective data governance. IBM Software Group (Hrsg.), Somers, New York.

ITIL.2013. ITIL Maturity Model and Self-assessment Service. User Guide. Axelos Ltd., Oktober 2013.

Kindling, M. und P. Schirmbacher. 2013. „Die digitale Forschungswelt als Gegenstand der Forschung“. *Information - Wissenschaft & Praxis* 64: 2-3, 137-148.

Kindling, M., P. Schirmbacher, und E. Simukovic. 2013. „Forschungsdatenmanagement an Hochschulen: das Beispiel der Humboldt-Universität zu Berlin“. *LIBREAS*. Library Ideas, Nr. 23.

Ludwig, J. und H. Enke (Hrsg.). 2013. Leitfaden zum Forschungsdaten-Management. Handreichungen aus dem WissGrid-Projekt. Online verfügbar unter https://escience.aip.de/ak-forschungsdaten/wp-content/uploads/2013/09/Leitfaden_Data-Management-WissGrid.pdf.

Mettler, T. 2010. „Supply Management im Krankenhaus: Konstruktion und Evaluation eines konfigurierbaren Reifegradmodells zur zielgerichteten Gestaltung“. Dissertation Nr. 3752, Universität Sankt Gallen, Sierke Verlag, Göttingen.

OMG. 2008. *Business Process Maturity Model* (BPMM). Version 1.0. Object Management Group, Inc. (Hrsg.). Juni 2008.

Open Science Policy Platform. 2016. European Open Science Policy Platform. Online verfügbar unter http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-policy-platform.

Paulk, M. C., B. Curtis, M. B: Chrissis, und C. V. Weber. 1993. *Capability Maturity Model for Software*. Version 1.1. Technischer Report. Software Engineering Institute, Carnegie Mellon Universität, Pittsburgh, Pennsylvania, Vereinigte Staaten, Februar 1993.

Peng, G., J. L.Privette, E. J. Kearns, N. A., Ritchey, S. Ansari. 2015. “A unified framework for measuring stewardship practices applied to digital environmental datasets”. *Data Science Journal* 13, 231-253.

Qin, J., K. Crowston, C. Flynn, und A. Kirkland. 2014. *A Capability Maturity Model for Research Data Management. Projektreport,* Version 1.0, Syracuse University, Vereinigte Staaten.

Recht, G. 2016. *Bundesdatenschutzgesetz (BDSG)*. Bundesministeriums der Justiz und für Verbraucherschutz (Hrsg.), 2. Auflage.

RfII. 2016. *Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland*. Rat für Informationsinfrastrukturen (Hrsg.), Göttingen.

RTD (Hrsg.). 2016. Draft European Open Science Agenda. Directorate-General for Research and Innovation (RTD). Online verfügbar unter: http://ec.europa.eu/research/openscience/pdf/draft_european_open_science_agenda.pdf.

SEI. 2011. Standard CMMI Appraisal Method for Process Improvement (SCAMPISM) A, Version 1.3: Method Definition Document. Software Engineering Institute (Hrsg.), März 2011. Online verfügbar unter: http://www.sei.cmu.edu/reports/11hb001.pdf

TU Darmstadt. 2015. Leitlinien zum Umgang mit digitalen Forschungsdaten an der TU Darmstadt. Stand: 16.12.2015.

Replikationen in den Wirtschaftswissenschaften - Stand und Perspektiven

Ralf Toepfer[1]

1 ZBW Leibniz-Informationszentrum Wirtschaft

Zusammenfassung. Zahlreiche Studien legen den Schluss nahe, dass auf empirischen Daten beruhende wirtschaftswissenschaftliche Forschungsergebnisse oftmals nicht reproduziert werden können. Die Notwendigkeit, die Reproduzierbarkeit empirischer Forschung zu erhöhen sowie Publikationsmöglichkeiten für Replikationsstudien zu schaffen, wird denn auch zunehmend gesehen. Ausgehend von einem Literaturüberblick über die Reproduzierbarkeit empirisch basierter, wirtschaftswissenschaftlicher Forschungsergebnisse sowie den Gründen mangelnder Reproduzierbarkeit (Stand), werden aktuelle Initiativen vorgestellt, die darauf abzielen Replikationen in den Wirtschaftswissenschaften zu erhöhen (Perspektiven).

Schlagwörter. Replikationen, Reproduzierbarkeit, Forschungsdaten, Wirtschaftswissenschaften, Fachzeitschriften

Einleitung

Ergebnisse empirischer Forschung müssen transparent dokumentiert, nachprüfbar und wiederholbar sein. Gerade die Robustheit von empirischen Befunden ist Grundlage für eine glaubwürdige Wissenschaft und adäquate Politikberatung. Es ist insofern kein gutes Signal, wenn etwa 70% der befragten Forschenden in einer Umfrage der Zeitschrift Nature angeben, dass sie es nicht geschafft hätten, die Experimente anderer Forscherinnen und Forscher zu reproduzieren und sogar mehr als die Hälfte der befragten in derselben Umfrage nicht einmal ihre eigenen Experimente reproduzieren konnten (Baker 2016). Vor dem Hintergrund dieser Zahlen verwundert es nicht, dass 90% der Forschenden eine „signifikante“ bzw. „leichte“ Replikationkrise ausmachen (Baker 2016, 452). Diese Replikationskrise hat alle empirisch arbeitenden Wissenschaftsdisziplinen erfasst; von der Medizin (Ioannidis 2005), über die Psychologie (Simmons, Nelson und Simonsohn 2011), die Soziologie (Gerber und Malhotra 2008), die Politikwissenschaft (Gerber, Green und Nickerson 2001) bis hin zur Betriebs- (Hubbard und Armstrong 1994) und Volkswirtschaftslehre (Chang und Li 2015).

In der Vergangenheit wurden Replikationen in den Wirtschaftswissenschaften eher kritisch beäugt und dementsprechend auch nur wenige Replikationsstudien publiziert. Ein wesentlicher Aspekt für den Mangel an publizierten Replikationsstudien ist, dass die Veröffentlichung solcher Studien mehr Risiken denn Chancen für die Reputation der Forschenden bedeutet und zwar sowohl für die Autorinnen bzw. Autoren der Originalstudie als auch für die Replizierenden (Park 2004, zitiert nach Fecher, Fäßdorf und Wagner 2016, 4). Vor dem Hintergrund der Replikationskrise und des damit einhergehenden Vertrauensverlustes in die Ergebnisse empirisch basierter

Forschung, scheint jedoch langsam ein Bewusstseinswandel einzutreten. Die Notwendigkeit, die Reproduzierbarkeit empirischer Forschung zu erhöhen sowie Publikationsmöglichkeiten für Replikationsstudien zu schaffen, wird zunehmend gesehen. Dies spiegelt sich in verschiedenen Initiativen wider, die in diesem Beitrag kurz dargestellt und diskutiert werden.

Der vorliegende Text gliedert sich in zwei größere Abschnitte. Ausgehend von einem Literaturüberblick über die Reproduzierbarkeit empirisch basierter, wirtschaftswissenschaftlicher Forschungsergebnisse sowie den Gründen mangelnder Reproduzierbarkeit wird in Abschnitt 2 zunächst der Stand von Replikationen in den Wirtschaftswissenschaften dargestellt. Darauf aufbauend werden in Abschnitt 3 aktuelle Initiativen vorgestellt, die darauf abzielen Replikationen in den Wirtschaftswissenschaften zu erhöhen. Im Einzelnen werden „The Replication Network“ (TRN), das Göttinger Replication Wiki, die Replication Section des E-Journal "economics" sowie ein Vorschlag zur Gründung eines reinen Replication-Journals vorgestellt. Darüber hinaus wird das von der DFG geförderte Projekt "International Journal of Economic Micro Data (IJEMD) - eine neuartige Informationsinfrastruktur zur Publikation von begutachteten Forschungsdaten in den Wirtschaftswissenschaften - präsentiert. In Abschnitt 4 wird ein kurzes Fazit gezogen.

Replikationen in den Wirtschaftswissenschaften (Stand)

Das Thema Replikationen ist für die Wirtschaftswissenschaften alles andere als neu. Bereits 1933 wies der damalige Herausgeber der Fachzeitschrift „Econometrics“ Ragnar Frisch in einem Editorial darauf hin, dass die Originaldaten zu den publizierten empirischen Arbeiten in der Regel mitpubliziert werden. Kane (1984) wies darüber hinaus auf die Bedeutung von reproduzierbarer Forschung hin:

> „Tedious though its requirements may be, reproducibility remains the touchstone of the scientific method. If an empirical finding is a fact, other researchers should be able to observe it, too. Successful and independent repetition of an econometric experiment increases professional confidence in the experiment’s alleged results.” (Kane 1984, 4)

Unbestritten in den Wirtschaftswissenschaften - und nicht nur dort - ist, dass Forschungsergebnisse reproduzierbar sein müssen und Replikationen einen Eckpfeiler des wissenschaftlichen Arbeitens bilden (McCullogh 2009, 118-119). Vor diesem Hintergrund stellen Untersuchungen, wie die von Dewald, Thursby und Anderson (1986), McCullough, McGeary und Harrison (2006), McCullough, McGeary und Harrison (2008) sowie Chang und Li (2015), die zeigen, dass die Ergebnisse empirischer Wirtschaftsforschung in der Regel nicht replizierbar sind, die Glaubwürdigkeit von und das Vertrauen in wirtschaftswissenschaftliche Forschung grundsätzlich in Frage. Es würde an dieser Stelle zu weit führen alle Untersuchungen, die die Replizierbarkeit wirtschaftswissenschaftlicher Forschung in Frage stellen ausführlich darzustellen. Exemplarisch wird deshalb hier nur auf eine aktuelle Untersuchung von Chang und Li (2015) kurz eingegangen. In ihrem Paper mit dem bezeichnenden Titel: „Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say ‚Usually Not‘“ legen Chang und Li dar, wie sie versucht haben die Ergebnisse von 67 Forschungsartikeln aus 13 angesehenen wirtschaftswissenschaftlichen Fachzeitschriften mit Hilfe der von den Autorinnen bzw. Autoren der Originalstudien bereitgestellten Forschungsdaten und Code-Files zu replizieren. Im Ergebnis konnten Chang und Li 22

von 67 (33%) Arbeiten ohne Hilfe der Autorinnen bzw. Autoren, also nur unter zu Hilfenahme der bereit gestellten Dokumentation (read-me-files), erfolgreich replizieren. Mit Hilfe der Autoren war es ihnen möglich 29 von 59 (49%) Artikeln zu replizieren. Chang und Li ziehen aus ihrer Untersuchung die Schlussfolgerung:

> „Because we successfully replicate less than half of the papers in our sample even with assistance from the authors, we conclude that economics research is usually not replicable." (Chang und Li 2015, 3)

Die Bestandsaufnahme der Replizierbarkeit wirtschaftswissenschaftlicher Forschung fällt somit ernüchternd aus und es stellt sich die Frage nach den Gründen für die mangelnde Replizierbarkeit. Um sich dieser Frage zu nähern, ist es unerlässlich zunächst die Frage zu klären, was Replikation in den Wirtschaftswissenschaften eigentlich meint. Der Terminus „Replikation" wird in den Wirtschaftswissenschaften nämlich nicht einheitlich definiert (Duvendack, Palmer-Jones, Reed 2016 sowie Clemens 2015). Hamermesh (2007) differenziert z. B. zwischen „pure replication", „statistical replication" und „scientific replication", wobei er unter „pure replication" das schlichte Reproduzieren einer Untersuchung mit denselben Daten und derselben Methode wie die in der Originalstudie verwendeten meint. „Statistical replication" bezieht sich auf die Reproduktion einer Untersuchung mit derselben Grundgesamtheit aber einer anderen Stichprobenwahl und unter „scientific replication" fasst Hamermesh die Replikation einer Untersuchung mit einer anderen Stichprobe, einer anderen Grundgesamtheit und einer möglicherweise ähnlichen, aber keineswegs derselben Methode (Hamermesh 2007, 1). In den oben genannten Untersuchungen zur Replizierbarkeit wirtschaftswissenschaftlicher Forschung wird in der Regel die Definition der „pure replication" zu Grunde gelegt. Clemens (2015) legt ebenfalls dar, dass kein Konsens bezüglich der Definition des Terminus „Replikation" herrscht und das dies einer der Gründe für die Nicht-Replizierbarkeit wirtschaftswissenschaftlicher Forschung ist.

> „Thus if a replication test gives discrepant results, under current usage of the term, this could mean a wide spectrum of things - from signaling a legitimate disagreement over the best method (science), to signaling incompetence and fraud (pseudoscience)." (Clemens 2015, 1)

Abseits definitorischer Fragen gibt es eine Reihe weiterer Gründe dafür, warum Arbeiten nicht repliziert werden können (Vlaeminck et al. 2013). Legt man die Definition von „pure replication" zugrunde, so ist ein offensichtlicher Grund für die mangelnde Replizierbarkeit, dass die Originaldaten nicht zur Verfügung stehen. Wie Andreoli-Versbach und Mueller-Langer (2014) zeigen ist das öffentliche Teilen von Forschungsdaten in den Wirtschaftswissenschaften eher die Ausnahme, denn die Regel. Von 488 empirisch arbeitenden Wirtschaftsforschenden in der Stichprobe von Andreoli-Versbach und Mueller-Langer teilen 16,8% sporadisch ihre Forschungsdaten und nur 2,46% teilen ihre Daten regelmäßig öffentlich. Die überwiegende Mehrheit, 394 Ökonominnen bzw. Ökonomen (80,74%), teilen ihre Forschungsdaten nicht (Andreoli-Versbach und Mueller-Langer 2014, 1621). Selbst wenn die der Veröffentlichung zugrunde liegenden Forschungsdaten zur Verfügung gestellt werden, ist damit aber keineswegs gesichert, dass die Forschungsergebnisse reproduziert werden können. Neben der Bereitstellung der Daten ist es erforderlich dem Replizierenden, die Berechnungsschritte, den Code sowie weitere Erläuterungen an die Hand zu geben (Colaresi 2016, 2 sowie McCullogh, McGeary and Harrison 2008). Genau diese Problematik adressieren informationstechnische Systeme wie Dataverse oder das ZBW

Journal Data Archive, in dem diese alle für eine Replikation erforderlichen Informationen zu einem Journalartikel in einem Repository zur Verfügung stellen.

Neben diesen eher „technischen“ Problemen der Reproduzierbarkeit von Forschungsergebnissen, spielen vor allem „kulturelle“ Aspekte eine Rolle. Mit der Replikation der Forschungsergebnisse Dritter kann keine Reputation gewonnen werden. Im Gegenteil, den Replizierenden wird eher ein Mangel an Kreativität vorgeworfen und Replikationen als Zeitverschwendung betrachtet. Darüber hinaus könnten Replikationen als mangelndes Vertrauen in die wissenschaftliche Integrität der Autorinnen bzw. Autoren der Originalstudie oder als persönlicher Disput interpretiert werden (Dewald, Thursby und Anderson 1986, 587). Kane (1984) bringt diese Aspekte auf den Punkt:

> “…uninventively verifying someone else’s empirical research is not a completely respectable use of one’s time. Choosing such a task is widely regarded as *prima facie* evidence of intellectual mediocrity, revealing a lack of creativity and perhaps even bullying spirit.” (Kane 1984, 3)

Ein ambivalentes Verhältnis zu Replikationen machen auch Fecher, Fräßdorf und Wagner (2016) in einer Umfrage unter Forschenden, die Daten des Sozioökonomischen Panels (SOEP) genutzt haben, aus. So stimmten 84% der Befragten der Aussage zu, dass Replikationen notwendig für die Verbesserung des wissenschaftlichen Fortschritts sind und 71% verneinten, dass Replikationen überflüssig seien. Andererseits hatten 58% der Befragten noch nie eine Replikation für eine Untersuchung durchgeführt, die auf SOEP-Daten beruht (Fecher, Fräßdorf und Wagner 2016, 7-8).

Darüber hinaus kommt erschwerend hinzu, dass Replikationsstudien kaum eine Chance haben in den wirtschaftswissenschaftlichen Fachzeitschriften publiziert zu werden. Dies liegt vor allem daran, dass die Journals vor allem an neuen Ergebnissen interessiert sind, da diese sich besser vermarkten lassen (Ulrich et al. 2016, 164). Van Witteloostuijn (2006) spricht passend von einem „novelty bias“ und weist zu Recht darauf hin, dass der obsessive Fokus auf neue, bahnbrechende Erkenntnisse, nicht der alleinige Motor für wissenschaftlichen Fortschritt sein sollte, da diese naturgemäß äußerst selten sind. „Much scholarly works involves transpiration, and not so much inspiration.” (van Witteloostuijn 2006, 485).

Vor diesem kurz skizzierten Hintergrund überrascht es wenig, dass Duvendack, Palmer-Jones und Reed (2015) in einer systematischen Auswertung im Zeitraum von 1977 bis 2014 nur insgesamt 162 veröffentlichte Replikationsstudien in den Top 50 volkswirtschaftlichen Fachzeitschriften ausfindig machen. Etwa 20% dieser Replikationsstudien wurden allein vom Journal of Applied Econometrics publiziert, sechs wirtschaftswissenschaftliche Fachzeitschriften zeichnen sich für etwa 60% aller veröffentlichten Replikationsstudien aus und nur 10 Journals haben jemals mehr als drei Replikationsstudien publiziert (Duvendack, Pamer-Jones und Reed 2015, 177). Trotz der zunehmenden Anzahl an wirtschaftswissenschaftlichen Fachzeitschriften mit einer Data Availibilty Policy (Vlaeminck und Herrmann 2015) ist die Anzahl der veröffentlichten Replikationsstudien nur marginal angestiegen (Duvendack, Palmer-Jones und Reed 2015).

Initiativen zur Erhöhung der Reproduzierbarkeit (Perspektiven)

Bedingt durch den Wandel in der Wissenschaft hin zu mehr Offenheit und Transparenz, wie er sich u.a. in den wissenschaftspolitisch motivierten Aktivitäten in Richtung "Open Science", z.B. der Europäischen Kommission abzeichnet, ist das Interesse an der Reproduzierbarkeit empirischer Wirtschaftsforschung sowie Publikationsmöglichkeiten für Replikationsstudien neu erwacht. Dies spiegelt sich in verschiedenen aktuellen Initiativen wider, die im Folgenden kurz dargestellt werden. Dabei kann grob zwischen folgenden Kategorien differenziert werden:

- Allgemeine Aufwertung des Themas durch Community Building
- Aufbau von alternativen Publikationsplattformen für Replikationen
- Verstärkte Veröffentlichung von Replikationsstudien bei bestehenden Fachzeitschriften
- Aufbau von dezidierten Replication Journals

Eine Initiative, die dem Community Building dient, ist „The Replication Network“ (TRN). Es spiegelt in gewisser Weise das neu entfachte Interesse der Wirtschaftswissenschaften am Thema Replikationen wider, versteht es sich doch als Kanal, um Aktualisierungen über den Stand der Replikationen in den Wirtschaftswissenschaften zu kommunizieren und darüber hinaus ein Netzwerk zum Teilen von Information und Ideen zu bilden. Ziel des TRN ist es, Wirtschaftsforschenden und Fachzeitschriften zu ermutigen Replikationsstudien zu publizieren. Das Netzwerk zählt aktuell immerhin 358 Mitglieder. In einem Blog berichten Gastwissenschaftlerinnen bzw. Gastwissenschaftler über ihre Erfahrungen und Publikationen im Kontext von Replikationen und die Kommentarfunktion des Blogs ermöglicht den Austausch und die Diskussion.

Einen Weg abseits der traditionellen Publikationspfade verfolgt das Göttinger Replikation-Wiki, dass u.a. auch in TRN integriert ist. Das Replication-Wiki umfasst zurzeit 330 Replikationen (Stand: 20.01.2017) und enthält Informationen zu Daten, Methoden, Software sowie weiterer Materialien zu mehr als 2.000 empirischen Studien, die somit zur Replikation zur Verfügung stehen. Im Mai 2016 zählte das Replication-Wiki mehr als 100 registrierte Nutzer und über 850.000 Seitenaufrufe (Höffler 2016). So interessant der Ansatz eine Replication-Wikis auch ist, so wenig eignet er sich allerdings für den Reputationsaufbau der Replizierenden, da die „Währung“ der Ökonominnen und Ökonomen Zitationen und Veröffentlichungen in möglichst hoch-gerankten Fachzeitschriften ist.

In der wirtschaftswissenschaftlichen Forschungscommunity werden deshalb vor allem Publikationsmöglichkeiten für Replikationen diskutiert, die sich in das bestehende „Wertesystem“ einpassen. So plädiert etwa Harzing (2016) für die Einführung von Rubriken für Replikationen in bestehenden Journalen mit dem Argument, das Replikationsstudien in genau dem richtigen „Field-Journal“ veröffentlicht werden könnten und damit genau das richtige Zielpublikum erreicht würde (Harzing 2016, 566). Auch Coffman, Niederle und Wilson (2017) machen sich für eine „replication section“ in bestehenden Fachzeitschriften stark und regen an, dass die „Top-Journals“ kurze („one-page“) „replication reports“ publizieren sollten. So plausibel diese Vorschläge auch sind, darf dabei nicht übersehen werden, dass alle Versuche der Einführung solcher Rubriken in der Vergangenheit erfolglos geblieben sind. Von 1977 bis 1999 bot beispielsweise das „Journal of Political Economics“ die Möglichkeit u.a. Replikationsstudien in der Rubrik „Confirmations and Contradictions“ zu veröffentlichen und die Zeitschrift „Labour Economics“ lud in ihrer Erstausgabe 1993 explizit dazu ein Replikationsstudien zu veröffentlichen. Nach einigen Jahren jedoch zog man dieses Angebot aufgrund der mangelnden Einreichungen zurück: “As the then-editor wrote, there was a, ‘lack of interest: we simply got no submissions. There is a structural lack of

interest in replication." (Hamermesh 2007, 9). Gleichwohl führte das E-Journal "Economics" 2016 nach intensiven Diskussionen unter den Herausgeberinnen bzw. Herausgebern eine "replication section" ein. Begründet wird die Einführung dieser section beim E-Journal „Economics" wie folgt:

> "Replications are an important public good to the economics community, and as such, they tend to be under-valued. By opening a replications section in our open access, open assessment journal, we hope to make it easier for authors to provide this important public good." (http://www.economics-ejournal.org/special-areas/replications)

Trotz der relativ hohen Anforderungen, die das E-Journal "Economics" an die Veröffentlichung von Replikationsstudien stellt, wurden seit 2016 zwei solche Arbeiten veröffentlicht (Stand: 20.01.2017).

Ausgehend von den negativen Erfahrungen bei der Etablierung von Replikations-Rubriken in bestehenden ökonomischen Fachzeitschriften in der Vergangenheit, plädieren z.B. Zimmermann (2015), aber auch van Witteloostuijn (2016, 486), für die Gründung von Fachzeitschriften, die ausschließlich Replikationen publizieren. Zimmermann (2015) argumentiert:

> "Existing journals could be more willing to accept replication studies, but it is clear reputational inertia is not providing the right incentives. Hence, I suggest the creation of a journal entirely dedicated to replication." (Zimmermann 2015, 3)

Genau an dieser Stelle der gegebenen, trägen Reputationsmechanismen, setzt das von Deutschen Forschungsgemeinschaft geförderte Projekt "International Journal of Economic Micro Data (IJEMD) - Ein neuartige Informationsinfrastruktur zur Publikation von begutachteten Forschungsdaten in den Wirtschaftswissenschaften" an. Ziel des Vorhabens ist es, eine offene Webplattform für die Publikation wirtschaftswissenschaftlicher Forschungsdaten zu konzipieren. Die konkrete Umsetzung soll dabei als ein Peer Review Data Journal erfolgen und neben Datenbeschreibungen vor allem auch Replikationsstudien veröffentlichen. Der aktuelle Arbeitstitel des Journals lautet „International Journal for Re-Views in Empirical Economics (IREE)." Mit dem Projekt wird die Absicht verfolgt, durch das Open Access Data Journal eine attraktive Publikationsmöglichkeit für Forschungsdaten und Replikationsstudien zu schaffen, da die Replizierenden in Form von Zitationen „belohnt" werden können, was wiederum im bestehenden „Wertesystem" zu einer Reputationssteigerung führen würde.

Fazit

Eine Bestandsaufnahme des Themas Replikationen in den Wirtschaftswissenschaften fällt ernüchternd aus. Zwar werden Replikationen als wichtig für den Erkenntnisfortschritt anerkannt, gleichzeitig sorgen die vorherrschenden Reputationsmechanismen jedoch dafür, dass für den individuellen Forscher bzw. die individuelle Forscherin keinerlei Anreize bestehen, Replikationen durchzuführen. Bedingt durch den Wandel in der Wissenschaft hin zu mehr Offenheit und Transparenz spricht jedoch einiges dafür, dass Replikationen in der einen oder anderen Form einen festen Platz im wissenschaftlichen Publikationsprozess erobern werden. Die oben genannten Initiativen und Projekte deuten meiner Auffassung nach eine positive Entwicklung an. Auch Christensen und

Miguel (2016) machen ein wachsendes Interesse an Transparenz und Reproduzierbarkeit in den Wirtschaftswissenschaften aus:

> „The rising interest in transparency und reproducibility in economics reflects broader global trends regarding these issues, both among academics and beyond. As such, we argue that 'this time' really may be different than earlier bursts of interest in research transparency within economics (…) that later lost momentum and mostly died down." (Christensen und Miguel 2016, 61)

Ob diesmal wirklich alles anders ist oder ob die oben skizzierten Initiativen, wie vergleichbare Aktivitäten in der Vergangenheit, an Vernachlässigung sterben („…died from neglect" Hamermesh 2007,11), kann zum jetzigen Zeitpunkt nicht abschließend beantwortet werden.

Literaturangaben

Anderson, R. G., und W. G. Dewald. 1994. „Replication and Scientific Standards in Applied Economics a Decade After the Journal of Money, Credit and Banking Project." *Federal Reserve Bank of St. Louis Review*, November/December: 79-83.

Andreoli-Versbach, Partrick, und Frank Mueller-Langer. 2014. „Open Access to data: An ideal professed but not practised." *Reserach Policy*. Online verfügbar unter http://dx.doi.org/10.1016/j.respol.2014.04.008. November 2014: 1621-1633. Zuletzt geprüft am 11.08.2017.

Baker, Monya. 2016. „Is there a Reproducibiliby Crisis?" *Nature*, 26. May: 452-454.

Chang, Andrew C., und Phillip Li. 2015 *Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Usually Not"*.

Christensen, Garret S., und Edward Miguel. 2016. *Transparency, Reproducibility, And The Credibility Or Economics Research.* NBER Working Paper 22989. Online verfügbar unter http://www.nber.org/papers/w22989. Zuletzt geprüft am 11.08.2017.

Clemens, Michael A. 2016. *The Meaning of Failed Replications: A Review and Proposal.* IZA.

Coffman, Lucas, Muriel Niederle, und Alistair J. Wilson. 2017. *Replications: A Proposal to Increase their Visibility and Promote them.*

Dewald, William G., Jerry Thursby, und Anderson Richard G. 1986. "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project." *American Economic Review*: 587-603.

Duvendack, Maren, Richard Palmer-Jones, und Robert W. Reed. 2016. *What is meant by "Replicatin" and Why Does It Encounter Resistance in Economics?* New Zealand: University of Canterbury.

Duvendack, Maren, Richard W. Palmer-Jones, und Robert W. Reed. 2015. "Replications in Economics: A Progress Report." *Econ Journal Watch* 12 (2): 164-191.

Frisch, Ragnar. 1933 "Editorial." *Econometrica*: 1-4.

Gerber, Alan S., Donald P. Green, und David Nickerson. 2001. „Testing for Publication Bias in Political Science." *Political Analysis* 9 (4), 2001: 385-92.

Gerber, Alan, und Neil Malhotra. 2008. „Publication Bias in Empirical Sociological Research Do Arbitrary Significance Levels Distort Published Results?" *Sociological Methods & Research* 37(1: 3-30. doi: 10.1177/0049124108318973

Hamermesh, Daniel S. 2007. *Replication in Economics.* NBER.

Höffler, Jan H. 2016 „Replication and Transparency in Economic Research." *Blog des Institute for New Economics Thinking*, 10. Mai. http://ineteconomics.org/ideas-papers/blog/replication-and-transparency. Zuletzt geprüft am 11.08.2017.

Hubbard, Raymond, und J. Scott Armstrong. 1994. „Replications and Extensions in Marketing: Rarely Published but Quite Contrary." *Int J Res Mark,* 11: 233-248.

Ioannidis, John P.A. 2005. „Why Most Published Research Findings Are False." *PLoS Med* 2 (8):e124. doi:10.1371/journal.pmed.0020124

Kane, Edward J. 1984. "Why Journal Editors Should Encourage The Replication Of Applied Econometric Reserach." *Quarterly Journal of Business and Economics*: 3-8.

Makel, Matthew C., und Jonathon A. Plucker. 2014. „Facts Are More Important Than Novelty: Replication in the Education Sciences." *Educational Researcher*, August / September: 304-316.

McCullogh, B.D. 2009. „Open Access Economics Journals and the Market for Reproducible Economic Research." *Economic Analysis & Policy* 39 (1): 117-126. Online verfügbar unter http://dx.doi.org/10.1016/S0313-5926(09)50047-1. Zuletzt geprüft am 11.08.2017.

McCullough, B.D., K.A. McGeary, und T. Harrison. 2006. „Lessons from the JMCB Archive." *Journal of Money, Credit and Banking*: 1093-1107.

McCullough, B.D., K.A. McGeary, und T.D. Harrison. 2008. „Do Economic Journal Archives Promote Replicable Research?" *Canadian Journal of Economics*: 1406-1420.

Pesaran, H. 2003 „Introducing a Replication Section." *Journal of Applied Econometrics* 18 (1): 111.

Simmons, Joseph P., Leif D. Nelson, und Uri Simonsohn. 2011. „False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359-1366.

Ulrich, Rolf, Edgar Erdfelder, Roland Deutsch, Bernhard Strauß, Anne Brüggemann, Bettina Hannover, Brunna Tuschen-Caffier, Clemens Kirschbaum, Gerhard Blickle, Jens Möller, und Winfried Rief. 2016. „Inflation von falsch-positiven Befunden in der psychologischen Forschung: Mögliche Ursachen und Gegenmaßnahmen." *Psychologische Runschau*, 67 (3): 163-174.

Vlaeminck, Sven, Gert F. Wagner, Joachim Wagner, Dietman Harhoff, und Olaf Siegert. 2013. „Replizierbare Forschung in den Wirtschaftswissenschaften erhöhen." *Libreas. Library Ideas* 23: 29-42.

Vlaeminck, Sven, und Lisa-Kristin Herrmann. 2015 „Data Policies and Data Archives: A New Paradigm for Academic Publishing in Economic Sciences?" In *New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science: Scale, Openness and Trust. Proceedings of the 19th International Conference on Electronic Publishing*, von B. Schmidt und M. Dobreva (Hrsg.).

Witteloostuijn, Arijen. 2016. „What happend to Popperian falsification? Publishing neutral and negative findings: Moving away from biased publication practices." *Cross Cultural & Strategic Management*: 481-508. Online verfügbar unter http://dx.doi.org/10.1108/CCSM-03-2016-0084 Zuletzt geprüft am 11.08.2017

Zimmermann, Christian. 2014. *On the Need for a Replication Journal.* St. Louis.

The GFBio Terminology Service: enabling research data management beyond data heterogeneity

Naouel Karam[1], Robert Harald Lorenz[2], Claudia Müller-Birn[3]

1,2,3 Institute of Computer Science, Freie Universität Berlin

Abstract. A primary goal of a research infrastructure for data management should be to enable efficient data discovery and integration of heterogeneous data. The German Federation for Biological Data (GFBio) was guided by this goal. The basic component, that enables such interoperability and serves as a backbone for such a platform, is the GFBio Terminology Service (GFBio TS). This service acts as a semantic platform for accessing, developing and reasoning about terminological resources within the biological and environmental domains. A RESTful API gives access to these terminological resources in a uniform way, regardless of their degree of complexity and whether they are internally stored or externally accessed through web services. Additionally, a set of widgets with an intrinsic API connection are made available for easy integration in applications and web interfaces. Based on the requirements of GFBios partners, we describe the added value that is provided by the GFBio Terminology Service with practical scenarios as well as the challenges we still face. We conclude by describing our current activities and future developments.

Keywords. Research data infrastructure, Interoperability, Terminology repository, Semantic Web, RESTful API, Widgets.

Introduction

Research practice has become more data-intensive over the last few decades, and this development is visible across many research disciplines. However, the sharing of research data beyond disciplinary borders is still a challenge. Thus, a research infrastructure for data management should allow for an efficient data integration and therefore, the discovery of heterogeneous research data.

The German Federation for Biological Data (GFBio) pursues this goal. GFBio aims at providing a data management platform and data archiving solutions for data capture, annotation, indexing, searching and storage in the area of biological and environmental research. The GFBio Data Portal[1] integrates existing data infrastructures such as PANGAEA[2] into the GFBio Repository Network.

Data generated in biodiversity and ecology research are extremely heterogeneous and pertaining to different scientific disciplines using various methods and technologies. The situation is further complicated by different understandings of employed terms within different scientific do-

1 http://www.gfbio.org
2 www.pangaea.de

mains. Developing interoperability and harmonizing data by using standards and terminological resources are crucial for data mobilization, integration, and discovery in the GFBio context.

The core component that enables this interoperability and serves as a backbone for the GFBio infrastructure is called the GFBio Terminology Service[3] (GFBio TS) (Karam et al. 2016). The GFBio Terminology Service acts as a semantic platform for accessing, developing, and reasoning over terminological resources. The GFBio TS focuses on integrating and giving access to terminologies developed by project partners as well as external terminologies defined and maintained by related communities. These terminologies can range from simple term lists to complex ontologies. Based on the requirements of the GFBio community, the Terminology Service provides access to over 20 terminologies so far, of which GFBios partners have contributed 10 terminologies. A well-defined RESTful API gives access to all terminologies in a uniform way regardless of their degree of complexity and whether they are internally stored or externally accessed through web services. The services provided by the GFBio TS can also be integrated easily within existing web applications with the help of widgets, which are small applications with limited functionality. We developed two exemplary widget prototypes so far: a term visualization and a search widget.

We will explain the advantage of using semantic technologies for data management and highlight the utility of the Terminology Service by practical use cases of semantically enhanced components. More specifically, we will differentiate between four main usage scenarios developed so far: *Explore*, *Access*, *Download* and *Contribute*. In the *Explore* scenario, researchers can reuse ontologies that are interesting for their research. In the *Access* scenario developers can use information in ontologies programmatically to provide semantically enriched applications and web services. In the *Download* scenario, information from the ontologies can be retrieved and stored to a local information system. In the *Contribute* scenario, we consider that scientists can store their terminologies in the TS to access all provided services automatically. Finally, we discuss existing challenges in this field that are often in the social-technical context.

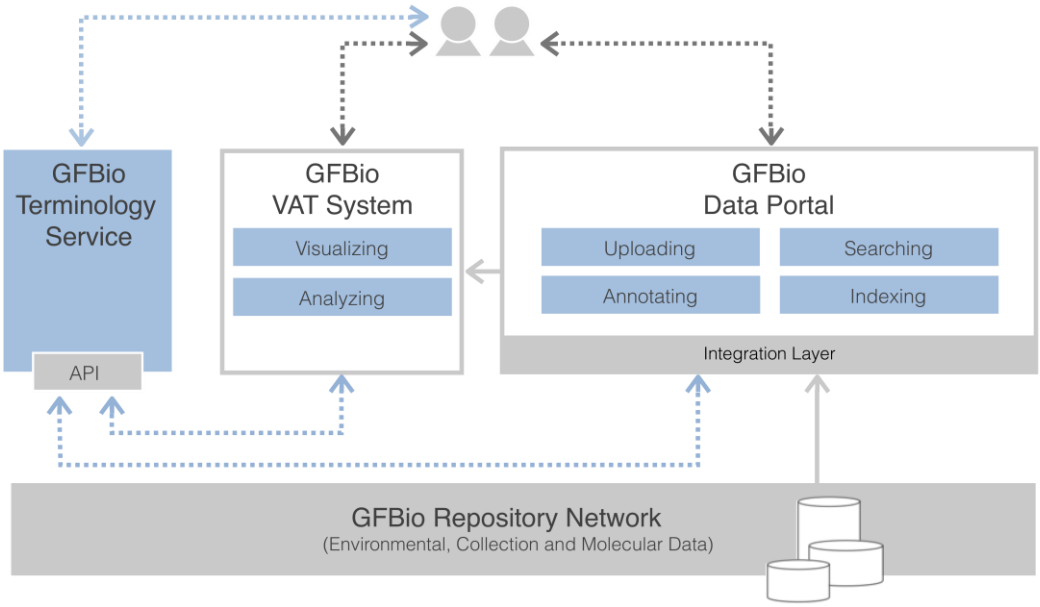


Figure 1. The GFBio components

3 https://terminologies.gfbio.org/

A common infrastructure for biological data

GFBio (Diepenbroek et al. 2014) is developing an infrastructure to enable biological and environmental scientists to share and discover data more efficiently. It aims at providing data management and data archiving solutions for data capture, annotation, indexing, searching and storage. These solutions range from tailored Excel spreadsheets to virtual research environments, such as the Diversity Workbench (Triebel et al. 1999), the Bexis system (Gerlach et al. 2015) or the EDIT Platform for Cybertaxonomy (Ciardelli et al. 2009). Figure 1 presents an overview of the research infrastructure of GFBio, consisting of four main components.

The GFBio Data Portal integrates existing data infrastructures into the GFBio Repository Network (bottom in Fig. 1). The latter comprises amongst others molecular data (EMBL-EBI[4]), environmental data (PANGAEA[5]), as well as natural history and culture collection data (e.g. MfN[6] , DSMZ[7] and SNSB[8]).

The data provided by portal users are indexed and semantically enriched, thereby providing the data with meaning. Analysis and visualization tools allow researchers to better understand the data,, for example, by using the GFBio VAT System (Visualization, Analysis & Transformation system) (Authmann et al. 2015). The possibility to enrich data with semantic information is provided by a fourth component - the GFBio Terminology Service. The semantic meaning is enabled by the provision and interlinking of ontologies and taxonomies.

There are existing systems providing a comparable terminology service. These systems can be either full platforms for terminology management (Noy et al. 2009; Côté et al. 1006; Suominen et al. 2014; Hoehndorf et al. 2015; Xiang et al. 2011) or frameworks for accessing terminologies (Adamusiak et al. 2011; Viljanen et al. 2012). We defined a set of requirements related to our project needs and analyzed to what extent existing systems meet those requirements (Karam et al. 2016). One requirement was to be able to integrate well established taxonomies like the World Register of Marine Species (WORMS[9]) or the Catalogue of Life (COL[10]). Those taxonomies are widely used in the domain for annotating species, for example, and they are a source of valuable hierarchical information. None of the existing systems integrate this type of terminologies. Additional requirements relate to our project's philosophy, to provide tools and inference mechanisms specifically tailored to the requirements of GFBio's partners. These derived insights motivated our decision to set up our own system – the GFBio Terminology Service – that is introduced in the next section.

4 The European Bioinformatics Institute (www.ebi.ac.uk)
5 www.pangaea.de
6 Naturkundemuseum (www.naturkundemuseum.berlin)
7 Leibniz-Institut - Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (www.dsmz.de)
8 Staatliche Naturwissenschaftliche Sammlungen Bayerns (www.snsb.mwn.de)
9 World Register of Marine Species (www.marinespecies.org)
10 Catalogue of Life (www.catalogueoflife.org)

The Terminology Service

We describe in this section the main building blocks of the GFBio Terminology Service. First, we introduce the basic concepts and define the meaning of *terminology* in the context of the GFBio project, then, we present the general architecture of the GFBio TS.

Basic Concepts

The term *terminology* refers to any terminological resource, this can be a formal ontology, a taxonomy, or any useful source of Semantic Web compliant collections of terms (e.g. locations available via a geographical database like Geonames[11]). It encompasses several meanings ranging from simple lists of terms to semantically rich ontologies. Unfortunately, there are currently no commonly accepted definitions of the different terminology types (in the biological domain), which leaves room for variation causing them to be used interchangeably depending on the context.

We introduce our concept of agreed terminology formality levels, with differing levels of specifications going from the most informal to the most formal level as described in Figure 2. The different levels are illustrated by the term *water* (*http://purl.obolibrary.org/obo/CHEBI_15377*) that is extracted from the CHEBI ontology[12] and depicted in Figure 3.

GFBio defines five different types or formality levels in terminologies. The less formal level contains a *Controlled Vocabulary*. It is the simplest type of terminology and consists of a finite list of terms. These labels have no definitions or hierarchical ordering. Based on the example, only the label *water* is part of the terminology.

The next formality level is *Glossary*. It is a list of term labels that additionally includes an informal definition of their meaning in natural language (i.e. human readable language). Since information expressed in natural language is typically not unambiguous, these specifications are not yet adequate for further processing by computer agents. In a glossary, the definition of the term *water* is partnered by its label.

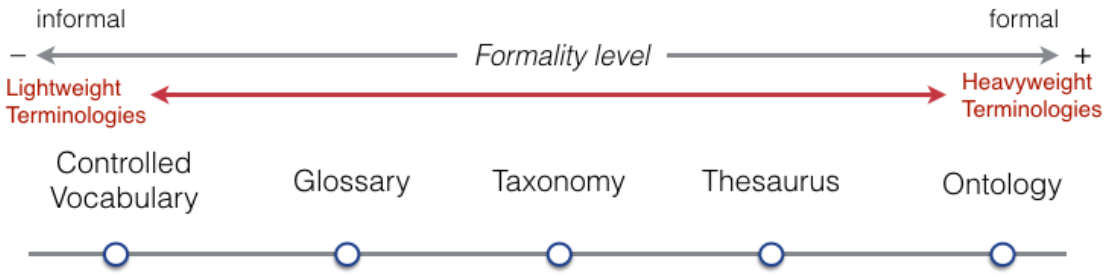


Figure 2. GFBio agreed terminology formality levels

In a *Taxonomy*, a term is a compound of a label, a definition and hierarchical information, e.g., by is-a relationships, thus providing additional semantics in the relations between the terms which can be interpreted by computer agents. The hierarchical structure depicted in Figure 3 would be part of a taxonomy describing the term *water*.

11 www.geonames.org
12 www.ebi.ac.uk/chebi/

A *Thesaurus* is a controlled vocabulary connected via relations between the terms expressing hierarchies (e.g., *narrower/broader term*), associations (e.g., *related term*), or synonym relationships. In the example, a thesaurus contains the information about the synonym *oxidane* of the term *water*.

The most formal terminology is an *Ontology*. A term consists of all the information provided at the lower levels augmented with complex relationships, allowing an unambiguous interpretation of terms and relationships according to logic-based rules. In our example, an ontology would contain the whole spectrum of relations we already considered in the other levels and additional complex or user defined relations like *has_role* and *is_conjugat_base_of*.

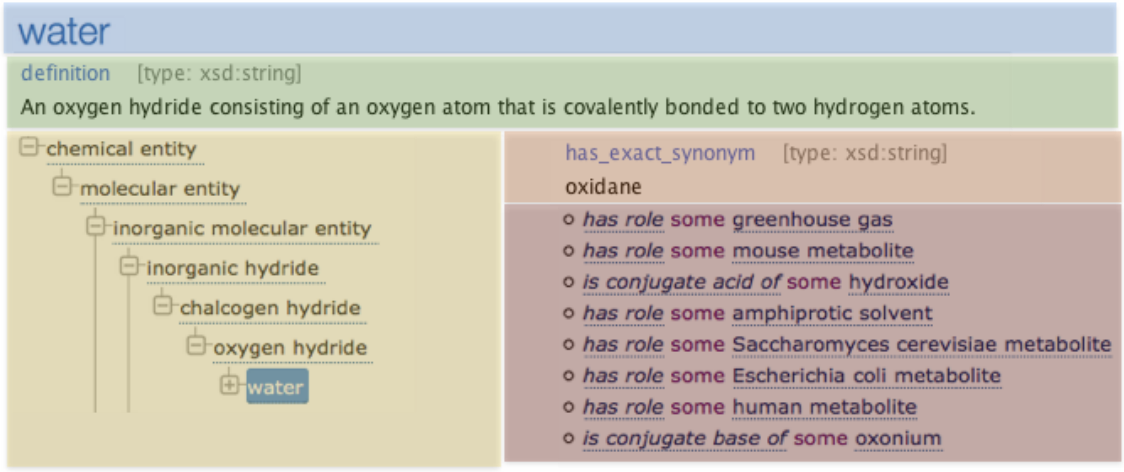


Figure 3. Excerpt of the definition of the term *water* of the CHEBI ontology

The Terminology Service Architecture

The general architecture of the Terminology Service is shown in Figure 4. In March 2017, the Terminology Service gives access to over 20 terminologies that have been requested by the GFBio partners so far. Those terminologies are either internally stored in a Semantic Web repository or remotely accessed via their web services. Internal terminologies are stored in a local RDF[13] store in a Semantic Web compliant format such as OWL[14] or SKOS[15]. Internal terminologies can be accessed directly via a Linked Data interface and a SPARQL[16] endpoint. The included terminologies are well established ones like the CHEBI ontology[17], for example, or ontologies provided by the GFBio community like the KINGDOM[18] ontology, describing a GFBio agreed list of species kingdoms. The complete list and actual status of included terminologies can be found in our technical report (Karam et al. 2017).

13 www.w3.org/RDF
14 www.w3.org/OWL
15 www.w3.org/2004/02/skos/
16 https://terminologies.gfbio.org/sparql
17 www.ebi.ac.uk/chebi/
18 https://terminologies.gfbio.org/describe/?url=http://terminologies.gfbio.org/terms/KINGDOM

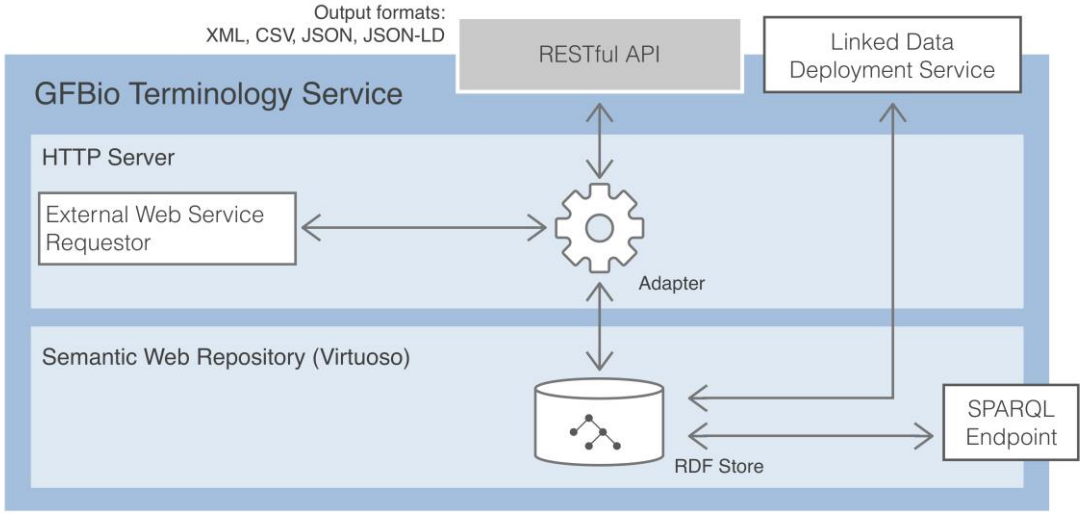


Figure 4. The GFBio Terminology Service architecture

The Terminology Service software is being developed using Java based on the Jena[19] Semantic Web framework. We implemented an external web service requestor for obtaining seven external taxonomies (such as the COL - Catalogue of Life). A key component of the TS is the adapter component (cf. the gear wheel in Figure 4) that enables the schema mapping of both internal and external terminological resources into a common output format. We defined a common schema for the Terminology Service output. A mapping to this schema is required for every underlying terminology or connected external service in order to achieve a harmonized API output. For instance, the COL attribute *name* is mapped to the GFBio TS attribute *label*. Thus, all terms and terminologies can be accessed via a common interface (the RESTful API), regardless of whether they are hosted internally or externally. The service output is delivered in four formats: JSON, XML, CSV, and JSON-LD. This interface allows developers who are not familiar with semantic technologies or Linked Data to easily access the provided terminologies efficiently.

Accessing the Terminology Service

The GFBio Terminology Service can be accessed either through a common interface - the RESTful API[20] - or using widgets we provide; these are small web applications with limited functionality which allow for user interactions. We describe in the following both ways to access the GFBio TS.

The Terminology Service API

The RESTful API of the TS can be used programmatically by connecting the service to other web services such as the GFBio Data Portal, the VAT (cf. Figure 1) or other applications. At the mo-

19 https://jena.apache.org/
20 Application Programming Interface

ment the API provides 14 endpoints that are organised into terminology-specific, term-specific, search, and hierarchy-oriented endpoints. Details about the call's signatures, the parameters and examples for using the service can be found in the API documentation section on our website (terminologies.gfbio.org). In the following, we describe each category briefly, a tabular description for each endpoint can be found in our technical report (Karam et al. 2017).

Terminology-specific endpoints

The four terminology-specific endpoints provide information on terminologies like the list of available terminologies and their metadata, such as the name, description and creation date.

Term-specific endpoints

Term-specific endpoints relate to particular terms from the terminologies. One can list all terms of a specific terminology, query the information about a term or get the list of its synonyms.

Search endpoints

Two search endpoints are provided, the first one returns all terms corresponding to a query string, the second is implemented for suggesting terms while users are typing.

Hierarchy-oriented endpoints

Hierarchy-oriented endpoints return information relative to the position of a term in the hierarchical structure of the terminology. Broaders and narrowers terms of a given term can be returned as well as the complete hierarchical path up to the top of the hierarchy.

The Terminology Service Widgets

The Terminology Service provides widgets – that are components, "chunks of web page" or small applications – intended to be used within web pages. The widgets deliver a restricted functionality, often for just one purpose, like displaying data or providing an interface. Typically, a widget contains a mixture of HTML, CSS and JavaScript where the complexity is ideally hidden so as to make it as easy as possible for developers to integrate the widgets in their application or website. All of our widgets use the Terminology Service API and thus, users can quickly expand their local service with all the functionalities provided by the GFBio TS API. Our goal is to provide reusable and easy to use widgets to be integrated and reused easily with none or little knowledge in web development. Furthermore, the widgets are licensed under an open source licence and will be published openly on Github soon. At the moment, we prototypically implemented two widgets: a term visualisation and a search widget. In the following, we take the latter as an example, to show the methodological approach for developing widgets.

The search widget allows users to search for terms from terminologies to determine their usefulness for their work, e.g. for annotating research data in the GFBio Data Portal. Before developing this widget, we examined 13 services which provide search functionalities in the same or related fields as ours. The majority (6) of the examined services allowing to look for classes (terms) *in* particular ontologies or vocabularies (Cropontology[21], Finto[22] (Suominen et al. 2014), Ontobee[23] (Xiang et al. 2011), Aber-Owl[24] (Hoehndorf et al. 2015), Bioportal[25] (Noy et al. 2009), OLS[26] (Côté et al. 2006)). The latter three are capable of searching *for* ontologies as well. Three services (Biosharing[27] (McQuilton et al. 2016), VEST[28] (Vest / AgroPortal map 2017), ANDS[29] (Australian National Data Service 2017)) looking for vocabularies, ontologies, policies or standards only and four (Datacite[30] (Datacite 2017), Dryad[31] (Dryad Digital Repostory 2017), F1000research[32] (F1000research 2017), Vertnet[33] (Vertnet 2017)) are for searching scientific papers and data resources. The appearance of the search interface differs a lot. From very simple interfaces to advanced ones with many search options and filter functionalities. We examined design criteria like the overall size of the widget, the position and layout of the submit button, the placeholder text of the search bar, the availability and presentation of advanced search functionalities and help sections. The main considerations are described in detail in our technical report (Karam et al. 2017), they resulted in the prototypical design depicted in Figure 5.

The development process included the investigation of a widget scaffold where the objective was twofold: (1) the development process for further widgets should be simplified and standardized, and (2) the process for developers to integrate the GFBio TS widgets into their websites should be supported. We then investigated three services (Google[34], Twitter[35] and ANDS (Australian National Data Service 2017)) that are providing customized widgets. With some kind of guidance users are able to click through options on the website to receive customized HTML code and references to JavaScript and style files to be embedded on their own website. As customisation is planned but not implemented yet, our goal is to deliver one JavaScript and one CSS file to be integrated in the users HTML via the corresponding HTML markups. Because our widgets will deliver a broad spectrum of functionality the scaffold consists next to the way how developers integrating it, of the module design pattern, used libraries, a shared layout file and partly shared functions.

21 www.cropontology.org
22 www.finto.fi
23 www.ontobee.org
24 www.aber-owl.net
25 http://bioportal.bioontology.org
26 www.ebi.ac.uk/ols
27 http://biosharing.org
28 http://vest.agrisemantics.org/vocabularies
29 http://vocabs.ands.org.au
30 www.datacite.org
31 www.datadryad.org
32 http://f1000research.com
33 http://portal.vertnet.org
34 https://developers.google.com
35 https://dev.twitter.com

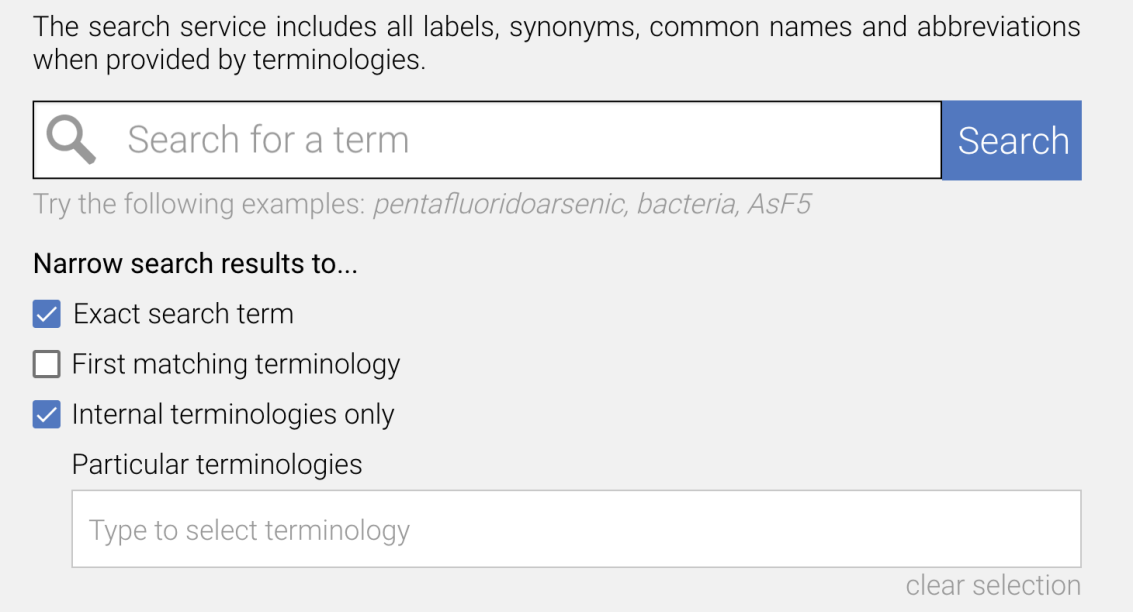


Figure 5. Screenshot of the GFBio TS search widget prototype

Using the Terminology Service within GFBio

Currently, the GFBio community uses the Terminology Service within four main scenarios. Each scenario has been defined and developed in cooperation with GFBios partners. Each partner provides discipline and context specific requirements to the GFBio TS. The development of these use cases is an ongoing process and further use cases will be provided in the near future.

In the *Browse* scenario, users (i.e. researchers) can peruse terminologies that are interesting for their research. For this, the visualization widget provides term details and shows a term's position within a tree structure, if the terminology is a taxonomy or in a graph structure, if the terminology is an ontology. In the GFBio Data Portal the visualization can be used in the research data submission process. When annotating the data in the submission process, the user can easily browse term details and explore existing term relations by type to identify those terms that describe their data best.

In the *Access* scenario developers can use information in terminologies programmatically to provide semantically enriched web services based on the GFBio TS. In the GFBio Data Portal, the TS allowed for developing a semantic search service for research data. Based on query expansion, the original search term is extended by related terms from different terminologies in order to provide a more comprehensive overview of existing research data.

In the *Consume* scenario, information from terminologies of the GFBio TS can be retrieved and stored to a local information system. In the GFBio context, this is needed for data management within small and medium scale projects that are carried out by virtual research environments such as BExIS (Gerlach et al. 2015) and Diversity Workbench (Triebel et al. 1999). In these contexts, the provided metadata from the terminologies of the TS can be pre-processed to support the data annotation process locally.

In the *Contribute* scenario we consider that researchers or data curators can store their individual terminologies in the GFBio TS. Instead of developing their own terminology management

system, this will allow them to access all services provided by the TS easily. For example, in the GFBio context, partners have already contributed ten terminologies. These terminologies are either internally stored like the KINGDOM ontology[36] or connected as external web services like the DTN Taxon Lists Services[37] or the Prokaryotic Nomenclature Up-to-Date[38] and interna. In GFBio, the mobilization of community-relevant terminologies is supported by an internal process. The terminology owner can register the terminology in the internal wiki and in collaboration with the terminology curator the needed metadata are provided. If the metadata are complete, a terminology is manually integrated into the TS.

Current activities and next steps

We introduced the GFBio TS that extends the GFBio infrastructure with semantic capabilities. This extension enables researchers to share their data despite their heterogeneous nature. After presenting the project context and the basic concepts, we described the general architecture of the Terminology Service and the way to access and integrate it using its public interface or via a set of downloadable widgets.

We described concrete use cases that support researchers at different levels in their research practice, for example, when searching for datasets or when using up-to-date terminologies in their virtual research environments.

At the moment, a high level application ontology, the GFBio ontology is being developed. It will enable interoperability between the various terminologies available by defining higher level links between them. Moreover, this ontology will serve mainly as a basis for annotations and automated faceted search.

We are working on the integration of the semantic annotation tool neonion (Müller-Birn et al. 2017) within the GFBio context. The aim is to allow scientists to annotate information in scientific texts with terminologies coming from the GFBio TS, and thus, research results and research data can be more closely connected.

The interoperability issue is due to different understandings of terms within different scientific domains or to the use of different labels to refer to the same term. This issue can be solved by annotating data with terms from the Terminology Service. Data can still be annotated using equivalent terms coming from different terminologies. In order to ensure interoperability the underlying terminologies should be interlinked. We are developing a semi-automated mapping service and interface based on a combination of matching algorithms.

The GFBio TS is continuously updated to meet partners needs. A set of tools is being developed to support terminologies selection based on query and text analysis as well as tools for transforming terminologies from text and tabular forms into a Semantic Web compliant format.

36 https://terminologies.gfbio.org/api/terminologies/KINGDOM/
37 http://www.diversitymobile.net/wiki/DTN_Taxon_Lists_Services
38 https://bacdive.dsmz.de/api/

References

Noy, Natalya Fridman, Nigam H. Shah, Patricia L. Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, et al. 2009. "BioPortal: ontologies and integrated data resources at the click of a mouse." *Nucleic Acids Research* 37, Web-Server-Issue: 170–173.

Côté, Richard G., Philip Jones, Rolf Apweiler, and Henning Hermjakob. 2006. "The Ontology Lookup Service: a lightweight cross-platform tool for controlled vocabulary queries." *BMC Bioinformatics* 7 (1): 1–7.

Suominen, Osma, Sini Pessala, Jouni Tuominen, Mikko Lappalainen, Susanna Nykyri, Henri Ylikotila, Matias Frosterus, and Eero Hyvönen. 2014. "Deploying National Ontology Services: From ONKI to Finto." In *Proceedings of the Industry Track at the Intl. Semantic Web Conference 2014*. Riva del Garda, Italy: CEUR Workshop Proceedings, October.

Hoehndorf, Robert, Luke Slater, Paul N. Schofield, and Georgios V. Gkoutos. 2015. "Aber-OWL: a framework for ontology-based data access in biology." *BMC Bioinformatics* 16 (1): 1–9.

Xiang, Zuoshuang, Chris Mungall, Alan Ruttenberg, and Yongqun He. 2011. "Ontobee: A Linked Data Server and Browser for Ontology Terms." In *Proceedings of the 2nd Intl. Conference on Biomedical Ontology*, Buffalo, NY, USA, July 26-30.

Adamusiak, Tomasz, Tony Burdett, Natalja Kurbatova, K. Joeri van der Velde, Niran Abeygunawardena, Despoina Antonakaki, Misha Kapushesky, Helen Parkinson, and Morris A. Swertz. 2011. "OntoCAT – simple ontology search and integration in Java, R and REST/JavaScript." *BMC Bioinformatics* 12 (1): 1–12.

Viljanen, Kim, Jouni Tuominen, Eetu Mäkelä, and Eero Hyvönen. 2012. "Normalized Access to Ontology Repositories." In *Proceedings of the Sixth International Conference on Semantic Computing* (IEEE ICSC 2012). Palermo, Italy: IEEE Press, September.

Karam, Naouel, Claudia Müller-Birn, Maren Gleisberg, David Fichtmüller, Robert Tolksdorf, and Anton Güntsch. 2016. "A Terminology Service Supporting Semantic Annotation, Integration, Discovery and Analysis of Interdisciplinary Research Data." *Datenbank-Spektrum* 16 (3): 195–205.

Diepenbroek, Michael, Frank Oliver Glöckner, Peter Grobe, Anton Güntsch, Robert Huber, Birgitta König-Ries, Ivaylo Kostadinov, et al. 2014. "Towards an Integrated Biodiversity and Ecological Research Data Management and Archiving Platform: The German Federation for the Curation of Biological Data (GFBio)." In *44. Jahrestagung der Gesellschaft für Informatik*, Stuttgart, Germany. Ausgabe 232. LNI. GI, isbn: 978-3-88579-626-8.

Triebel, Dagmar, Gregor Hagedorn, Stefan Jablonski, and Gerhard Rambold (eds.). 1999. "Diversity Workbench: A virtual research environment for building and accessing biodiversity and environmental data." Online available http://www.diversityworkbench.net.

Gerlach, Roman, David Blaa, Javad Chamanara, Martin Hohmuth, Nafiseh Navabpour, Sven Thiel, and Birgitta König-Ries. 2015. "BEXIS 2: A platform for managing heterogeneous biodiversity data and projects." In *TDWG Annual Conference*.

Ciardelli, Pepé, Patricia Kelbert, Andreas Kohlbecker, Niels Hoffmann, Anton Güntsch, and Walter G. Berendsohn.2009. "The EDIT Cyberplatform for Taxonomy and the Taxonomic Workflow: Selected Components." In *39. Jahrestagung der Gesellschaft für Informatik e.V. (GI)*, Lübeck, Germany, 625–638.

Authmann, Christian, Christian Beilschmidt, Johannes Drönner, Michael Mattig, and Bernhard Seeger. 2015. "VAT: A System for Visualizing, Analyzing and Transforming Spatial Data in Science." *Datenbank-Spektrum* 15 (3): 175–184.

"Crop Ontology curation and annotation tool – 2011 Generation Challenge Programme, Bioversity International as project implementing agency." Accessed: 2017-01-06.

McQuilton, Peter, Alejandra Gonzalez-Beltran, Philippe Rocca-Serra, Milo Thurston, Allyson Lister, Eamonn Maguire, and Susanna-Assunta Sansone. 2016. *BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences*. Database 2016 (2016): baw075.

"Australian National Data Service website." Online available www.ands.org.au. Accessed: 2017-01-06.

"VEST / AgroPortal - Map Of Standards website." Online available http://vest.agrisemantics.org/vocabularies. Accessed: 2017-01-06.

"Datacite website." Online available www.datacite.org. Accessed: 2017-01-06.

"Dryad Digital Repository website." Online available http://www.datadryad.org. Accessed: 2017-01-06.

"F1000Research website." Online available https://f1000research.com. Accessed: 2017-01-06.

"VERTNET: Distributed Databases with backbone website." Online available http://portal.vertnet.org. Accessed: 2017-01-06.

Morville, Peter, and Jeffery Callender. 2010. *Search Patterns: Design for Discovery*. I–X, 1–180. O'Reilly, isbn: 978-0-596-80227-1.

Turbek, Steve. 2008. "Advancing advanced search." Online available http://boxesandarrows.com/advancing- advanced-search. Accessed: 2017-01-13.

Müller-Birn, Claudia, Tina Klüwer, André Breitenfeld, Alexa Schlegel, and Lukas Benedix. 2015. "neonion: Combining Human and Machine Intelligence." In *18th ACM Conference on Com-*

puter Supported Cooperative Work & Social Computing, CSCW 2015, Vancouver, BC, Canada, March 14-18, 2015, Companion Volume, 223–226.

Karam, Naouel, Robert Harald Lorenz, and Claudia Müller-Birn. 2017. "The GFBio Terminology Service: Enabling a research data management beyond data heterogeneity". *Technical Report* Ser. B TR-B-17-01. Freie Universität Berlin, Institut für Informatik, March.

Distributed Research Data Management - Plädoyer für eine verteilte Forschungsdaten-Infrastruktur

Reiko Kaps[1]

1 Leibniz Universität Hannover

Zusammenfassung. Klassisches Forschungsdatenmanagement sieht die Forschenden eher als Kunden denn als Mitstreiter. Beratung und Angebote etwa zur Veröffentlichung von Forschungsdaten orientieren sich an Service-Konzepten, die Informationen zentral verteilen und vorhalten. Diese Einstellung spiegelt sich auch bei Forschungsdaten-Repositorien wider.

Diese Publikationsplattformen eignen sich damit zwar gut als Schaufenster, weniger oder oft gar nicht als Arbeits- und Austauschplattform für Forschungsdaten. Zudem sind Forschungsdaten-Repositorien für die dringendsten Probleme der Forschenden nur unzureichend gerüstet: Dazu zählen die Sicherung der Urheberschaft, die Wahrung der Unversehrtheit von veröffentlichten Daten sowie die dauerhafte Verknüpfung mit Metadaten und Lizenzinformationen. Herkömmliche IT-Dienste gewährleisten solche Zusicherungen nur innerhalb ihres Systems. Verlassen die veröffentlichten Daten dieses Refugium, können Informationen zu Urheberschaft und Lizenz in Gefahr geraten.

Jenseits des klassischen Client-Server-Ansatzes existieren andere Verfahren, um Daten und Dienste im Internet zu verbreiten und deren Authentizität sicherzustellen. Diese dezentralen Ansätze setzen auf Peer-to-Peer-Techniken und neuerdings auch auf Blockchain-Verfahren, die die mathematische Grundlage für Kryptowährungen bilden. Blockchain-Verfahren sichern Transaktionen, stellen "Einigkeit" unter den beteiligten Knoten her und eignen sich so für den sicheren, vertrauenswürdigen und nachvollziehbaren Austausch beliebiger Daten – also auch von Forschungsdaten.

Inzwischen stehen Anwendungen bereit, die Dateien dezentral verteilen (IPFS), Daten strukturiert ablegen (BigchainDB) und untrennbar mit Metainformationen zu Urheber, Lizenz etc. verbinden (Mediachain/IPDB). Diese Programme benötigen keine zentrale Instanz, sichern die Datenintegrität, verfolgen Änderungen und verteilen effektiv Datenbestände. Jeder kann damit nicht nur Daten anbieten und abrufen, sondern auch auf dieser Basis eigene Anwendungen entwickeln.

Das folgende Papier stellt diese Ansätze vor, zeigt deren Vorteile gegenüber klassischen Client-Server-Angeboten und skizziert, wie sie das Datenmanagement für veröffentlichte Forschungsdaten verbessern und vereinfachen können.

Datarefuge

Das Musterland des Forschungsdatenmanagement erlebt derzeit eine einzigartige Bewegung. Seit Monaten treffen sich fast jedes Wochenende ganze Scharen von Wissenschaftlern, Programmierern und Aktivisten in Universitätsbibliotheken, Hackerspaces und anderen Orten, um öffentlich zugängliche Forschungsdaten zu retten. Dabei handelt es sich um Daten aus der Klimaforschung öffentlicher Bundeseinrichungen wie der Umweltbehörde EPA und der NASA, die diese auf ihren eigenen Servern bislang zum Download anbieten.

Abbildung 1. Datarefuge auf Twitter (Screenshot 2017)

Diesen Daten droht durch den neuen US-Präsidenten Gefahr, denn bereits im Wahlkampf äußerte sich Donald Trump sehr ablehnend gegen die Forschung zum Klimawandel und deren Erkenntnisse[1]. Das noch unter seinem Vorgänger wichtige Thema verschwand sehr bald nach Trumps Amtsübernahme von einigen offiziellen Webseiten und ist nun nur noch im Web-Archiv des Weißen Hauses zu finden[2].

Seit der Wahl versuchen zahleiche US-Wissenschafler daher, diese drohenden Verluste abzuwenden: Dazu müssen sie die Daten mühsam von den Behörden-Webservern laden und auf externen, oft in Kanada stehenden Servern auslagern.

Da die Daten oft nicht einfach kopiert werden können, sondern sich hinter aufwendigen Webanwendungen verstecken, ist diese Rettungsaktion kein leichtes Unterfangen. Die an #Datarefuge Beteiligten müssen zu diesem Zweck Webseiten und Daten analysieren und Skripte für den Download und die Konvertierung programmieren.

Client-Server-Dilemma

Warum müssen die US-Wissenschalfter und -Aktivisten jedoch diesen Aufwand treiben? Neben den politischen Rahmenbedingungen drängt sich dabei eine grundsätzliche Frage auf: Liegt der eigentliche Grund für diesen massiven Rettungsaufwand womöglich in der Art, wie wissenschaft-

1 http://www.stern.de/politik/ausland/donald-trump--wissenschaftler-retten-klima-daten-vor-seinem-amtsantritt-7241850.html
2 http://www.zeit.de/politik/ausland/2017-01/donald-trump-website-weisses-haus-klimawandel

liche Einrichtungen in der Vergangenheit und aktuell ihre Forschungsdaten speichern und veröffentlichen?

Wissenschaftliche Einrichtungen veröffentlichen derzeit Forschungsdaten und deren Ergebnisse über Plattformen, die dem Client-Server-Modell entsprechen. Das heißt, ein oder auch mehrere zentrale Server verteilen auf ihnen vorgehaltene Daten an Interessenten (Clients). Diese Architektur ist vergleichsweise anfällig für Ausfälle (single point of failure). Wer dabei auf Redundanz und Verfügbarkeit wert legt, muss sich beide Punkte teuer erkaufen. Das erschwert die Skalierbarkeit, die ausschließlich der Server-Betreiber sicherstellen kann. Das gilt besonders dann, wenn der bereitgestellte Dienst tatsächlich erfolgreich ist: Steigt die Nachfrage nach diesen Daten, müssen sie an jeden Client einzeln ausgeliefert werden: Abhängig von den vorhandenen Netzwerkressourcen droht der Dienst an seinem eigenen Erfolg zu scheitern.

Gleichzeitig forciert der Client-Server-Ansatz Diensteinseln. Interoperabilität und leichter Datenaustausch mit anderen bleiben nachrangige Aufgaben oder sie müssen durch umfangreiche Regelwerke sichergestellt werden (RFC/IETF, ISO). Diese Diensteinseln orientieren sich zudem an kommerziellen Angeboten, behandeln ihre Nutzer fast immer als Kunden und betrachten ihre Daten als Ausstellungsgegenstand.

Auf dem Client-Server-Modell aufsetzende Dienste haben in der Vergangenheit massgeblich zum Wachstum des Internets beigetragen. Die steigende Verbreitung des Internets und neue Herausforderungen bei Datenhaltung und -verteilung lassen dieses Konzept jedoch an seine Grenzen stoßen.

Neue Herausforderungen

Datacenter- und Infrastruktur-Betreiber stehen vor der Aufgabe, immer größere Datensätze bereitzustellen und an immer mehr Teilnehmer zuverteilen. So hat sich der Datenverkehr am deutschen Internet-Knoten DE-CIX in den vergangenen 5 Jahren mehr als verdoppelt[3].

Während im kommerziellen Bereich dabei besonders Videodaten die Entwicklung vorantreiben, sind es in der Wissenschaft digitale Forschungsdaten: Zu den Messreihen in überschaubaren Textdateien haben sich längst Bilder, Videos, Simulationen, Visualisierungen und Social-Media-Logs gesellt, die den Umfang digitaler Forschungsdaten massiv nach oben treiben. Anders als bei anderen Daten unterliegen Forschungsdaten jedoch Ansprüchen, die weit über das Speichern und Verteilen hinausgehen.

Im Kern handelt es sich dabei um die Forderungen der öffentlichen Förderer, die Forschungsdaten für wenigstens 10 Jahre erhalten und sie für die Öffentlichkeit nutzbar machen wollen. Forschungsdaten müssen daher ausreichend dokumentiert sein sowie in Formaten vorliegen, deren Aufbau offengelegt ist. Außerdem muss ihre Herkunft gesichert sein und ihre Entstehungs- und Bearbeitungswege nachvollzogen werden können (Integrität, Versionierung, Verknüpfung). Ohne diese Randbedingungen sind veröffentlichte Forschungsdaten kaum oder nur mit hohen Aufwand durch Interessierte nutzbar – selbst wenn diese der derselben Fachdisziplin angehören.

3 https://www.de-cix.net/en/locations/germany/frankfurt/statistics

Auswege

Die skizzierten Probleme lassen sich bereits mit vorhandenen und erprobten Techniken lösen, die sowohl auf Peer-to-Peer-Prinzipien als auch das tradierte Client-Server-Modell setzen.

Unter dem Begriff Peer-to-Peer (P2P) versammeln sich Techniken, die ohne die zentralen Instanzen des Client-Server-Modells auskommen. In Netzwerken arbeiten solche Systeme gleichzeitig als Server und als Client, sodass sich ein Netz aus (grundsätzlich gleichberechtigten) Knoten aufspannt. Diese Knoten verteilen die für die Peer-to-Peer-Anwendung nötigen Informationen.

Eine der bekanntesten Peer-to-Peer-Anwendungen dürfte das im Jahr 2001 entwickelte Bittorrent-Protokoll sein, das große Dateien zuverlässig und effektiv verteilt. Im Unterschied zu den von Standardisierungsgremien gepflegte HTTP oder FTP nutzt Bittorrent auch die ansonsten ungenutzten Upload-Kapazitäten der jeweiligen Download-Knoten, sodass eine Datei nicht nur vom Anbieter selbst sondern auch von denjenigen verteilt wird, die diese Datei herunterladen[4]. Anhand von Prüfsummen stellt Bittorrent dabei sicher, dass die übertragene Datei dem Orginal entspricht. Dank dieser Fähigkeiten erlangte Bittorrent einerseits eine durchaus zweifelhafte Berühmtheit als Dateitauschbörse für digitale Medien wie Musik und Spielfilme, andererseits stellte das Protokoll damit auch sein Fähigkeiten als effektiver Dateiverteiler unter Beweis[5].

Für den Umgang mit Dateien und Inhalten steht inzwischen die Dateiversionierungssoftware Git bereit, die ebenfalls auf Peer-to-Peer-Techniken setzt. Versionsverwaltungen wie Git protokollieren Änderungen an Dateiinhalten, erlauben verteilte und nicht-lineare Arbeitsabläufe und gewährleisten damit Nachvollziehbarkeit. Viele Funktionen von Git lassen auch ohne permanente Internet-Verbindung einsetzen.

Eine weitere Technikentwicklung der vergangenen Jahre erlaubt eine lückenlose Buchführung von Aktionen: Die Blockchain ist eine Datenbank mit mathematischem Integritätsansatz. Sie speichert den Hashwertes (Integritätsgarant) eines Datensatzes im Hashwert des jeweils nachfolgenden. Die Technik ähnelt dem Journal in der Buchführung und stellt damit die Grundlage für Kryptowährungen bereit[6]. Mittels einer Blockchain lassen sich sowohl die Transaktionsicherheit als auch die Nachvollziehbarkeit in verteilten Systemen erheblich vereinfachen und verbessern.

IPFS

Jede der bereits genannten Peer-to-Peer-Techniken löst nur Teile der eingangs geschilderten Probleme. Allerdings steht seit dem Jahr 2014 das Interplanetary File System (IPFS) als quelloffene Implementierung bereit, die Web-, Bittorrent-, Git- und Blockchain-Funktionen in einem verteilten Dateisystem vereint[7]. Das IPFS-Konzept und die in Go geschriebene Referenzimplementierung stammen von Juan Benet[8].

IPFS stellt ein vollständig verteiltes Netzwerk-Dateisystem bereit. Darin arbeiten alle Knoten sowohl als Server als auch als Client und jeder Knoten ist mit jedem anderen verbunden (siehe

4 https://de.wikipedia.org/wiki/BitTorrent
5 https://torrentfreak.com/bittorrent-dominates-internet-traffic-070901/
6 https://de.wikipedia.org/wiki/Buchf%C3%BChrung#Journal_.28Grundbuch.29
7 https://ipfs.io/
8 https://github.com/ipfs/papers/raw/master/ipfs-cap2pfs/ipfs-p2p-file-system.pdf, https://github.com/ipfs/ipfs, https://twitter.com/juanbenet

Abbildung 1). IPFS adressiert Inhalte über Hashes, macht Dateien über lesbare Namen auffindbar (IPNS) und dedupliziert Dateien innerhalb seines Dateisystems. Abgerufene Inhalte anderer IPFS-Knoten hält IPFS in einem lokalen Cache vor - bei Bedarf sogar dauerhaft (Pinning). Lädt IPFS fremde Inhalte, versucht es sie von möglichst vielen und nahe gelegenen Knoten abzuholen. Jeder angefragte Knoten liefert dabei Teilstücke einer Datei an den Anfragenden aus. Ähnlich wie Git protokolliert IPFS Änderungen an seinen Dateien respektive Inhalten.

Research Data Federation

IPFS stellt damit Funktionen bereit, aus denen Inhalteproduzenten und -anbieter, Forschende, Infrastrukturbetreiber und Bibliotheken sowie digitale Archive Profit ziehen können. Mit IPFS und verwandten Techniken lässt sich eine globale Infrastruktur für Forschungsdaten aufbauen, bei der jeder Nutzer zum Teil der Infrastruktur wird. Anders als Client-Server-Konzepte zeichnet sie sich durch eine hohe Zensurresistenz und Ausfallsicherheit aus. Sie nutzt Netzwerkressourcen optimal, verbessert die Netzsicherheit und erlaubt eigene Anwendungen, die auf IPFS aufsetzen.

Beispielsweise kann IPFS dabei helfen, bessere und günstigere Lösungen für das Problem der Zweitkopie zu finden. Forschungsdaten-Repositorien und andere digitale Archive verhindern mit einer Zweitkopie die ungewollte Alterung und Verfälschung ihrer Archivobjekte, die etwa durch Bitfehler verursacht werden können. Diese Maßnahme verdoppelt jedoch die Kosten für jede gespeicherte Datei, denn es sind dafür zusätzliche Speichermedien und Standorte nötig. Gerade für kleinere wissenschaftliche Einrichtungen dürfte der Betrieb mehrerer Systeme an unterschiedlichen Standorten ein nur schwer lösbares Problem sein.

In einer auf IPFS aufsetzenden Forschungsdaten-Infrastruktur können jedoch andere Teilnehmer die Aufbewahrung der Zweitkopie über das beschriebene Pinning übernehmen. Das eröffnet große Freiheiten beim Speicherort und kann Kosten senken. Diese Zweitkopie-Delegation lässt sich etwa mittels des Gegenseitigkeitsprinzips (Peering) oder über Mietmodelle vertraglich regeln.

Testbed: Call for Participation

Angesichts der genannten Argumente für ein verteiltes Forschungsdaten-Dateisystem und der Notwendigkeit Forschungsdaten effizient vorzuhalten und zu verteilen, schlagen wir einen Feldversuch in Form eines Testbeds vor: Dieses Experiment soll IPFS und ähnliche Techniken auf ihre Praxistauglichkeit untersuchen und Möglichkeiten der Zusammenarbeit mittels dieser Techniken erproben.

Wir wollen dabei sowohl Anwendungen testen, die Forschende direkt an die beschriebene Infrastruktur anbinden und sich damit besser in den wissenschaftlichen Workflow einfügen, als auch Infrastruktur-Konzepte erproben, mit denen Wissenschaftsorganisationen Forschungsdaten leicht und kostengünstig aufbewahren und verteilen können. Zu diesen Punkten können Interessenten jederzeit weitere Themen beitragen.

Dank der Peer-to-Peer-Struktur liegen die Hürden für die Teilnahme niedrig: Neben den klassischen Rechenzentren und Diensteanbietern wie Bibliotheken kann jeder mitmachen, der einen Rechner im Internet betreibt. Das sind im Idealfall ständig laufende Server, können aber auch PCs

und Notebooks sein, die nur temporär arbeiten. Darüber hinaus benötigen die Teilnehmer jedoch den Willen, sich mit den Interna von Peer-to-Peer-Techniken wie IPFS auseinanderzusetzen und dieses Wissen mit anderen zu teilen.
Weitere Fragen beantwortet der Autor gern per E-Mail oder Twitter[9].

9 E-Mail: kaps@luis.uni-hannover.de; Twitter: https://twitter.com/reik_kaps

Herausforderung Forschungsdatenmanagement- Unterstützung der Hochschulen durch eine einrichtungsübergreifende Kooperation in NRW

Constanze Curdt[1], Volker Hess[2], Ania Lopez[3], Benedikt Magrean[4], Dominik Rudolph[5], Johanna Vompras[6]

1 Geographisches Institut, Regionales Rechenzentrum, Universität zu Köln
2 Zentrum für Informations- und Mediendienste , Universität Siegen
3 Universitätsbibliothek, Universität Duisburg-Essen
4 IT-Center, RWTH Aachen University
5 Zentrum für Informationsverarbeitung, Westfälische Wilhelms-Universität Münster
6 Universitätsbibliothek, Universität Bielefeld

Zusammenfassung. Forschungsdatenmanagement ist ein Thema, das aktuell unter anderem insbesondere Hochschulen betrifft. Aufbauend auf einer 2015 kooperativ durchgeführten Bestandsaufnahme zu Forschungsdatenmanagement (FDM) an den Hochschulen in Nordrhein-Westfalen, hat sich 2016 das sogenannte "Fachteam Forschungsdatenmanagement" gebildet, eine Expertengruppe zusammengesetzt aus Vertretern von Bibliotheken und Rechenzentren der Hochschulen. Ziel der Arbeit des Fachteams ist es, die Hochschulen für das Thema FDM zu sensibilisieren und damit hochschulübergreifende Kooperationen für die Entwicklung von Verfahren zu institutionellem FDM zu fördern. Gleichzeitig sollen Konzepte und Handlungsempfehlungen für die Etablierung von flächendeckendem und nachhaltigem FDM an den Hochschulen des Landes erarbeitet werden. Der Beitrag stellt die bisher erzielten Ergebnisse der Arbeit des Fachteams vor und ordnet diese in die deutschlandweiten Initiativen zu FDM ein.

Schlagwörter. Forschungsdatenmanagement, Digitalisierung, Research Data Management, eResearch

Einleitung

Im Zuge der Digitalisierung der Forschung ist in den letzten Jahren ein enormes quantitatives und qualitatives Wachstum von Forschungsdaten – also z.B. Messdaten, Daten aus Erhebungen, Textdaten, Daten aus medizinischen Proben – entstanden. Gleichzeitig wächst das Bewusstsein für den langfristigen Wert dieser Daten, die nicht länger nur als flüchtiges „Abfallprodukt“ des Forschungsprozesses, sondern als wichtige Ressource für zukünftige Forschungsvorhaben zur Vermeidung von kostenintensiver redundanter Forschung und zur Qualitätssicherung und Kontrolle erzielter Ergebnisse dienen können. Dies spiegelt sich auch in den Forderungen und Richtlinien wichtiger Organisationen und Einrichtungen wie der Deutschen Forschungsgemeinschaft (DFG), der Hochschulrektorenkonferenz (HRK), der Allianz der deutschen Wissenschaftsorganisationen und des Rates für Informationsinfrastrukturen (RfII) wider. Alle Hochschulen stehen damit gemeinsam vor der Herausforderung, Strukturen für ein professionelles Forschungsdatenmanagement (FDM) zu schaffen und auszubauen (z.B. technische (Speicher-) Infrastrukturen, Beratungs-

angebote, Richtlinien…), um die Forschenden wirksam zu unterstützen. Aufgrund der gemeinsamen Problemstellung erscheinen hochschulübergreifende Kooperationen der beste Weg zu diesem Ziel zu sein.

Die aktive Förderung von hochschulübergreifenden Kooperationen kann derzeit in einigen Bundesländern beobachtet werden (z.B. Hessen und Baden-Württemberg). In Nordrhein-Westfalen sind bislang erst wenige landesweite Initiativen zu erkennen, etwa der Ausbau von Speicherstrukturen zur landesweiten Nutzung im FDM, der derzeit von drei Konsortien beantragt wird. Es besteht große Heterogenität zwischen den einzelnen Hochschulen bei den bisherigen Fortschritten zur Einrichtung von institutionellem FDM. Insgesamt mangelt es an einer landesweiten Sichtweise bzw. Strategie auf das Thema genauso wie an zentralen Angeboten und Unterstützungsmöglichkeiten für Hochschulen, die das Thema FDM vollständig abdecken. Damit liegt NRW im Vergleich zu anderen Bundesländern in der Entwicklung zurück.

Um hier Abhilfe zu schaffen und zur Unterstützung des kooperativen Gedankens wurde vom DV-ISA (heute DH-NRW) das Fachteam Forschungsdatenmanagement im Frühjahr 2016 aus Vertretern von sechs Hochschulen aus NRW gebildet, das im vergangenen Jahr zahlreiche Fortschritte erzielen konnte. Im Folgenden sollen die Ziele und Erkenntnisse der bisherigen Arbeit des Fachteams dargestellt werden.

Forschungsdatenmanagement in NRW 2015/2016

Im Kooperationsverbund der Digitalen Hochschule NRW (ehemals DV-ISA, Arbeitskreis DV-Infrastruktur der Hochschulen in NRW) haben sich 33 Hochschulen des Landes Nordrhein-Westfalen zusammengeschlossen, um gemeinsam mit dem Ministerium für Innovation, Wissenschaft und Forschung (MIWF) des Landes Nordrhein-Westfalen in den Themenbereichen Information, Kommunikation und Medien (IKM) zusammenzuarbeiten und digitale Initiativen weiter voranzutreiben.

Aufbauend auf der Erfahrung in der hochschulübergreifenden Kooperation zu IKM-Themen, ist das Thema Forschungsdatenmanagement im Frühjahr 2015 von DH-NRW (damals noch DV-ISA) aufgegriffen worden, im Wesentlichen motiviert durch die HRK-Empfehlungen (Hochschulkonferenz 2015). So wurde 2015 eine Arbeitsgruppe „Vorstudie FDM“ (AG Vorstudie) ins Leben gerufen, die eine Bestandsaufnahme zum Forschungsdatenmanagement an den Hochschulen in NRW durchgeführt hat (DV-ISA 2016). Wie auch in ganz Deutschland, gab es 2015 nur einzelne Hochschulen in NRW, die bereits aktive Schritte in Richtung FDM gemacht hatten. Universitäten wie Münster und Aachen hatten ihre Bedarfe durch entsprechende Umfragen (zum Teil angelehnt an die Umfragen der HU Berlin, diese wiederum an Umfragen britischer Universitäten angelehnt (Simukovic, Kindling und Schirmbacher 2013)) geklärt und Vorprojekte initiiert. Im Fokus der Vorstudie standen also eher erfassende als konzeptionelle Aktivitäten.

Eine der Erkenntnisse bereits nach der Vorstudie war, dass die Unterstützung der landesweiten Einführung von FDM an allen Hochschulen ein Spagat zwischen den beiden Polen ‘eine Hochschule hat die Wichtigkeit von FDM und/oder ihre Rolle darin noch nicht erkannt’ und ‘eine Hochschule hat bereits begonnen, Unterstützungsaktivitäten für ihre WissenschaftlerInnen zu entwickeln’ ist.

Roadmap 2016

Aufbauend auf den Handlungsempfehlungen der Vorstudie hat DH-NRW eine Roadmap für das Thema Forschungsdatenmanagement an den Hochschulen in NRW für 2016 skizziert. Als wichtigste Punkte können aufgeführt werden:

- DH-NRW etabliert sich als zentraler Ansprechpartner zum Thema FDM für die Hochschulen in NRW, es sollen daher
  - ein nachhaltiges FDM-Angebot und die dazugehörigen Koordinierungsaktivitäten erarbeitet werden,
  - die Kooperation und Abstimmung der Hochschulen in NRW zusammen mit anderen Akteuren sichergestellt werden,
  - Empfehlungen und Lösungen, die in einer nationalen Gesamtstrategie erfolgreich eingebracht werden können, erarbeitet werden.
- Ende 2016 soll ein Konzept für nachhaltiges FDM an den Hochschulen in NRW vorliegen.

Ressourcen

Für das Umsetzen der Roadmap wurde das sogenannte Fachteam Forschungs-datenmanagement gebildet, eine Expertengruppe zusammengesetzt aus Vertretern von Bibliotheken und Rechenzentren von NRW Hochschulen.[1] Für die Leitung des Fachteams wurden Personalmittel (½ Vollzeitäquivalenz, wiss. MitarbeiterIn) seitens von DH-NRW zur Verfügung gestellt, was zu einer entsprechenden Abordnung einer Mitarbeiterin der Universitätsbibliothek der Universität Duisburg-Essen führte. Alle weiteren Mitarbeiter des Fachteams wurden von ihren Einrichtungen freigestellt und haben die Arbeit im Fachteam neben ihrer einrichtungsinternen Arbeit wahrgenommen.

Zusätzlich wurden Sachmittel zur Verfügung gestellt, in Form von Reisekosten, Catering für Veranstaltungen und Marketingmaterial. Organisatorisch unterstützt wurde die Arbeit des Fachteams durch die Geschäftsstelle der DH-NRW, inhaltlich angebunden wurden die Arbeitsergebnisse an die Arbeit des „Kernteams“ der DH-NRW, welches aus gewählten Vertretern der Arbeitsgemeinschaft der IKM-Verantwortlichen (z.B. CIOs) der Hochschulen von NRW besteht.

Lösungsansätze

Ziel der Arbeit des Fachteams war und ist es, die Hochschulen in NRW für das Thema FDM zu sensibilisieren und damit hochschulübergreifende Kooperationen für die Entwicklung von Verfahren zu institutionellem FDM zu fördern. Gleichzeitig sollen Konzepte und Handlungsempfehlungen für die Etablierung von flächendeckendem und nachhaltigem FDM an den Hochschulen des Landes erarbeitet werden.

1 die auch AutorInnen dieses Beitrags sind.

Aufbauend auf der oben skizzierten Roadmap für 2016, wurde diese vom Fachteam in die drei Aktionsebenen unterteilt:

1. Kommunikation:
   Die wichtigsten Stakeholder sollten für das Thema sensibilisiert werden: hochschulpolitische Gremien, fachliche Experten und WissenschaftlerInnen.
2. Zentrales Informationsangebot:
   Geplant wurde eine Webseite, die die Aktivitäten zu FDM (technisch und organisatorisch/administrativ) an Hochschulen in NRW und somit eine Fortführung der Vorstudie von 2015 darstellen sollte. Zusätzlich sollte ein Informationspaket oder sog. Werkzeugkoffer erarbeitet werden, der Materialien zur Nachnutzung auf strategischer und operativer Ebene zur freien Nachnutzung für die Hochschulen bereitstellt.
3. Operationalisierung der HRK-Empfehlungen:
   Die HRK-Empfehlungen zu institutionellem Forschungsdatenmanagement sollten auf Prozessebene heruntergebrochen werden, und es sollte herausgearbeitet werden, welche Möglichkeiten es für kooperative Lösungen zu FDM gibt, und u.a. die Frage beantwortet werden, welche Services gebraucht und welche umgesetzt werden sollten.

Umsetzung

Zu den drei identifizierten Aktionsebenen fanden 2016 mehrere Aktivitäten statt, die nachfolgend beschrieben werden:

Kommunikation

In 2016 fanden Motivationsvorträge seitens des Fachteams auf Sitzungen verschiedener Gremien statt, damit wurden folgende Stakeholder erreicht: Prorektoren Forschung der Universitäten NRWs, Vizepräsidenten Forschung der Fachhochschulen NRWs, IKM-Verantwortliche (z.B. CIOs) der Hochschulen NRWs, Leiter der Rechenzentren der Universitäten NRWs (ARNW) und Leiter der Universitätsbibliotheken NRWs (AGUB).

Im Mai 2016 wurde die Veranstaltung „Jour Fixe“ als monatlich wiederkehrendes Format ins Leben gerufen und vom Fachteam organisiert. An der Universität Duisburg-Essen (da aufgrund der zentralen Lage gut erreichbar) wurden für jeweils zwei Stunden Vertreter der Hochschulen eingeladen, um sich zu einem informell gehaltenen Austausch zu treffen. Dies geschah zum Termin vor Ort oder aber virtuell über Adobe Connect. Es wurden jeweils 1-2 Vorträge zu verschiedenen Facetten des Themas Forschungsdatenmanagement angeboten. Dies waren Erfahrungsberichte von Vertretern der Hochschulen (seitens der zentralen Einrichtungen mit institutionellem Blick, oder fachlich motivierte Best-Practice-Beispiele von WissenschaftlerInnen) oder aber Vorstellungen von aktuellen Projekten (RADAR, DMPOnline, GFBio, etc.). Abgerundet wurde jede Veranstaltung durch viel Raum für Diskussion und Erfahrungsaustausch.

Angeboten wurde auch, durch Impulsvorträge seitens des Fachteams das Thema Forschungsdatenmanagement an den einzelnen Hochschulen zu positionieren. Dies wurde beispielsweise

durch Einladung zu einem allgemeinen Kick-Off-Workshop an der FH Bielefeld in Anspruch genommen. Angefragt wurde ein ähnliches Format für die TH Köln.

Über allgemein einführende Vorträge zum Thema FDM hinaus, wurde ein Impulsvortrag innerhalb des Kick-Off-Workshops des INF-Projekts des SFB "Medien der Kooperation" an der Universität Siegen eingebracht.

Zentrales Informationsangebot

Eine Webseite mit Hintergrundinformation zu den Aktivitäten des Fachteams wurde im Rahmen des DH-NRW Webauftrittes eingerichtet (Digitale Hochschule NRW 2017). Auf eine Fortführung der Auflistung der Aktivitäten aus der Vorstudie innerhalb dieser Webseite wurde verzichtet. Dies zum einen da die Ausrichtung und Zielgruppe eines solchen Inhalts fraglich war, zum anderen weil dem Fachteam bekannt war, dass entsprechende andere allgemeine Informationskanäle zum Thema FDM in Erarbeitung waren (Forschungsdaten.info 2016) oder bereits existierten und weiter erarbeitet werden (Forschungsdaten.org 2016), deren Nachnutzung sinnvoller erschien. So decken beide Angebote einerseits die Bedarfe nach allgemeiner Information von WissenschaftlerInnen ab, bzw. werden perspektivisch als Wiki seitens einer nationalen Arbeitsgruppe[2] für das Auflisten von Aktivitäten, Angeboten und AnsprechpartnerInnen zu FDM ausgebaut.

Als erster Teil des sog. Werkzeugkoffers wurden Folien zur Einführung in das Thema FDM zur freien Nachnutzung und Veränderung (CC-0-Lizenz) bereitgestellt (Curdt et al. 2016). Diese sind für Vertreter der Hochschulen gedacht, die das Thema institutionell platzieren möchten, aber sich (noch) nicht sehr intensiv mit dem Thema auseinandergesetzt haben. Durch entsprechende Notizen werden die Folien erklärt.

Geplant ist derzeit die Entwicklung eines Motivationsfilms, der in das Thema FDM einführt und sich an (Nachwuchs-)WissenschaftlerInnen richtet.

Operationalisierung der HRK-Empfehlungen

Der Austausch untereinander und mit den VertreterInnen anderer Hochschulen erhöhte innerhalb des Fachteams die Kenntnis hinsichtlich der in NRW bereits existierenden Infrastrukturangebote im FDM, die in Zukunft gegebenenfalls zur NRW-weiten Nutzung ausgebaut werden können. Hier stieg auch die Transparenz hinsichtlich geplanter Infrastrukturaktivitäten im Bereich Speicherstrukturen. So fanden im September entsprechende themenspezifische Workshops zwischen Vertretern der Hochschul-rechenzentren in Köln und Essen statt, bei denen die den Aktivitäten der Hochschulen in NRW zugrundeliegenden Konzepte vorgestellt wurden. Hier wurde festgestellt, dass im Zusammenwirken der drei vorliegenden Konzepte ein großer Schritt in Richtung einer gemeinsamen technischen Infrastruktur für das FDM für die Hochschulen im Land NRW als Teil der NFDI[3] gelingen kann.

---

2 Gemeint ist die im Oktober gegründete Unter-AG "Gelbe Seiten" der DINI-nestor-AG Forschungsdaten, s.a.http://www.forschungsdaten.org/index.php/DINI-nestor-WS6#Diskussion_der_Gruppe_.E2.80.9CGelbe_Seiten_Forschungsdatenmanagement.E2.80.9D.2C_Netzwerke bzw. http://www.forschungsdaten.org/index.php/UAG_Gelbe_Seiten_der_AG_Forschungsdaten, Letzter Zugriff: 29.12.2016.

3 NFDI - Nationale Forschungsdateninfrastruktur, ein vom Rat für Informationsinfrastrukturen vorgeschlagenes Konzept für eine national verteilte Forschungsdateninfrastruktur (RfII 2016).

Darüber hinaus wurde seitens des Fachteams ein eintägiger Workshop zum Thema HRK-Empfehlungen zu FDM für Vertreter der Hochschulen NRWs durchgeführt. Herr Meyer-Doerpinghaus als Vertreter der HRK hat dabei die wesentlichen Punkte der HRK-Empfehlungen für die anwesenden Hochschulen erläutert, in anschließender Gruppenarbeit wurden Punkte zum idealtypischen Stufenplan zur Einführung von FDM an Hochschulen erarbeitet. Ziel der Veranstaltung war insbesondere die zuvor vom Fachteam erarbeiteten Handlungsempfehlungen, die die Prozessebene in der Umsetzung der HRK-Empfehlungen adressieren, mit den TeilnehmerInnen zu diskutieren. Es sollte dabei geklärt werden, was den Hochschulen fehlt, um institutionelle FDM-Services und technische Infrastruktur umzusetzen.

Lessons Learned

Schon aus dem Vorprojekt 2015 war ersichtlich geworden, dass der Kenntnisstand zu FDM an den Hochschulen NRWs sehr unterschiedlich ist. Es war also wichtig, sowohl die Hochschulen, die sich noch nicht dem Thema FDM zugewandt haben, als auch die Hochschulen, die schon im Einführungsprozess waren, zu motivieren, ihre Ergebnisse auszutauschen und damit auch als „best practices“ für andere Hochschulen zur Verfügung zu stellen.

Das Fachteam 2016 wurde mit großen und auch forschungsstarken Universitäten (Aachen, Bielefeld, Duisburg-Essen, Köln, Münster und Siegen) besetzt, da diese bereits mit dem Thema FDM befasst waren. Diese Auswahl und die Größe der Gruppe erwiesen sich als einer der Erfolgsfaktoren. Die hier geübte Transparenz hatte sehr positive Auswirkungen auf die gemeinsame Arbeit. So konnte erreicht werden, dass im Herbst 2016 geplante Speicherinfrastrukturprojekte, an denen die Hochschulen des Fachteams beteiligt waren, transparent vorgestellt und in der Folge auch abgestimmt wurden.

Die monatlichen Fachteam-Treffen wurde durch öffentliche Jour Fixe FDM Veranstaltungen am Nachmittag ergänzt. Hier wurde durch Impulsvorträge eines der vielen Themen von FDM vorgestellt und diskutiert. Dieses Format erlebt durchweg eine positive Resonanz und wird auch 2017 weitergeführt. Im Jahr 2016 haben regelmäßig 20-30 TeilnehmerInnen (vor Ort oder virtuell) an den 7 Treffen teilgenommen. So wurden 19 Hochschulen NRWs erreicht. Das entstandene Kommunikationsformat ist einmalig in der Zusammensetzung aus Vertretern von Bibliotheken, Rechenzentren und Forschungsreferenten über Hochschulgrenzen hinweg.

Die größte Herausforderung der Roadmap für 2016 war die Erstellung einer Handlungsempfehlung als Konzept für die nachhaltige Umsetzung von FDM in NRW (Operationalisierung der HRK-Empfehlungen). Diese hat sich als zu anspruchsvoll erwiesen. Es hat sich gezeigt, dass die Ausgangslage, Bedürfnisse und Interessen der einzelnen Hochschulen in Bezug auf Forschungsdatenmanagement sehr heterogen sind. So ist ein wesentlicher Unterschied zwischen Universitäten und Fachhochschulen die IT-Ausstattung und die damit einhergehende Unterstützung des Forschungsprozesses. Gleichzeitig sind die Bedürfnisse einzelner Hochschulen je nach fachlichen Forschungsschwerpunkten sehr heterogen. Während es für einzelne Fachdisziplinen bereits in den wissenschaftlichen Communities etablierte Lösungen gibt, die von Hochschulen nachgenutzt werden, besteht für einige Fächer ein hoher Bedarf an Entwicklungsarbeit für Lösungen rund um das Handling der Daten.

Gleichzeitig haben sich die Möglichkeiten und Dienste rund um institutionelles FDM weiterentwickelt und der Bedarf an Unterstützung in der Umsetzung ist erheblich gestiegen. Als Folge

daraus wurden seitens des Fachteams die Arbeiten an der Handlungsempfehlung nicht weiter verfolgt. Stattdessen wurden konkrete Aktivitäten, die hochschulübergreifend als landesweites Projekt umgesetzt werden können, beschrieben.

Zusammenfassung und Ausblick

Das Thema Forschungsdatenmanagement (FDM) hat in den letzten Jahren wesentlich an Bedeutung gewonnen. Einrichtungen und Organisationen wie beispielsweise die DFG, die HRK oder der RfII haben Empfehlungen für den Umgang mit Forschungsdaten erlassen und den Aufbau von nachhaltigen Infrastrukturlösungen gefordert. Folglich stehen auch Hochschulen allein oder in Kooperation in der Verantwortung ein professionelles FDM zu schaffen und aufzubauen.

In diesem Beitrag wurden Ziele, Umsetzung und Erkenntnisse des Fachteams FDM dargestellt, dass vom DV-ISA (heute DH-NRW) im Frühjahr 2016 eingerichtet wurde, bestehend aus Vertretern von Rechenzentren und Bibliotheken aus NRW Hochschulen. Entsprechend der skizzierten Roadmap für 2016 wurden seitens des Fachteams verschiedenen Maßnahmen etabliert (z.B. monatlicher Jour Fixe, Workshop) zur Sensibilisierung der NRW Hochschulen zum Thema FDM, sowie zur Schaffung von hochschulübergreifenden Kooperationen. Ergänzend wurden seitens des Fachteams Informationsmaterialen zur Nachnutzung durch Hochschulen erarbeitet.

Der Umsetzungsstatus und das Bewusstsein für die Wichtigkeit von FDM sind an den verschiedenen NRW Hochschulen auch nach Durchführung der unterschiedlichen FDM Aktivitäten seitens des Fachteams sehr groß. Während es an einigen Hochschulstandorten bereits etablierte Leitlinien zum Umgang mit Forschungsdaten (z.B. RWTH Aachen, Universität Bielefeld, HHU Düsseldorf, BU Wuppertal) und es entsprechende zentrale Unterstützungsangebote für die Wissenschaftler gibt (z.B. RWTH Aachen, Universität Bielefeld), muss an anderen Standorten noch Überzeugungsarbeit für die Relevanz von FDM geleistet werden.

Darüber hinaus gibt es in NRW einige hochschulübergreifende Kooperationen, die einzelne Bausteine für flächendeckende Angebote für nachhaltiges FDM an alle Hochschulen des Landes liefern können. Dazu gehört die Kollaborationsplattform Sciebo (Owncloud-basiert, Nutzung durch 25 Hochschulen in NRW), die TSM-Archivierungssoftware (Konsortialvertrag von 13 Hochschulen in NRW, Basis für die Entwicklung eines Archivierungsdienstes "SimpleArchive" an der RWTH Aachen), die Landeslizenz von Rosetta von ExLibris (Werkzeug zur Langzeitarchivierung publikationsnaher Forschungsdaten), Projektantrag RD-Storage (Speicherung von Daten basierend auf Object-Store-Technologie, Konsortium von 6 Hochschulen), Pilotprojekt zur standortübergreifenden Archivierung von Forschungsdaten (Kooperation der Universitäten Düsseldorf, Siegen und Wuppertal) und das Pilotprojekt für eine kollaborative Speicher- und Service-Infrastruktur auf Basis von quelloffener freier Software (Projektantrag in einem Konsortium von 5 Hochschulen).

Auch nach Ende der skizzierten Roadmap für das Fachteam bis März 2017 wird das Thema FDM weiterhin eine wichtige Rolle in der Strategie der DH-NRW spielen. Daher ist derzeit ein dreijähriges Projekt in Planung, in dem ein Unterstützungsangebot für die NRW Hochschulen erarbeitet werden soll. Dieses soll u.a. Musterrichtlinien zum Umgang mit Forschungsdaten und eine Handreichung zu rechtlichen Fragen im Umgang mit Forschungsdaten umfassen. Des Weiteren ist geplant eine Übersicht von „best practices “ zu Policy-Enforcement zusammenzustellen und Ergebnisse aus deutschlandweiten hochschul- und landesweiten Bedarfsumfragen aufzuberei-

ten. Weiterhin ist der Aufbau einer zentralen Anlaufstelle für Detail- und Beratungsfragen, sowie Beratung und Unterstützung an den Standorten und der Aufbau eines zentralen Angebots für verschiedene Tools (z.B. Datenmanagementpläne und Metadaten) für NRW-Hochschulen geplant. Darüber hinaus sollen die verschiedenen schon existierenden oder sich im Aufbau befindenden hochschulübergreifenden Initiativen im Hinblick auf die technische Umsetzung von FDM NRW-weit koordiniert und begleitet werden, sowie Awareness- und Sensibilisierungsaktivitäten weiterhin fortgeführt werden (z.B. regelmäßiger Jour Fixe und ganztägige Workshops). In der bisherigen Planung soll der bottom-up-Charakter der Fachteam-Arbeit erhalten bleiben, personell unterstützt durch Projektmitarbeiter.

Die Einführung von FDM wird sowohl in der Veröffentlichung des RfII als auch in den Aussagen der HRK als nationale Aufgabe adressiert.[4] Wissenschaft passiert ohne institutionelle oder Ländergrenzen. Kooperative Forschungsstrukturen nicht nur bundesland- sondern auch staatenübergreifend sind gewollt und notwendig geworden. Daher ist FDM eigentlich kein Thema für einzelne Bundesländer, bisherige Initiativen und Aktivitäten ergeben sich aber aus der föderalen Struktur in Deutschland. Da bereits regionale Aktivitäten existieren, erscheint es in NRW daher sinnvoll, diese vorhandenen Strukturen und die räumliche Nähe zu nutzen, um kooperativ an institutionellem FDM zu arbeiten. Alle aktuellen und geplanten NRW-Aktivitäten werden in nationale und falls möglich internationale Initiativen (z.B. DINI/nestor AG Forschungsdaten, Research Data Alliance) eingebracht, um auch hier Synergieeffekte zu nutzen.

Anders als Baden-Württemberg und Hessen wird in NRW mit den bisherigen und geplanten Aktivitäten kein zentral initiiertes top-down FDM verfolgt. Stattdessen wird auf den bottom-up-Charakter der institutionalisierten Zusammenarbeit mehrerer Hochschulen innerhalb der DH-NRW gesetzt um somit Kompetenzen und bereits vorhandene Aktivitäten zu bündeln, um sie dann den anderen Hochschulen zur Verfügung zu stellen. Dabei ist der kontinuierliche und aktive Austausch zwischen Vertretern der verschiedenen Organisations- und Infrastruktureinrichtungen (Bibliotheken, Rechenzentren, Forschungsreferaten) über Hochschulgrenzen hinweg in dieser intensiven Form in Bezug auf FDM in Deutschland einmalig und für die inhaltliche Arbeit sehr bereichernd. Die aktive Einbindung und Beteiligung von Vertretern von Fachhochschulen bildet zurzeit in Deutschland eine positive Ausnahme.

Die bisherigen Aktivitäten adressierten bislang die Vertreter der Infrastruktureinrichtungen. Dies sehen die Autoren als guten Startpunkt. Wichtig ist nun, WissenschaftlerInnen bzw. Fachcommunities zu adressieren und zu involvieren. Austausch und Gespräche gab es bisher dort, wo natürliche Brücken zwischen Infrastruktureinrichtungen und WissenschaftlerInnen durch Forschungskooperationen (z.B. INF-Projekten) bestehen. Nur durch engen Austausch zwischen Infrastruktureinrichtungen und wissenschaftlichen Communities können nachhaltige Lösungen für FDM entstehen, die nicht „an den WissenschaftlerInnen vorbei“ entwickelt werden. Auch in Hinblick auf die vom RfII vorgeschlagene NFDI, die fachliche Schwerpunkte für das Entstehen von Kompetenz- und/oder Datenzentren skizziert, ist eine Fokussierung auf das Thema FDM mit seiner großen fachspezifischen Heterogenität unabdingbar.

4 Siehe dazu auch ganz aktuell die Pressemitteilung der HRK zum HRK-Workshops zur Zukunft des For schungsdatenmanagements in Bonn (Hochschulrektorenkonferenz 2016).

Literaturangaben

Curdt, Constanze, Krämer, Florian, Hess, Volker, Lopez, Ania, Magrean, Benedikt, Rudolph, Dominik, und Johanna Vompras. 2016. „Einführung in Forschungsdatenmanagement.“ Letzter Zugriff 29. Dezember 2016. doi: 10.5281/zenodo.165126.

Digitale Hochschule NRW, DH-NRW. „Forschungsdatenmanagement.“ Letzter Zugriff 27. März 2017. https://www.dh-nrw.de/handlungsfelder/forschung/forschungsdatenmanagement.

DV-ISA Arbeitskreis DV-Infrastruktur der Hochschulen in NRW. „Umgang mit digitalen Daten in der Wissenschaft: Forschungsdatenmanagement in NRW - Eine erste Bestandsaufnahme.“ Letzter Zugriff 29. Dezember 2016. doi: 10.5281/zenodo.200429.

Forschungsdaten.info. „Forschung und Daten managen.“ Letzter Zugriff 29. Dezember 2016. https://www.forschungsdaten.info.

Forschungsdaten.org. Letzter Zugriff 29. Dezember 2016. http://www.forschungsdaten.org/index.php/Hauptseite.

Hochschulrektorenkonferenz. 2015. „Wie Hochschulleitungen die Entwicklung des Forschungsdatenmanagements steuern können. Orientierungspfade, Handlungsoptionen, Szenarien, Empfehlung der 19. Mitgliederversammlung der HRK am 10. November 2015 in Kiel.“ Vorgestellt bei der 19. Mitgliederversammlung der HRK, Kiel, 10. November 2015. Letzter Zugriff 29. Dezember 2016. http://www.hrk.de/uploads/tx_szconvention/Empfehlung_Forschungsdatenmanagement__final_Stand_11.11.2015.pdf.

Hochschulrektorenkonferenz 2016. „Pressemitteilung. Forschungsdatenmanagement: Deutschland muss aufholen – Impulse von Bund und Ländern unverzichtbar.“ 19. Dezember 2016. https://www.hrk.de/presse/pressemitteilungen/pressemitteilung/?tx_ttnews[tt_news]=4092&cHash=e1e913c76f0305b63530804e33ff0a76.

RfII – Rat für Informationsinfrastrukturen. 2016. *Leistung aus Vielfalt - Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagementes in Deutschland*. Göttingen. Letzter Zugriff 29. Dezember 2016, http://www.rfii.de/?wpdmdl=1998.

Simukovic, Elena, Maxi Kindling, und Peter Schirmbacher. 2013. "Umfrage zum Umgang mit digitalen Forschungsdaten an der Humboldt-Universität zu Berlin.“ Letzter Zugriff 12. Dezember 2016. http://edoc.hu-berlin.de/docviews/abstract.php?id=40341.

Der Aufbau einer „Entity Collection“ der Forschungsleistung der TU Dortmund

Hans-Georg Becker[1], Kathrin Höhner[2]

1,2 Universitätsbibliothek der Technischen Universität Dortmund

Zusammenfassung. Mit einer systematischen Kuratierung der lokalen Daten will die Universitätsbibliothek das Ziel erreichen, die Forschungsleistung der Technischen Universität Dortmund möglichst vollständig sichtbar und nachnutzbar zu machen. Um Data Curation effizient zu gestalten, verfolgt die Universitätsbibliothek Dortmund ein umfassendes Gesamtkonzept. So werden ein Metadaten-Managementsystem (MMS) und eine Datenplattform entwickelt, die auf die Nachnutzung möglichst vieler bereits existierender und verlinkbarer Identifikatoren zielen. Zudem bietet das Metadaten-Managementsystem die Möglichkeit, unterschiedlichste Datentypen mit verschiedenen Relationen zu erfassen. Es ist daher für das Monitoring der an der Technischen Universität Dortmund produzierten und nachzuweisenden Forschungsdaten besonders geeignet. Auch wird es eine Synchronisation mit dem Katalog plus der Universitätsbibliothek geben.
Für die Wissenschaftlerinnen und Wissenschaftler wird es für die unterschiedlichen Anwendungsszenarien eine einzige Oberfläche als Einstieg geben. Hierüber können sie nicht nur Publikationen für die Hochschulbibliographie melden, Volltexte in das Repositorium hochladen oder Publikationslisten für Webseiten generieren, sondern auch finanzielle Unterstützung für Open Access-Publikationen beantragen.

Schlagwörter. Datenkuratierung, Metadatenmanagement

Aufbau einer “Entity collection“

Im Laufe des gesamten Forschungsprozesses werden Daten generiert. Deren Analyse und Interpretation erfolgt einerseits im Rahmen von klassischen Publikationen in wissenschaftlichen Zeitschriften oder Monographien, andererseits im Rahmen von (Zwischen)-Berichten, Software, Kunstwerken oder anderem. All dies stellt ebenso wie die erhaltenen Primärdaten den Forschungsoutput einer Universität dar.

Die zugehörigen Metadaten in strukturierter Form abzulegen, zu präsentieren und durchsuchbar zu machen, ist eine Herausforderung für Bibliotheken, die sie mit ihrer Kompetenz für die Strukturierung von Daten bereits jetzt annehmen. Klassischerweise wird das Publikationsaufkommen einer Universität in Fachzeitschriften und Monographien ermittelt und z. B. in Form einer Hochschulbibliographie oder als Publikationslisten für einzelne Wissenschaftlerinnen und Wissenschaftler präsentiert. Jahn und Horstmann dagegen haben bereits 2010 eine „disziplinsensitive Strukturierung der bibliographischen Information“ gefordert. So ist es auch das Anliegen der Universitätsbibliothek (UB) der TU Dortmund, die gesamten Forschungsergebnisse ihrer Wissenschaftlerinnen und Wissenschaftler bibliothekarisch aufzubereiten, sie zu präsentieren und dabei auch Hierarchien und Vernetzungen darstellen zu können. Ein Forschungsinformationssystem ist

an der TU Dortmund noch nicht vorhanden, weswegen Metadaten zu Projekten etc. nur unzureichend bekannt sind.

So entstand die Idee, eine Entity Collection aufzubauen, in der alle zur Darstellung des Forschungsoutputs der TU Dortmund benötigten Metadaten erfasst sind.

An der UB wurden bereits 2015 folgende Prinzipien des Metadatenmanagements aufgestellt, die als Leitlinien für den Aufbau der Entitäten-Kollektion dienen.

- Jeder Datensatz hat ein Mastersystem und ein Masterformat.
- Datensätze werden automatisch verteilt und nicht mehrfach erfasst.
- Linked Data ist ein Kernkonzept.
- Es werden so wenig Daten wie möglich und so viele wie nötig erfasst.
- Es gibt ein Repositorium für digitale Objekte.

Basierend auf diesen Prinzipien wurde im Laufe der letzten beiden Jahre zunächst gemeinsam mit der UB der Ruhr-Universität Bochum, später nur in Dortmund, ein Metadatenmanagementsystem (MMS) entwickelt, mit dessen Hilfe die Entitäten-Kollektion aufgebaut, angepasst aber auch deren Daten weiterverwendet werden können. Als erster Schritt werden die beiden bisher vollkommen voneinander unabhängigen Systeme Hochschulbibliographie und Repositorium für Volltexte und andere digitale Objekte durch das MMS verknüpft.

Datenmodell und Datenstruktur

Bei der Konzeption des Datenmodells mussten zwei unterschiedliche Ansätze berücksichtigt werden: einerseits müssen Metadaten zu Publikationen unterschiedlichster Art erfasst werden können, um verschiedene Zielsysteme von Webseiten bis zu (Fach)Repositorien bedienen zu können, andererseits waren die Daten der Personen und Organisationseinheiten einschließlich temporärer Projekte der Universitäten zu berücksichtigen, um Relationen darstellen zu können.

Dieses Datenmodell ist zugleich so flexibel, dass in Zukunft weitere benötigte Kategorien (z. B. Projektangaben) eingebaut werden können. Der u.a. vom Wissenschaftsrat und von der DFG für Forschungsinformationssysteme (FIS) empfohlene Metadatenstandard Cerif findet dabei Berücksichtigung.

So basieren die Entitätstypen auf denen des Cerif-Modells und auch die Relationen unter ihnen orientieren sich an diesem Modell. Für die optimale Darstellung des Forschungsoutputs ist es notwendig, Organisationseinheiten, Personen, Arbeitsgruppen und Projekte sowie unterschiedliche Publikationsformen abzubilden. Letztere unterscheiden sich im Cerif-Modell in Publikationen, Patente und Produkte. Dabei versteht man unter Produkten insbesondere nicht-textuelle Objekte.

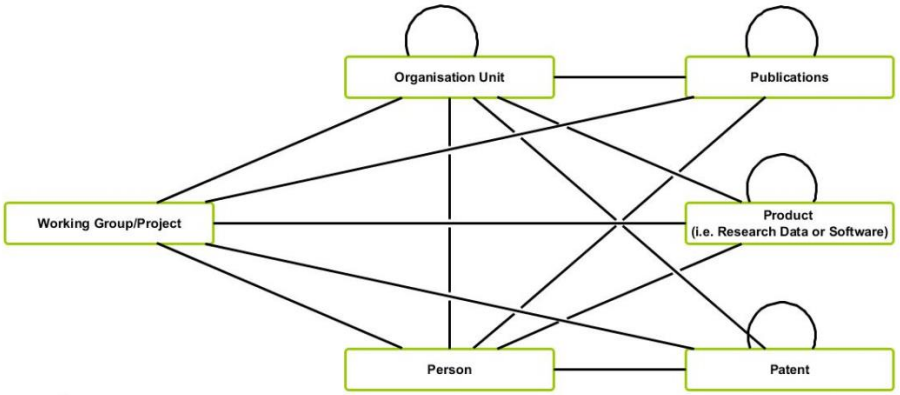


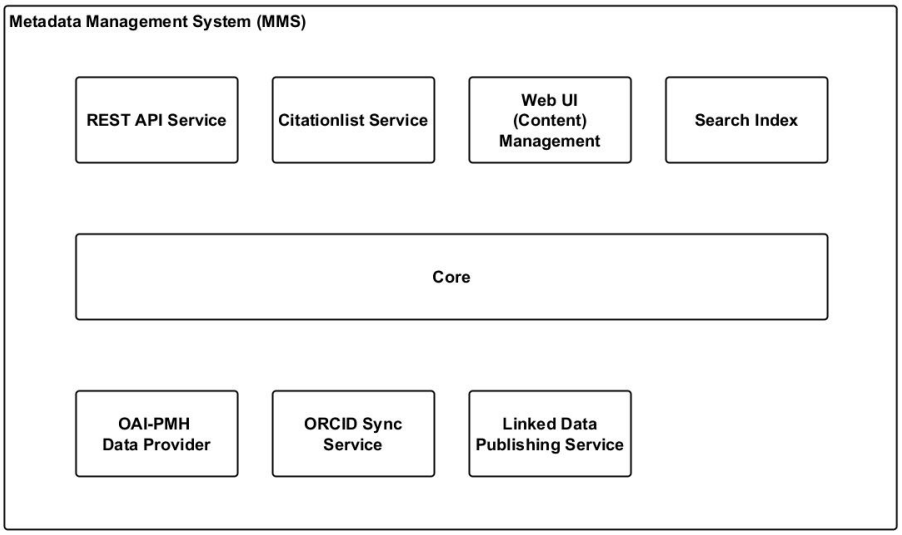
Abbildung 1.1. An CERIF orientiertes Datenmodell des MMS (Becker 2017)

Im MMS werden die drei Entitäten Publikationen, Patente und Produkte unter „Works“ zusammengefasst und mittels Typisierung durch kontrolliertes Vokabular unterscheidbar und Cerif-kompatibel gemacht.

Alle Entitäten bekommen im Datenmodell der UB Dortmund eindeutige Identifikatoren (IDs), wozu soweit wie möglich bereits vorhandene IDs nachgenutzt und möglichst wenige selbst generiert werden sollen. Für Personen werden z. B. die Open Researcher and Contributor ID (ORCID iD), für Personen, Organisationen und deren Untereinheiten die Gemeinsame Normdatei ID (GND ID) sowie DOIs oder andere persistente Identifikatoren als IDs für Werke unterschiedlicher Art verwendet. Insbesondere die Verwendung von Normdaten und kontrollierten Vokabularen erleichtert die Umsetzung des linked data-Prinzips.

Software-Architektur

Das MMS besteht aktuell aus sechs Services, die auf eine mit einem application programming interface (API) versehenen Persistenzschicht aufsetzen.



Abiildung 1.2 Architektur des Metadatenmanagementsystems (Becker 2017)

Der Service „Web UI (Content) Management“ liefert die Werkzeuge, um die unterschiedlichen Entitäten durch Metadaten zu beschreiben und mit anderen Entitäten zu verknüpfen. Überall dort, wo Verknüpfungen erzeugt werden, gibt es Assistenten, die aus den bereits im System befindlichen Entitäten Vorschläge unterbreiten. Eine Auswahl erzeugt dann eine Verknüpfung der IDs der Entitäten.

Auf Basis der im System erfassten Metadaten wird im Hintergrund ein Suchindex (Search Index) erzeugt. Dabei werden die durchsuchbaren Felder befüllt sowie der entstandene Datensatz als JavaScript Object Notation (JSON) abgelegt. Ferner werden hier die „einfache Suche“ sowie die Vorschlagsassistenten mit Daten befüllt. Auf den Suchindex kann direkt lesend zugegriffen werden; so lassen sich unterschiedliche Arten von Präsentationsanwendungen entwickeln. Auch die Webanwendung der Hochschulbibliographie der TU Dortmund verwendet diesen Index. Zusätzlich werden die Ergebnisse innerhalb der Hochschulbibliographie mittels schema.org im JSON-LD-Format (JSON for Linked Data) angereichert, was der Erhöhung der Sichtbarkeit dient.

Das MMS verfügt über ein Representational State Transfer application programming interface (REST API), mit dem die unterschiedlichen Entitätstypen angelegt, aktualisiert, gelöscht und ausgelesen werden können. Weitere Funktionen sind geplant, die es ermöglichen sollen, dass „Web UI (Content) Management“ nur noch über die REST-Schnittstelle auf die Daten zugreifen kann. Außerdem können damit ohne direkten Zugriff auf den Suchindex „Retrieval“-Anwendungen implementiert werden Die REST-Schnittstelle ist auch der Service, der die Integration in weitere Systemarchitekturen ermöglichen soll.

Der Publikationslisten-Service (Citationlist Service) liefert Zitationslisten von Personen, Organisationseinheiten oder Arbeitsgruppen und Projekten in frei wählbaren Zitationsstilen und Konfigurationen z. B. für die Verwendung in Webseiten.

Für die Nachnutzung der Daten wird neben dem Index auch ein Open Archives-Initiative Protocol for Metadata Harvesting (OAI-PMH) Data Provider zur Verfügung gestellt. Über diesen Service werden z. B. die für OpenAIRE notwendigen Daten zum Harvesting bereitgestellt, so dass

die Publikationen sichtbarer werden, die unter Verantwortung von Angehörigen der TU Dortmund entstanden sind.

Auch die Synchronisation der Daten mit der ORCID-Plattform erhöht die Sichtbarkeit der Daten. Die Synchronisation läuft hierbei in beide Richtungen: vorausgesetzt, eine Wissenschaftlerin / ein Wissenschaftler hat den Zugriff erlaubt, werden auf der einen Seite Daten aus der ORCID-Plattform in die Hochschulbibliographie übernommen, auf der anderen Seite neue oder angereicherte Daten aus dem MMS in die ORCID-Plattform übertragen.. Dabei werden verschiedene Mechanismen zur Vermeidung von Duplikaten angewendet.

Der Linked Data Publishing Service wird die Daten aus dem MMS als an die Cerif-Ontologie angelehnte Resource Description Framework (RDF)-Daten in die Linked Open Data Plattform der UB Dortmund publizieren.

Implementiert ist das System in der Programmiersprache Python unter Verwendung des Webframeworks Flask. Für den Service „Web UI (Content) Management“ wird zusätzlich das WTForms-Framework für die Definition von Webformularen verwendet. Dieses erlaubt es, objektorientiert passende Formulare generieren zu lassen, sowie deren Inhalte ohne weiteren Aufwand als JSON-Daten nachzunutzen.

Für die Persistenz wird ein Apache Solr-Index verwendet, in dem neben den suchbaren Feldern auch der vollständige Datensatz als JSON enthalten ist.

Zielsysteme

Das im ersten Abschnitt beschriebene MMS dient als zentrale „Datendrehscheibe“ für unterschiedliche Zielsysteme, wie in Abb. 2.1 gezeigt wird. Gerade vor dem Hintergrund, dass eine Hochschulbibliographie oder ein universitäres Repositorium von Wissenschaftlerinnen und Wissenschaftlern wenig als primäre Rechercheinstrumente verwendet werden, wird die Sichtbarkeit der Metadaten insbesondere durch Implementierung verschiedener Schnittstellen zur möglichst breiten Bedienung unterschiedlicher Zielsysteme erhöht.

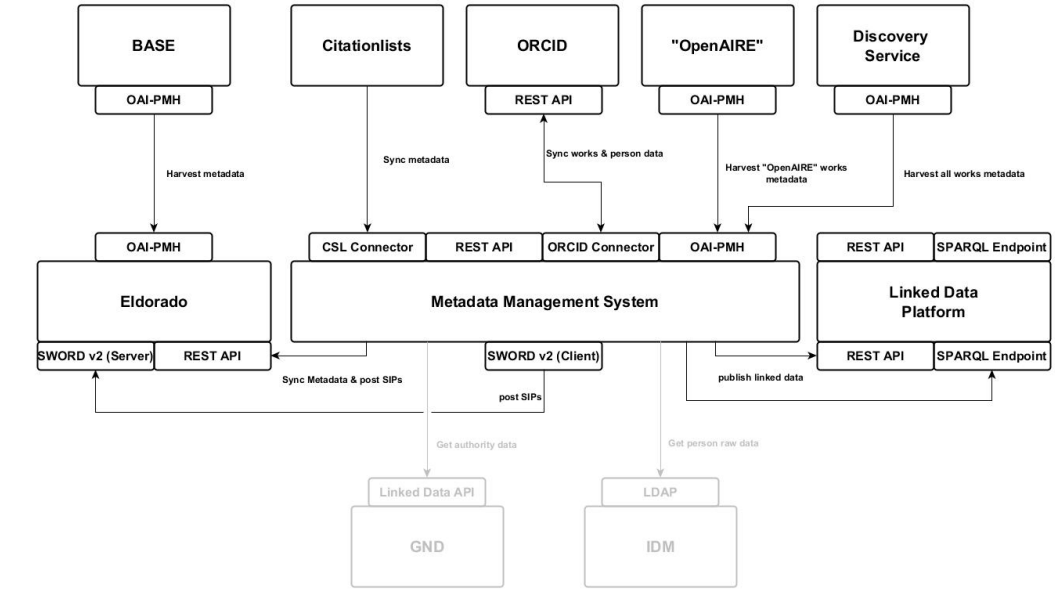


Abbildung 2.1. Datenfluss zwischen den Systemen an der UB Dortmund (Becker 2017)

Quellen für das MMS sind ORCID, die GND und für Personendaten das Identity-Managementsystem der TU Dortmund. Für die initiale Verknüpfung des TU-eigenen Repositoriums „Eldorado" mit dem MMS dient dieses ebenfalls als Quelle.

„Eldorado" und die ORCID-Plattform sind gleichzeitig auch Zielsysteme des MMS. Weitere Zielsysteme sind die Webseiten von Angehörigen der TU Dortmund, der Katalog plus der TU Dortmund (in der Grafik als Discovery Service bezeichnet) sowie OpenAIRE als Portal für die Forschungsergebnisse EU-geförderter Projekte.

Die meisten Zielsysteme bedienen die etablierten Schnittstellen REST, OAI-PMH oder SWORD (Simple Web-service Offering Repository Deposit). Für ORCID dagegen musste ein „Connector" geschrieben werden. Da das MMS unterschiedliche Schnittstellen bedient, ist ein Szenario der Datenlieferung an Fach-Repositorien ebenfalls umsetzbar.

Die verschiedenen Zielsysteme werden im Folgenden detaillierter beschrieben. Dabei werden BASE und OpenAIRE nicht weiter berücksichtigt, da die Datenlieferung an diese beiden Plattformen ausschließlich der Erhöhung der Sichtbarkeit der Forschungsergebnisse der TU Dortmund dient und keine TU-Dortmund- spezifischen Anwendungen darstellen.

Publikationslisten

Die Darstellung der Forschungsleistung der Angehörigen der TU Dortmund soll auf unterschiedlichen Wegen erfolgen. Ein wichtiges Ziel ist die Generierung strukturierter Publikationslisten („Citation lists") für die Webseiten der einzelnen Forscherinnen und Forscher, von Lehrstühlen, Instituten oder auch ganzen Fakultäten. Die individuelle Strukturierung der Listen ermöglicht eine „hochgradig kondensierte Darstellungsweise wissenschaftlicher Erkenntnis" und trägt damit zur Optimierung der Visualisierung der eigenen Forschungsergebnisse bei. Denn nur wenn Forschende einen Mehrwert in der Meldung der Publikationen an die Universitätsbibliothek erkennen,

werden sie bereit sein, ihre Verlags-Publikationen ebenso wie ihre Forschungsarbeiten und grauen Publikationen der UB zu übermitteln. Für einzelne Projekte oder Arbeitsgruppen ist es ebenso notwendig, spezifische Publikationslisten generieren zu können - sei es für die Präsentation auf der Webseite oder für Förderanträge und Berichte zu Förderanträgen, wie z. B. für Sonderforschungsbereiche. Aus diesem Grunde erfasst die UB im Rahmen ihres MMS auch Arbeitsgruppen und Projekte als eigene Entitäten.

Hochschulbibliographie

Die Hochschulbibliographie dient als zentrales Nachweisinstrument für die an der TU Dortmund entstandenen Publikationen. Darüber hinaus werden hier auch andere Forschungsleistungen, wie z. B. Patente, Vorträge oder auch Projektberichte erfasst. Die Datenlieferung an die Hochschulbibliographie beruht auf freiwilliger Basis. Um von Wissenschaftlerinnen und Wissenschaftlern akzeptiert zu werden, müssen die Inhalte attraktiv dargestellt und ein niedrigschwelliger Einstieg der Publikationsmeldung ermöglicht werden, der sich an unterschiedlichen Bedürfnissen der Forschenden orientiert. Zum einen können Forschende ihre Publikationen selbst manuell eintragen, zum anderen wird es möglich sein, Listen aus Literaturverwaltungsprogrammen im BibTex- oder RIS-Format zu importieren oder auch ganz klassisch Word- oder Excel-Listen an die UB melden. Letztere werden von den Mitarbeitern der UB manuell in die Hochschulbibliographie eingetragen.

Gerade für Wissenschaftlerinnen und Wissenschaftler der Geistes-, Kunst- und Kulturwissenschaften, die in den klassischen Abstract- und Zitationsdatenbanken häufig unterrepräsentiert sind, bieten die Hochschulbibliographie oder auch die unter 2.1 beschriebenen Publikationslisten auf ihren Webseiten eine gute Möglichkeit, ihre Forschungsleistung zu präsentieren.

Für Forschende aller Disziplinen bietet die Hochschulbibliographie die Möglichkeit, all jene Forschungsergebnisse zu präsentieren, die nicht klassische Text-Publikationen sind, z. B Software oder auf Repositorien publizierte Forschungsdaten.

Repositorium - Eldorado

Mit dem seit Mitte 1997 betriebenen Repositorium „Eldorado“, das auf der Software DSpace basiert, bietet die UB Dortmund eine Plattform, um Dokumente sowohl primär als auch als Zweitveröffentlichung zu publizieren. Dabei spielt „Eldorado“ nicht nur für die digitale Veröffentlichung von Dissertationen und klassischer grauer Literatur, wie z. B. Zwischen-, Jahres- oder Abschlussberichte, sondern auch für Dokumente der Universitätsverwaltung eine wichtige Rolle. Auch Periodika und Schriftenreihen werden genuin hier veröffentlicht, z. B. die Projektberichte der Fakultät für Raumplanung oder die Preprints der Fakultät für Mathematik. „Eldorado“ ist von der Deutschen Initiative für Netzwerkinformation zertifiziert worden und bietet alle notwendigen Schnittstellen, um dort abgelegte Dokumente dauerhaft zur Verfügung zu stellen und eine Indexierung in wissenschaftlich relevanten Suchmaschinen zu ermöglichen. Zudem werden als persistente Identifikatoren DOIs vergeben. Die Anbindung des Repositoriums an das MMS ermöglicht zum einen eine intuitivere Bedienung für Wissenschaftlerinnen und Wissenschaftler und dient zum anderen der Verbesserung der Sichtbarkeit der Publikationen des Repositoriums, da das MMS umfassendere Metadaten erlaubt und insbesondere Verlinkungen, z. B. mit der ORCID-iD

ermöglicht. Die Verwendung von schema.org im MMS und damit auch für die Repositoriums-Daten trägt ebenso zur Erhöhung der Sichtbarkeit und damit zur Optimierung des Repositoriums bei. In den Empfehlungen der „Repositories Early Adopters Expert Group“, werden fünf „required“ Aspekte genannt, von denen die TU Dortmund vier erfüllt; einzig die Forderung 4 nach unterschiedlichen Granularitätslevels für persiste Identifikatoren kann bisher nicht erfüllt werden. Mit der Anbindung des Repositoriums an das MMS werden zudem die drei „recommended“ Aspekte erfüllt - so liefert das MMS hinreichende Metadaten für eine gute Zitation, die Sichtbarkeit wird mittels schema.org erhöht und HTML Meta Data Tags werden bereits verwendet.

ORCID

ORCID ist für die Bibliothek der TU Dortmund ein attraktives System, weil es über die eindeutige Verknüpfung von Personen mit ihren Publikationen vollständigere Aussagen über die Publikationstätigkeiten der Angehörigen der eigenen Universität ermöglicht. Gleichzeitig strebt die TU Dortmund an, dass ihre Wissenschaftlerinnen und Wissenschaftler die Institution „Technische Universität Dortmund“ in normierter, international anerkannter Form bezeichnen. Dazu hat die TU Dortmund bereits 2016 eine Mitgliedschaft bei ORCID abgeschlossen, um sich von ihren Wissenschaftlerinnen und Wissenschaftlern sowohl das Recht einräumen zu lassen, Daten aus dem Profil zu lesen, als auch Publikationsdaten sowie die Bezeichnung der TU Dortmund in deren Profile einzuspielen. Hieraus ergeben sich Vorteile alle Beteiligten: Daten, die aus ORCID in die Hochschulbibliographie eingespielt werden, fließen hierüber in die Publikationslisten für die Webseiten ein. Damit wird den Wissenschaftlerinnen und Wissenschaftlern das Melden ihrer Publikationen erspart, sofern sie beim Einreichen jeder einzelnen Publikation ihre ORCID ID mit angeben. Die meisten internationalen Verlage verknüpfen die ORCID ID mit dem DOI der Publikation. Wurde zuvor den Organisationen DataCite und Crossref ein „auto-update“ erlaubt, werden alle Publikationen, die einen DOI als Identifikator bekommen, automatisch in das ORCID-Profil eingespielt. Für die Bibliothek ergibt sich der Vorteil einer vollständigeren Hochschulbibliographie. Die Mitarbeiter der UB bereiten die von ORCID gelieferten Daten bibliothekarisch auf und spielen sie in das ORCID-Profil zurück.

Katalog plus (Discovery Service)

Der Katalog plus der UB Dortmund ist das primäre Suchinstrument der Bibliothek. Neben dem Nachweis des eigenen analogen und digitalen Bestandes sollen alle Publikationen von Angehörigen der TU Dortmund hier zu finden sein - unabhängig davon, ob sie analog, digital oder gar nicht bereitgestellt werden können. Daher ist geplant, die Daten des MMS mittels OAI-PMH in den Katalog plus zu transferieren.

Ausblick

Das Potential des MMS, unterschiedliche Datentypen mit verschiedenen Relationen zu erfassen, ist insbesondere für das Monitoring der an der TU Dortmund produzierten und nachzuweisenden

Forschungsdaten geeignet - unabhängig von deren Publikation und gegebenenfalls deren Publikationsort.

An der TU Dortmund wurde im Frühjahr 2016 unter den Forschenden der TU Dortmund abgefragt, ob und in welcher Form Unterstützung beim FDM gewünscht wird. Basierend auf den Ergebnissen dieser Bedarfsabfrage wurde eine OAIS-konforme (Open Archival Information System) Architektur entwickelt, die sich im Wesentlichen aus bereits existierenden Softwarekomponenten zusammensetzt und die die maximale Integration und Nachnutzung bereits etablierter Systeme zum Ziel hat.

Das hier vorgestellte MMS kann als Content-Management-System eingesetzt werden, wenn Forschungsdaten, die in beliebigen Arbeitsumgebungen der Forschenden verwaltet wurde, zur Archivierung in Form eines submission information package (SIP) übergeben werden. Das MMS stellt in diesem Kontext unter anderem ein Werkzeug dar, mit dessen Hilfe leichtgewichtig spezielle Formulare mit FDM-spezifischen Metadaten angeboten werden können. Dies erlaubt eine hohe Flexibilität bei der Erfassung von individuellen oder projektbezogenen Metadaten für Forschungsdaten, die archiviert werden sollen.

Gleichzeitig dient das MMS als Verknüpfung zwischen FDM und einem einzuführenden CRIS. Im Datenmodell wurde bewusst der CERIF-Standard berücksichtigt, um Daten automatisiert in ein CRIS zu übertragen.

Die Modularität des MMS erlaubt es, über die bestehenden Anwendungen an der TU Dortmund und der Ruhr-Universität Bochum hinaus, Szenarien des Metadatenmanagements zu unterstützen. Im Sinne der Openness sind Code und Dokumentation über GitHub verfügbar.

Literaturangaben

Depping, Ralf. 2014. „Publikationsservices im Dienstleistungsportfolio von Hochschulbibliotheken. Eine (Neu-)Verortung in der wissenschaftlichen Publikationskette“, *o-bib - Das offene Bibliotheksjournal* 1: 71-91. Online verfügbar unter http://dx.doi.org/10.5282/o-bib/2014H1S71-91. Zuletzt geprüft am 13.03.2017.

Fenner et al. 2016. „A Data Citation Roadmap for Scholarly Data Repositories”, bioRxiv preprint first posted online Dec. 28, 2016. Online verfügbar unter http://dx.doi.org/10.1101/097196. Zuletzt geprüft am 08.03.2017.

Horstmann, Wolfram und Najko Jahn. 2010. „Persönliche Publikationslisten als hochschulweiter Dienst - eine Bestandsaufnahme“, *Bibliothek, Forschung & Praxis*, 34 (1): 185-193. Online verfügbar unter http://doi.org/10.1515/bfup.2010.032. Zuletzt geprüft am 13.03.2017.

Open Science bei Fraunhofer – Serviceentwicklung und Realisierung einer Forschungsdateninfrastruktur für Open Data

Tina Klages[1], Andrea Wuchner[2]

1,2 Fraunhofer-Informationszentrum Raum und Bau

Zusammenfassung. Die Digitalisierung hat als Megatrend inzwischen die Wissenschaft erreicht und führt in diesem Zusammenhang sowohl zu vielen Potenzialen, die sich durch die vorhandenen und zukünftigen technischen Möglichkeiten ergeben, aber auch zu Herausforderungen für Wissenschaftsorganisationen weltweit. In diesem Kontext ist Open Science ein wichtiges Schlagwort, da es inzwischen möglich ist, den gesamten Wissenschaftsprozess von der ersten Idee bis hin zur Veröffentlichung der Forschungsergebnisse offen zu legen. Ziel ist es, neben Transparenz und Nachvollziehbarkeit wissenschaftlicher Ergebnisse und Daten auch deren Nachnutzbarkeit zu gewährleisten, wo dies möglich ist. Kern und Ausgangspunkt von Open Science sind der offene Zugang zu Publikationen (Open Access) und vor allem zu den Forschungsdaten (Open Data). Die Fraunhofer-Gesellschaft hat sich des Themas Open angenommen und steht dabei als Wissenschaftsorganisation der angewandten Forschung vor der Herausforderung, dass ein Großteil der Forschungsprojekte gemeinsam mit Industrieunternehmen durchgeführt wird. Daher sind vor allem Geheimhaltungsinteressen der Unternehmen bei der Ausgestaltung von Open Science für die Fraunhofer-Gesellschaft zu berücksichtigen. Auf das Thema Open Access wirkt sich diese Situation nicht aus, da Publikationen, ob offen zugänglich oder nicht, ohnehin publiziert sind. In Bezug auf das Thema Forschungsdaten ist die Situation eine andere. Hier fallen die Geheimhaltungsinteressen der Unternehmen stark ins Gewicht, was vor allem den offenen Zugang zu Forschungsdaten angeht, der inzwischen z.B. im Zuge des Förderprogramms Horizon2020 von der europäischen Kommission gefordert wird. Daher verfolgt die Fraunhofer-Gesellschaft das Prinzip, Forschungsdaten so offen wie möglich, aber auch so geschlossen wie nötig zu behandeln. Die Umsetzung einer fraunhofer-spezifischen Lösung stößt auf verschiedene Herausforderungen: Kultur, Information, Qualifikation, Prozessentwicklung, Serviceentwicklung, Nachweisbarkeit und Interoperabilität, Disziplinspezifische Kulturen, rechtliche Rahmenbedingungen und Qualität.
Die technische Infrastruktur für den Nachweis und die Veröffentlichung von Forschungsdaten soll mit der bestehenden Publikationsinfrastruktur der Fraunhofer-Gesellschaft verknüpft und in bestehende Publikations- und Beratungsprozesse integriert werden. Darüber hinaus soll die Infrastruktur die Erfüllung von Anforderungen der Forschungsförderer zu Open Data automatisiert erbringen.

Durch die bedarfsgerechte Entwicklung und Etablierung einer Forschungsdateninfrastruktur, die an die heterogenen Ansprüche der 70 Institute der Fraunhofer-Gesellschaft angepasst ist, wird ein wichtiger Beitrag zu der Umsetzung von Open Science erbracht. Die Infrastruktur wird durch entsprechende Services und Unterstützungsangebote für die Fraunhofer-Wissenschaftler begleitet. Durch eine Fraunhofer-Policy zum Umgang mit Forschungsdaten wird das Thema auf strategischer Ebene ebenfalls in die Fraunhofer-Gesellschaft getragen. Somit wird ein ganzheitlicher Ansatz in Bezug auf Forschungsdaten bei Fraunhofer verfolgt.

Schlagwörter. Fraunhofer Gesellschaft, Forschungsdatenmanagement, Forschungsdateninfrastruktur

Ausgangssituation

Die Digitalisierung hat als Megatrend inzwischen neben der Wirtschaft auch die Wissenschaft erreicht und führt in diesem Zusammenhang zu großen Veränderungen im wissenschaftlichen Arbeiten. Durch die Möglichkeiten des Internets und anderen Innovationen im Bereich der Information und Kommunikation arbeiten Wissenschaftler weltweit zusammen an Forschungsfragen und kooperieren dabei über Plattformen, teilen Forschungsergebnisse und beschleunigen dadurch den wissenschaftlichen Fortschritt. In diesem Kontext ist Open Science ein wichtiges Konzept, das es ermöglicht, den gesamten Wissenschaftsprozess von der ersten Idee bis hin zur Veröffentlichung der Forschungsergebnisse offen zu legen. Ziel ist es, neben Transparenz und Nachvollziehbarkeit wissenschaftlicher Ergebnisse bzw. Daten auch deren Nachnutzbarkeit zu gewährleisten, sofern dies z.B. aus rechtlichen Gründen oder Geheimhaltungsründen möglich ist. Gleichzeitig sollen dadurch Innovationszyklen verkürzt werden, um den Forschungs- und Wirtschaftsstandort Europa zu stärken[1]. Das Konzept der Open Science umfasst mehrere Teilbereiche, wie z. B. Open Source, Citizen Science[2]. Kern von Open Science ist jedoch der offene Zugang zu Publikationen (Open Access) und zu Forschungsdaten (Open Access to Research Data). Dieser Paradigmenwechsel von einem analogen und geschlossenen Forschungsprozess zu dessen Digitalisierung und Öffnung ändert das wissenschaftliche Arbeiten von Grund auf. Dadurch ergeben sich viele Potenziale, aber auch Herausforderungen in Bezug auf die Ausgestaltung von Open Science, denen sich Wissenschaftsorganisationen weltweit stellen müssen.

Auch die Fraunhofer-Gesellschaft hat sich des Themas Open Science angenommen. Die Fraunhofer-Gesellschaft ist die führende Organisation für angewandte Forschung in Europa. In Kooperation mit anderen Wissenschaftsorganisationen und Wirtschaftsunternehmen erforscht sie Themen, die sich eng an den Bedarfen der Gesellschaft orientieren: Gesundheit, Sicherheit, Kommunikation, Mobilität, Energie und Umwelt. Mit insgesamt 69 Instituten und 24 500 Mitarbeitern werden dabei vom jährlichen Forschungsvolumen von 2,1 Milliarden Euro ca. 1,9 Milliarden Euro über Vertragsforschung erbracht. Davon werden insgesamt ca. 70 Prozent durch Kooperationsprojekte mit Partnern aus der Wirtschaft und durch öffentlich geförderte Forschungsprojekte erwirtschaftet[3]. Ziel ist neben der Forschung für die Gesellschaft die Beförderung der Innovationsfähigkeit des Wirtschaftsstandortes Deutschland bzw. Europas. Durch ihre besondere Position im Innovationssystem als Bindeglied zwischen Grundlagenforschung und Entwicklung sieht sich die Fraunhofer-Gesellschaft bei der Ausgestaltung von Open Science dabei als Wissenschaftsorganisation der angewandten Forschung vor die Herausforderung gestellt, dass ein Großteil der Forschungsprojekte gemeinsam mit Industrieunternehmen durchgeführt wird. Ziel dieser Kooperationsprojekte ist die gemeinsame Wertschöpfung durch die Erarbeitung von individuellen Lösungen für Fragestellungen aus der Praxis. Ergebnis dieser Aktivitäten sind Innovationen, die auf technologischem Fortschritt basieren und die Marktposition der Kooperationspartner stärken. Aus diesen Gründen ist es essentiell, die Geheimhaltungs- bzw. Verwertungsinteressen der Unternehmen zu berücksichtigen. Die Öffnung von Forschungsprozessen sowie die offene Bereitstellung von Forschungsergebnissen in Form von Publikationen und Forschungsdaten sind in vielen Fällen

1 „EU Digital Market Strategy“, accessed March 29th 2017, https://ec.europa.eu/commission/priorities/digital-single-market_en
2 Homepage der AG Open Science, accessed March 29th, 2017, http://www.ag-openscience.de/
3 „Fraunhofer Zahlen und Fakten“, accessed March 29th 2017, https://www.fraunhofer.de/de/ueber-fraunhofer/profil-selbstverstaendnis/zahlen-und-fakten.html

bei Fraunhofer somit nicht ohne weiteres möglich und müssen fallweise entschieden werden. Die Voraussetzungen für Open Science sind für Organisationen der angewandten Forschung andere als für Einrichtungen, die öffentlich finanzierte Grundlagenforschung betreiben, da deren Ergebnisse meist ohne Einschränkung für die Öffentlichkeit bestimmt sind.

Betrachtet man im Kontext von Open Science den offenen Zugang zu Publikationen und Forschungsdaten, so sind Publikationen in diesem Kontext nicht problematisch, da sich diese grundsätzlich an die Öffentlichkeit richten. Hier wurde die Entscheidung der Veröffentlichung bereits im Vorfeld getroffen. In Bezug auf Daten ist die Situation eine andere: Hier fallen die Geheimhaltungsinteressen der Unternehmen stark ins Gewicht, was vor allem den offenen Zugang zu Forschungsdaten betrifft, der inzwischen z. B. im Zuge des Förderprogramms Horizon 2020[4] von der Europäischen Kommission gefordert wird. Daher verfolgt die Fraunhofer-Gesellschaft das Prinzip, Forschungsdaten so offen wie möglich, aber auch so geschlossen wie nötig zu behandeln.[5]

Herausforderungen und institutioneller Lösungsansatz

Der offene Zugang kann nicht ohne weiteres durch jeden Wissenschaftler einzeln sichergestellt werden. Auch das Management der Forschungsdaten kann, je nach Umfang und Anforderungen, nicht jeder Wissenschaftler im Rahmen seiner Forschungstätigkeit eigenständig leisten. Für Forschungsdaten besteht weiterhin die Anforderung der Nachnutzbarkeit, die durch ein gewisses Maß an Standardisierung über Disziplinen hinweg ermöglicht werden muss. Hier ergibt sich somit der Bedarf einer Bündelung von Kompetenzen im Umgang mit Forschungsdaten in der Fraunhofer-Gesellschaft.

Darüber hinaus ist der Aufbau einer institutionellen Infrastruktur für den Nachweis von Forschungsdaten notwendig, die sowohl die Veröffentlichung als auch den Nachweis von Datensätzen erlaubt, soweit dies im Kontext von Open Data gewünscht ist. Hierbei stellt die direkte Verknüpfung von Datensätzen mit den zugehörigen Publikationen eine wesentliche Anforderung dar. Die Infrastruktur muss dazu mit der bestehenden Publikationsinfrastruktur der Fraunhofer-Gesellschaft verknüpft und in bestehende Publikations- und Beratungsprozesse integriert werden. Darüber hinaus soll die Infrastruktur die Erfüllung von Anforderungen der Forschungsförderer zu Open Data über Schnittstellen automatisiert erbringen, wie beispielsweise zum OpenAIRE Portal der Europäischen Kommission[6]. Ziel ist es, den Umgang mit und den Nachweis bzw. die Publikation von Forschungsdaten durch die Wissenschaftler bestmöglich zu unterstützen. Die Ausgestaltung dieser Vorhaben wird durch das Competence Center Research Services & Open Science geleistet. Dabei handelt es sich um einen zentralen Dienstleister der Fraunhofer-Gesellschaft, der sich u. a. mit Themen beschäftigt, die sich aus der Digitalisierung der Wissenschaft ergeben. Die Entwicklung einer Fraunhofer-spezifischen Lösung birgt eine Reihe von Herausforderungen, die einerseits im Bereich des Umgangs mit Forschungsdaten im Allgemeinen begründet sind, sich andererseits jedoch auch durch die spezifische Struktur der Fraunhofer-Gesellschaft ergeben.

4 Open Data Pilot im Forschungsprogramm Horizon2020, accessed March 29th 2017, https://www.openaire.eu/opendatapilot

5 EARTO Paper on OpenX, accessed March 29th 2017, http://www.earto.eu/fileadmin/content/Website_2/EARTO_Paper_on_Open_X_-_13_November_2015_-_Final.pdf

6 OpenAIRE Portal der Europäischen Kommission, accessed March 29th, 2017 https://www.openaire.eu

Kultur:
Eine Infrastruktur ist zwar eine notwendige, jedoch keine hinreichende Bedingung für den verantwortungsvollen Umgang mit Forschungsdaten. Dazu ist vor allem die Entwicklung einer Kultur des nachhaltigen Umgangs mit Forschungsdaten und deren Management erforderlich. Durch die Digitalisierung und die Entwicklung hin zu Open Science bzw. Open Data hat sich die Bedeutung von Forschungsdaten, offen oder geschlossen, stark gewandelt. Sie stellen inzwischen eine wertvolle Ressource für Wissenschaft und Wirtschaft dar, die einen sorgfältigen Umgang erfordert, um alle damit einhergehenden Potenziale voll ausschöpfen zu können. Hier ist also ein Umdenken der Wissenschaft notwendig, das durch die Schaffung von Awareness unterstützt werden sollte.

Information:
Umfangreiche Informationsaktivitäten sind essentiell, um die veränderte Bedeutung von Forschungsdaten im Kontext von Open Science deutlich machen. Hierbei ist es wichtig, sowohl die damit einhergehenden Herausforderungen als auch die Potenziale von Forschungsdatenmanagement und Open Data zu kommunizieren. Darüber hinaus müssen die Wissenschaftler über den Umgang mit Forschungsdaten informiert werden.

Qualifikation:
Um einen qualifizierten Umgang mit Forschungsdaten in den unterschiedlichen Disziplinen in Einklang mit den Anforderungen der wissenschaftlichen Integrität zu erreichen, bedarf es Maßnahmen für die Qualifizierung der Fraunhofer-Wissenschaftler[7]. Neue Kompetenzprofile sind notwendig, um die Wissenschaftler für die Herausforderungen im Umgang mit Daten, vor allem in Bezug auf große Datenmengen, vorzubereiten bzw. Spezialisten in diesem Bereich auszubilden. Hier existieren unterschiedliche Konzepte und Berufsbilder, z. B. Data Curator, Data Manager, Data Scientist, Data Analyst und Data Librarian.

Prozessentwicklung:
Die Heterogenität der Disziplinen spiegelt sich innerhalb der Fraunhofer-Institute auch in der Heterogenität von Prozessen wieder, dies wird u. a. in Bezug auf Publikationen deutlich. Für einen nachhaltigen Umgang mit Forschungsdaten ist es daher von großer Bedeutung, entsprechende Prozesse innerhalb der Institute zu etablieren. Teil dieser Prozesse müssen qualifizierte Personen sein, die das Thema wissenschaftlich begleiten und die Wissenschaftler beraten.

Serviceentwicklung:
Die Entwicklung und Bereitstellung von Unterstützungsangeboten für die Fraunhofer-Institute und deren Wissenschaftler zu allen Fragestellungen rund um die Themen Forschungsdaten und deren Management bzw. Open Data sowie den damit einhergehenden politischen Anforderungen ist essentiell. Hierzu gehört ein Forschungsdatensupport, der z. B. in Hinblick auf Datenmanagementpläne (wie in Horizon 2020 seit 2017 verpflichtend gefordert) und Forschungsdatenmanagement im Zusammenhang mit Projekten berät.

7 DFG: Leitlinien zur guten wissenschaftlichen Praxis, accessed March 29th 2017, http://www.dfg.de/foerderung/grundlagen_rahmenbedingungen/gwp/index.html

Nachweisbarkeit und Interoperabilität:
Um Forschungsdaten nachzuweisen und diese auch auffindbar zu machen, ist es unerlässlich ein Metadaten-Profil festzulegen, mithilfe dessen es möglich ist, die Datensätze nachhaltig zu beschreiben und in Nachweissystemen zu erfassen. Hier ist es sinnvoll, bereits bestehende und verbreitete Standards nachzunutzen, und diese den Anforderungen der Fraunhofer-Gesellschaft entsprechend anzupassen. So ist die Interoperabilität zu anderen Nachweis-Systemen unter Berücksichtigung der Fraunhofer-spezifischen Besonderheiten gewährleistet.

Disziplinspezifische Kulturen:
Die einzelnen wissenschaftlichen Fachdisziplinen zeichnen sich durch einen unterschiedlichen Umgang mit Forschungsdaten aus. So existiert eine Vielzahl verschiedener Datenformate und vorherrschender Standards, welche disziplinspezifisch zum Einsatz kommen. Auch die anfallenden Datenmengen sind zwischen den jeweiligen Fachdisziplinen unterschiedlich. Diese Heterogenität, die sich bei Fraunhofer widerspiegelt, muss beim Aufbau der Infrastruktur und bei der Entwicklung von Serviceangeboten berücksichtigt werden.

Rechtliche Rahmenbedingungen der Nachnutzung von veröffentlichten Daten:
Um Forschungsdaten nachnutzen zu können, bedarf es das Urheberrecht betreffend klarer Regelungen, die im Rahmen der derzeitigen Rechtsprechung noch nicht geklärt sind. Dennoch hat der Erzeuger der Daten (z. B. über die Vergabe von CC Lizenzen) die Möglichkeit, eindeutige Nachnutzungsrechte zu gewähren.

Qualität:
Die Fraunhofer-Publikationsinfrastruktur, bestehend aus der Publikationsnachweisdatenbank Fraunhofer-Publica und dem Repositorium Fraunhofer-ePrints, unterliegt in Bezug auf den Nachweis von Publikationen und Volltexten einer Qualitätskontrolle, um sicherzustellen, dass die Möglichkeit der Auffindbarkeit der Inhalte besteht. Dies ist auch für den Nachweis von Forschungsdaten unerlässlich, um einen gesicherten Standard festzulegen, der für alle Datennachweise und Datensätze gleichermaßen gilt.

Projekt FORDATIS

Um die bei Fraunhofer generierten Forschungsdaten standardisiert nachzuweisen und zu veröffentlichen, wurde das Fraunhofer-interne Projekt „FORDATIS – Forschungsdateninfrastruktur“ beauftragt.

Das Data Curation Continuum Model[8], das an der Monash University in Australien entwickelt wurde, eignet sich, um zu verdeutlichen, auf welchen Ausschnitt der generierten Forschungsdaten das Projekt abzielt.

8 Treloar, A. & Harboe-Ree, C., 2008. Data management and the curation continuum. How the Monash ex perience is informing repository relationship., accessed March 29th 2017, http://www.valaconf.org.au /vala2008/papers2008/111_Treloar_Final.pdf

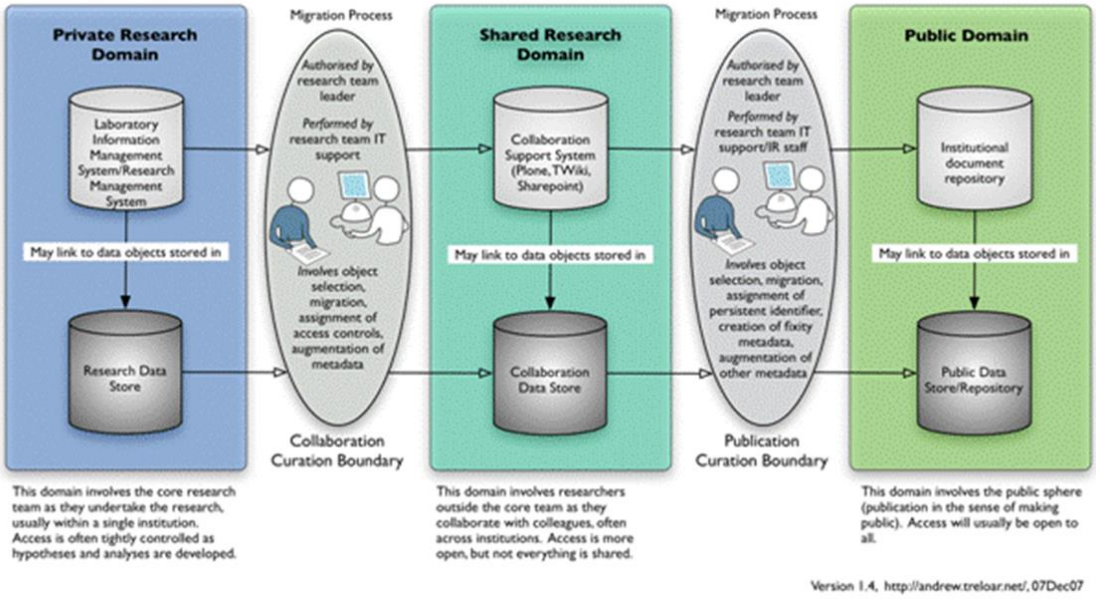


Abbildung 1. Data Curation Continuum (Treloar, A. & Harboe-Ree, 2008)

Das Modell beschreibt die verschiedenen Domänen, in denen Forschungsdaten sich im Verlauf ihres Lebenszyklus bewegen. Erzeugt und erstmalig ausgewertet, werden sie in der Private Research Domain. Dort befinden sich die Daten in sogenannten Research Management-Systemen. Sie werden in die Shared Research Domain überführt, wenn sie anderen Forschern oder Vorgesetzten zugänglich gemacht werden. In dieser Phase sind also Systeme notwendig, die eine Kollaboration von Wissenschaftlern anhand der Forschungsdaten ermöglichen. Mit dem Abschluss der Forschungsarbeiten erfolgt die Veröffentlichung ausgewählter Daten in die sogenannte Public Domain. Darunter wird in diesem Fall verstanden, dass die Daten einer breiten Öffentlichkeit zugänglich gemacht werden. Die Forschungsergebnisse, zu denen auch die Daten zählen, sind erstmalig ausgewertet, können veröffentlicht und für die Nachnutzung zur Verfügung gestellt werden. Dazu werden die Forschungsergebnisse in Repositories migriert. Zusätzlich muss eine Verlinkung mit Persistent Identifier und mit Metadaten versehenen Datenobjekten erfolgen, die sich in einem öffentlichen Forschungsdaten-Repository befinden[9].

An dieser Stelle setzt das Projekt „FORDATIS“ an: Ziel ist es, eine Forschungsdateninfrastruktur für veröffentlichte und zu veröffentlichende Forschungsdaten aufzubauen. Die technische Infrastruktur soll von verschiedenen Services, wie z. B. Support und Schulungen begleitet werden[10], was langfristig zu einer Erhöhung der Awareness im Umgang mit Forschungsdaten führen soll. Betrachtet man den internen Datenfluss bei Fraunhofer, ergeben sich Parallelen zum Data Curation Continuum:

Die Daten werden in Forschungsprojekten an den einzelnen Instituten erzeugt und erstmalig ausgewertet. Hierbei handelt es sich um verschiedene Ausprägungen von Forschungsdaten:

- Forschungsdaten, die unveränderlich sind und in Beziehung zu einer Publikation stehen,
- Forschungsdaten, die unveränderlich für sich stehen, oder aber
- Forschungsdaten, die sich in dieser Phase noch verändern.

9 Büttner, Stephan, Hans-Christoph Hobohm, and Lars Müller, ed. 2011. Handbuch Forschungsdatenmana gement. Bad Honnef: Bock + Herchen ,S. 29-32

10 Eine Erläuterung der Services findet sich auf S. 8f.

Zur Fraunhofer-internen Nachnutzung können unveränderliche Forschungsdaten, sowohl eigenständig als auch im Zusammenhang mit einer Publikation bereitgestellt werden. Das gibt den verschiedenen Fraunhofer-Instituten die Möglichkeit der frühen und internen Nachnutzung. Allerdings haben die verschiedenen Fraunhofer-Institute Kompetenz-Portfolios, die sich teilweise überschneiden, so dass eine interne Veröffentlichung der Daten unter Umständen mit verschiedenen Vorbehalten verbunden ist.

Nach der Erstauswertung der Daten und Abschluss des Forschungsprojektes können diese entweder in Zusammenhang mit Publikationen oder als eigenständige Publikation veröffentlicht werden, was die Sichtbarkeit der Ergebnisse erhöht und es gleichzeitig ermöglicht, der Nachweispflicht bei Förderanforderungen nachzukommen. Um diese Veröffentlichung zu ermöglichen, soll das Forschungsdatenrepositorium „FORDATIS“ an die Publikationsnachweisdatenbank Fraunhofer-Publica angebunden werden. Dadurch wird die direkte Verknüpfung von Forschungsdaten mit den dazugehörigen Publikationen ermöglicht.

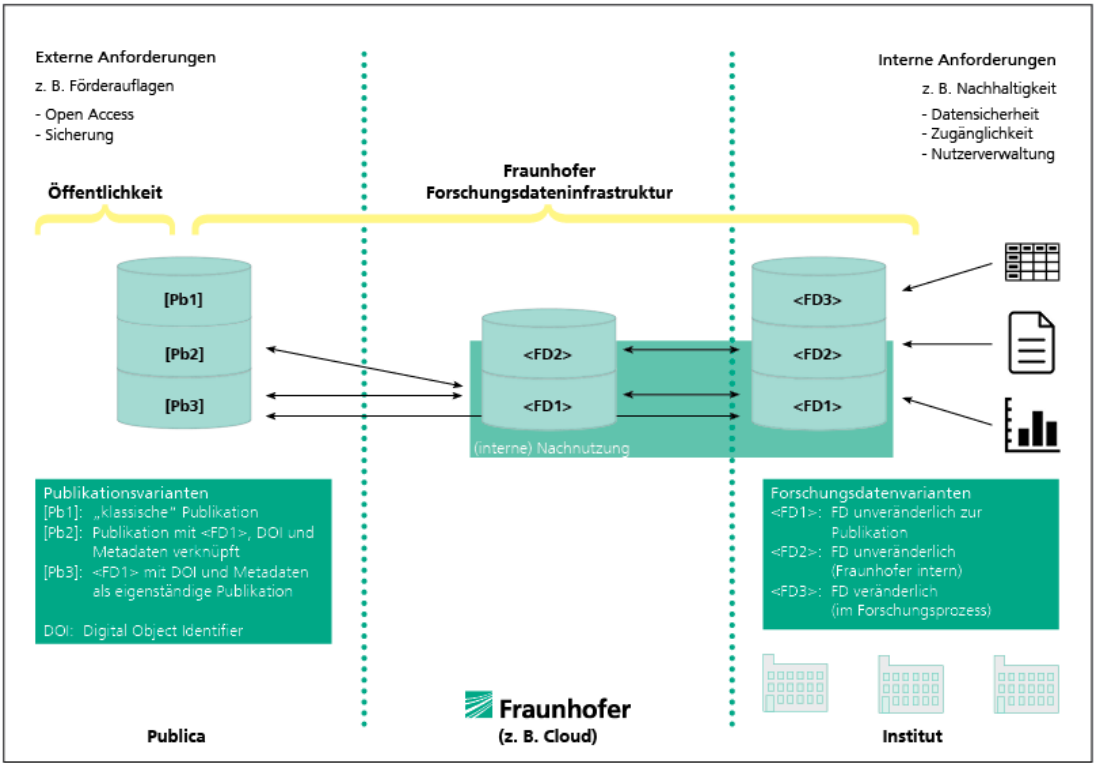


Abbildung 2. Forschungsdaten bei Fraunhofer (Spiecker, Claus 2016)

Umsetzung

Das Projekt „FORDATIS“ wird in dem Zeitraum von Juni 2016 bis Juni 2018 am Fraunhofer-Informationszentrum Raum und Bau (Fraunhofer IRB) bearbeitet.
Dazu beinhaltet das Projekt acht Arbeitspakete:

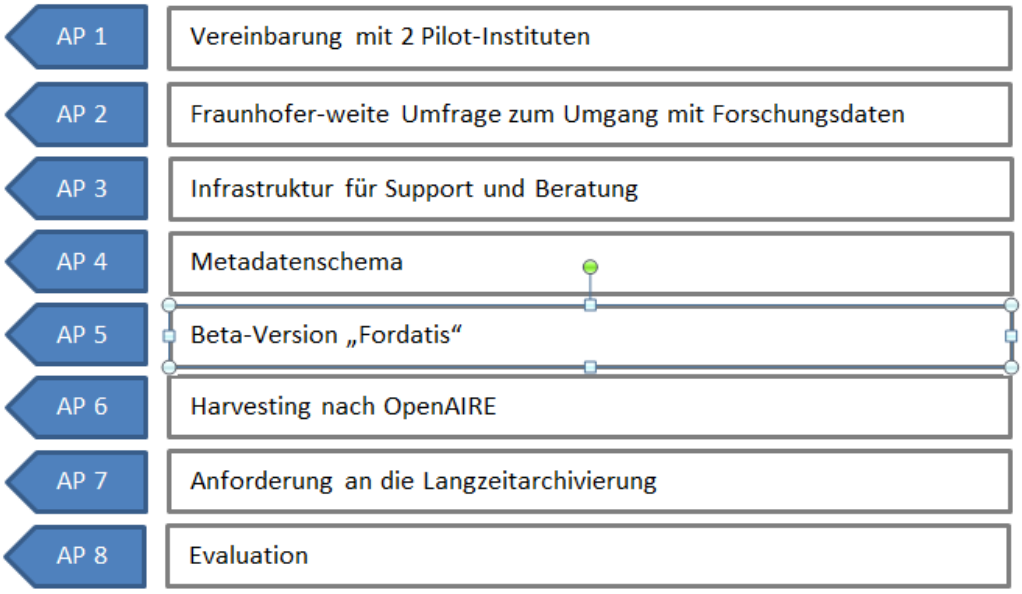


Abbildung 3. Arbeitspakete in FORDATIS

A1: Vereinbarung mit 2 Pilot-Instituten: Zwei Pilot-Institute werden bei der Umsetzung der Auflagen für Forschungsdaten begleitet, beraten und unterstützt. Die gewonnenen Erkenntnisse und Umsetzungen dienen als Basis für Fraunhofer-weite Lösungen in der zweiten Phase.
AP2: Fraunhofer-weite Umfrage: Die Umfrage soll nähere Erkenntnisse der Institute zum Umgang mit Forschungsdaten bringen und so weitere Bedarfe konkretisieren.
AP3: Infrastruktur für Support und Beratung: Elemente des Basisangebots sind Informationsbereitstellung, Schulung, Beantwortung von Anfragen, Beratung speziell für EU-Antragsteller. Hierzu ist die Zusammenarbeit mit internen und externen Multiplikatoren und Netzwerken sowie anderen Wissenschaftsorganisationen notwendig.
AP4: Application Profile: In diesem Arbeitspaket geht es um die Erarbeitung eines Metadatenprofils als Grundlage für eine standardisierte Erfassung von Forschungsdaten.
AP5: Beta-Version „Fordatis“: Programmierung, Test und Realisierung eines Prototyps des erweiterten Forschungsdaten-Repositoriums. Daneben wird ein Meldeworkflow für eine Forschungsdaten-Registrierung entworfen. Im Rahmen dieses Meldeworkflows findet auch die Qualitätskontrolle statt, die für standardisierte Metadaten und Daten sorgt.
AP6: Harvesting nach OpenAIRE: Die Compliance-Anforderungen der OpenAIRE-Plattform zum Forschungsdatennachweis bei EU-Projekten werden überprüft und realisiert.
AP7: Anforderung an die Langzeitarchivierung: In diesem Arbeitspaket sollen die Anforderungen an die Langzeitarchivierung definiert werden.
AP8: Evaluation: Das Projekt soll evaluiert und weiterer Handlungsbedarf und Entwicklungsdesiderate sollen abgestimmt und zur möglichen Projektweiterführung in Phase 2 vorbereitet werden.

Verschiedene Arbeitspakete werden dabei auch im Rahmen des EU-Projekts „JERRI“, Joining Efforts for Responsible Research and Innovation, mitbearbeitet. Das Projekt hat das Ziel fünf Dimensionen von Responsible Research and Innovation an der Fraunhofer-Gesellschaft und der Niederländischen Organisation für Angewandte Naturwissenschaftliche Forschung (TNO) zu betrachten, weiterzuentwickeln und zu implementieren. Das Fraunhofer IRB ist dabei mit der

Dimension „Open Access“ vertreten. Open Access wird dabei als offener Zugang zu Publikationen und Forschungsdaten verstanden.

Kern von FORDATIS ist der technische Prototyp der Forschungsdateninfrastruktur. Dieser soll folgende Funktionalitäten zur Verfügung stellen:

- DOI-Vergabe, um Forschungsdaten eindeutig referenzieren zu können. Die Registrierung der Forschungsdaten wird über die Technische Informationsbibliothek (TIB) Hannover automatisiert erfolgen.
- Forschungsdaten sollen analog zu Publikationen nachgewiesen werden und recherchierbar sein. Der Nachweis von Forschungsdaten erfolgt entsprechend dem Metadaten-Standard DataCite4.0 . Dieser Standard wird von der EU zur Erfüllung der Veröffentlichungspflicht von Forschungsdaten im Rahmen des Förderprogramms Horizon2020 empfohlen . Er ist auch die Grundlage für die Lieferung der Metadaten für die DOI-Vergabe .
- Die Metadaten der Forschungsdaten sollen an das OpenAIRE-Repositorium über die OAI-PMH-Schnittstelle geharvestet werden.
- Die Forschungsdaten sollen analog zu den Volltexten abgelegt und mit den Metadaten verknüpft werden.
- Die Forschungsdaten sollen mit den bibliographischen Angaben der Publikationsnachweisdatenbank Fraunhofer-Publica verknüpft werden. Dies soll über einen Identifier innerhalb der Publikationsnachweise geschehen.

Darüber hinaus soll der technische Prototyp von verschiedenen Service- und Beratungsangeboten flankiert werden. Diese werden bei Fraunhofer traditionell sehr bedarfsorientiert entwickelt. Denkbar sind folgende Angebote:

- Etablierung von Meldeworkflows, Qualitätskontrolle und individuellen Forschungsdaten-Prozessen an den Instituten, welche darauf abzielen, den Umgang mit Forschungsdaten zu begleiten und damit die Veröffentlichung und Langzeitarchivierung zu vereinfachen.
- Beratung zu Forschungsdatenmanagement bei Projektantragsstellung: Die Förderorganisationen stellen verschiedene Anforderungen an den Umgang mit Forschungsdaten. Um den einzelnen Wissenschaftler zu entlasten sollen Beratung und Support während der Projektantragsphase etabliert werden.
- Beratung zum Umgang mit Forschungsdaten im Rahmen von Kooperationsprojekten mit Industriepartnern: Bei Projekten, die Fraunhofer-Wissenschaftler mit Industriepartnern oder im Auftrag aus der Industrie durchführen, gibt es meist ein Geheimhaltungsinteresse, da wissenschaftliche Ergebnisse kommerziell verwertet werden sollen. Dieses Geheimhaltungsinteresse erstreckt sich auch auf Forschungsdaten, die in diesem Fall nicht veröffentlicht werden können. Wissenschaftler sollen im Rahmen von Best-Practices sensibilisiert und beraten werden, wie in diesem Fall mit Forschungsdaten zu Verfahren ist, um einerseits den Anforderungen der Industriepartner nachzukommen, aber auch gemeinsam mit ihnen mögliche Potenziale offener Daten auszuloten.
- Bereitstellung von Templates für Datenmanagementpläne auf Deutsch und Englisch für Projekte des Förderprogramms „Horizon 2020 “. Im Rahmen dieses Förderprogramms müssen Datenmanagementpläne spätestens 6 Monate nach Projektstart eingereicht und während der Projektlaufzeit fortgeschrieben werden. Dazu werden Templates bereitgestellt.
- Schulungen der Fraunhofer-Wissenschaftler im Bereich „Forschungsdaten“: Analog zu bereits etablierten Schulungen im Bereich „Publikationsmanagement“ soll ein Schulungsan-

gebot zum Thema „Forschungsdaten“ etabliert werden. Dieses soll sich an die Fraunhofer-Wissenschaftler wenden.

Fazit und Ausblick

Die Fraunhofer-Gesellschaft kann mit dem Projekt „FORDATIS“ einem Großteil der zuvor skizzierten Herausforderungen im Zusammenhang mit Forschungsdatenmanagement bzw. Open Data begegnen:

Information: Durch Supportangebote ist es möglich, den Wissenschaftlern die Herausforderungen, aber auch die Potenziale der Veröffentlichung von Forschungsdaten zu kommunizieren und sie in der Praxis zu beraten.

Prozessentwicklung: Im Rahmen des Projekts kann die Entwicklung von Prozessen zum Umgang mit Forschungsdaten an den Instituten unterstützt werden. Dies kann exemplarisch mit den Pilotinstituten erfolgen.

Serviceentwicklung: Das Projekt sieht die Entwicklung und Durchführung oben skizzierter Services im Bereich „Forschungsdaten“ vor. Diese sollen langfristig verstetigt werden.

Nachweisbarkeit und Interoperabilität: Durch die qualitätsgesicherte Erfassung von Metadaten wird die Nachweisbarkeit von Forschungsdaten gewährleistet. Die Metadaten basieren auf dem Standard DataCite 4.0, wodurch die Interoperabilität der Daten abgedeckt ist.

Qualität: Die Qualität der Metadaten wird durch die manuelle Qualitätskontrolle der erfassten Metadaten gewährleistet.

Qualifikation: Dieser Herausforderung kann im Rahmen von FORDATIS nur zu einem Teil entsprochen werden. Für Wissenschaftler werden Schulungen zum Thema Forschungsdatenmanagement angeboten. Die Qualifizierung von Data Scientists und Data Analysts muss in einem anderen Rahmen erfolgen.

Rechtliche Rahmenbedingungen der Nachnutzung: Die EU macht diesbezüglich Empfehlungen hin zu den Lizenzen CC-BY und CC-O.[11] Für Fraunhofer könnten rechtliche Rahmenbedingungen in einer Forschungsdaten-Policy festgelegt werden.

Disziplinspezifische Kulturen: Durch die Zusammenarbeit mit verschiedenen Pilotinstituten soll der disziplinspezifische Umgang mit Forschungsdaten bei Fraunhofer näher untersucht werden. Die Ergebnisse werden in die Ausgestaltung von FORDATIS mit einbezogen.

Die Arbeiten an Fraunhofer-FORDATIS werden nach Abschluss der Projektphase 1 (Juni 2018) voraussichtlich noch nicht beendet sein. Der Projekt-Antrag sieht eine weitere Phase vor, in der die gewonnenen Erkenntnisse aus Umfragen und Zusammenarbeit mit Pilot-Instituten umgesetzt und die aufgebaute Infrastruktur zum Beispiel durch Anbindung an weitere, im Aufbau befindlichen Infrastrukturen, wie die Nationale Forschungsdateninfrastruktur[12] oder Europäische Open Science Cloud[13] erweitert werden soll. In diesem Rahmen sind weitere Standardisierungsvorgaben „Top-Down“ zu erwarten. Zentrale Services, wie Beratung oder Schulungen sollen optimiert und verstetigt werden. Die Feinplanung der Phase 2 ist abhängig von den Erkenntnissen

11 Open Research Data Pilot, accessed 29th March 2017, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

12 RfII – Rat für Informationsinfrastrukturen: Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland, Göttingen 2016, S. 2

13 European Open Science Cloud, accessed 29th March 2017, http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

aus Phase 1, insbesondere der Umfrage und den Erfahrungen mit den Pilotinstituten sowie den Erkenntnissen aus JERRI.

Ein weiterer, großer Themenbereich für Phase 2 ist die Langzeitarchivierung der Forschungsdaten. Die Anforderungen, die in Phase 1 definiert werden, sollen dabei umgesetzt werden. Angestrebt ist eine Langzeitarchivierung entsprechend dem OAIS-Modell[14]. Es müssen Fragen hinsichtlich der Datenmenge und der Dauer der Archivierung geklärt werden.

Literaturangaben

„EU Digital Market Strategy“. Online verfügbar unter https://ec.europa.eu/commission/priorities/digital-single-market_en. Zuletzt geprüft am 29.03.2017.

Homepage der AG Open Science. Online verfügbar unter http://www.ag-openscience.de/. Zuletzt geprüft am 29.03.2017.

„Fraunhofer Zahlen und Fakten“. Online verfügbar unter https://www.fraunhofer.de/de/ueber-fraunhofer/profil-selbstverstaendnis/zahlen-und-fakten.html. Zuletzt geprüft am 29.03.2017.

“Open Data Pilot im Forschungsprogramm Horizon2020”. Online verfügbar unter https://www.openaire.eu/opendatapilot. Zuletzt geprüft am 29.03.2017

“EARTO Paper on OpenX”. Online verfügbar unter http://www.earto.eu/fileadmin/content/Website_2/EARTO_Paper_on_Open_X_-_13_November_2015_-_Final.pdf. Zuletzt geprüft am 29.03.2017.

“OpenAIRE Portal der Europäischen Kommission”. Online verfügbar unter https://www.openaire.eu. Zuletzt geprüft am 29.03.2017.

„DFG: Leitlinien zur guten wissenschaftlichen Praxis“. Online verfügbar unter http://www.dfg.de/foerderung/grundlagen_rahmenbedingungen/gwp/index.html. Zuletzt geprüft am 29.03.2017.

“Data management and the curation continuum. How the Monash experience is informing repository relationship”. Online verfügbar unter http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf. Zuletzt geprüft am 29.03.2017.

Büttner, Stephan, Hans-Christoph Hobohm, and Lars Müller (Hrsg.). 2011. *Handbuch Forschungsdatenmanagement*. Bad Honnef: Bock + Herchen: 29-32.

“Open Research Data Pilot”. Online verfügbar unter http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf. Zuletzt geprüft am 29.03.2017.

14 Nestor Handbuch Forschungsdatenmanagement, accessed 29th March 2017, http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_368.pdf

RfII – Rat für Informationsinfrastrukturen. 2016. *Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland*. Göttingen: 2.

European Open Science Cloud. Online verfügbar unter http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud. Zuletzt geprüft am 29.03.2017.

Nestor Handbuch Forschungsdatenmanagement. Online verfügbar unter http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_368.pdf. Zuletzt geprüft am 29.03.2017.

Erfassung und Speicherung von Forschungsdaten im Fachbereich Chemie: Bereitstellung moderner Forschungs-infrastrukturen durch ein elektronisches Laborjournal mit Repositorium-Anbindung

Nicole Jung[1], Pierre Tremouilhac[2], Claudia Kramer[3], Jan Potthoff[4]

1 Institut für Organische Chemie, Institut für Toxikologie und Genetik, KIT
2 Institut für Toxikologie und Genetik, KIT
3 KIT-Bibliothek
4 Steinbuch Centre for Computing, KIT

Zusammenfassung. Eine moderne Infrastruktur zur Erfassung, Bearbeitung und Speicherung von Forschungsdaten bildet die Voraussetzung für eine strukturierte Dokumentation der Forschung zur effizienten Nachnutzung der Ergebnisse und der Daten. Für den Fachbereich Chemie wurden zwei Komponenten (elektronisches Laborjournal [ELN] und Repositorium) identifiziert, durch deren Entwicklung und Vernetzung Forschungsdaten über das bisher mögliche Verfahren hinaus bewahrt und nutzbar gemacht werden sollen. Beide Komponenten des Chemotion-Projektes werden am KIT programmiert und als Open Source Infrastruktur anderen Forschern zur Verfügung gestellt. Die Entwicklung des Laborjournals ist zunächst auf die Bedürfnisse organischer Chemiker ausgelegt und umfasst in der aktuellen Version umfassende Möglichkeiten für eine moderne Arbeitsweise im synthetischen Labor. Neben den Basisfunktionen wurden verschiedene Elemente eingebaut, die über die Dokumentation von Forschungsergebnissen hinaus nachhaltiges Arbeiten und eine eindeutige, korrekte Strukturierung wissenschaftlicher Arbeit unterstützen. Durch die Beteiligung des Rechenzentrums werden Strukturen geschaffen, die eine parallele, nachhaltige Sicherung der Rohdaten zu den jeweiligen Experimenten erlauben. Durch die Kombination des ELNs mit einem Repositorium, dessen Schnittstelle die Datenstruktur der ELN Inhalte aufgreift, wird es dem Forscher ermöglicht, Daten ohne weitere Vorbereitung auf direktem Weg anderen Forschern zur Verfügung zu stellen. Durch die parallele Entwicklung des ELN und des Repositoriums werden mehrere Aspekte des Forschungsdatenmanagements adressiert. Darüber hinaus ermöglicht dies eine effiziente Suche fachspezifischer Forschungsdaten. Das neue Modell aus Kombination von ELN und Repositorium sieht eine Kombination aus Erfassung und Dokumentation sowie Publikation und Speicherung von Datensätzen vor. Die Indexierung, welche aktuell erst nach der Publikation von Daten über Journale oder Patente stattfindet (z.B. über den Chemical Abstracts Service, CAS), soll schon direkt nach der Entstehung vorgenommen werden können. Damit werden bereits die Rohdaten mit semantischen Informationen sowie Metadaten versehen, die insbesondere bei der Verwendung des OAIS-Referenzmodells (Open Archival Information System) von entscheidender Bedeutung für eine nutzerorientierte Verwendung der archivierten Daten sind. So kann in allen Publikationsprozessen auf eine einheitliche, eindeutige Quelle verwiesen werden. Der freie Zugang und die kostenfreie Nutzung für akademische Forscher sollen in Zukunft die Grundlage für gemeinsames Arbeiten bilden, eine Bereitstellung der Forschungsdaten ermöglichen und den Austausch unter den Wissenschaftlern fördern.

Schlagwörter. elektronisches Laborbuch, Repositorium, Datenmanagement, Chemie, Synthese

Einleitung

Im Fachbereich Chemie besteht, wie in anderen naturwissenschaftlichen Disziplinen, bisher ein Mangel an frei nutzbarer, fachspezifischer Software für einen zeitgemäßen Umgang mit Forschungsdaten (Winkler-Nees 2013). Hierdurch wird besonders die Bereitstellung von Forschungsinformationen erschwert. In den letzten 15 Jahren konnte der Zugang zu publizierten Forschungsergebnissen stark verbessert werden. Dies liegt vor allen Dingen an der Verfügbarkeit von Publikationen in online-Versionen der wissenschaftlichen Journale und der web-basierten Bereitstellung von Informationen durch Datenbanken wie SciFinder oder Reaxys (Beilstein).[1] [2]Während diese Entwicklungen die Suche nach veröffentlichten Informationen erleichtert haben, fehlen Lösungen für eine umfassende digitale Verfügbarkeit anderer Forschungsdaten. Ein Grund dafür sind die anspruchsvollen Anforderungen an die Forschungsinfrastruktur und Software, da abhängig vom Forschungsgebiet mehrere Aspekte der Laborarbeit in den elektronischen Datenerfassungs- und Lagersystemen abgebildet werden müssen. Besondere Herausforderungen müssen im Forschungsbereich Chemie überwunden werden, da die Zeichnung und Verarbeitung von chemischen Strukturen ein entscheidender Schritt für die Korrelation von Forschungsdaten mit der entsprechenden chemischen Umsetzung oder Struktur ist (Coles et al. 2013). Während in den letzten Jahren mehrere ELN (wie SciNote, Biovia ELN, EMEN, Open BIS-ELN LIMS, LabFolder) [3] [4] [5] (Rees et al. 2013; Barillari et al. 2016; Rubacha et al. 2011) entwickelt wurden, die intelligente Lösungen für die Dokumentation von Forschungsdaten bieten, sind im Bereich chemische Forschung nur wenige Produkte verfügbar. Beispiele für geeignete Systeme in der Chemie sind das PerkinElmer E-Notebook für Chemie, Indigo-ELN oder LabTrove - und Open Enventory, von denen nur die letzteren drei Produkte kostenlos zur Verfügung stehen (Day et al. 2015; Willoughby et al. 2014; Milsted et al. 2013; Rudolphi & Goosen 2012) [6] [7]. Für einen Einsatz in der Hochschule, der Flexibilität der Software-Struktur und möglichst kostenlose Verfügbarkeit erfordert, sollten entsprechende elektronische Journale vorzugsweise zusätzlich als Open Source zur Verfügung stehen. Die derzeit verfügbaren Systeme wurden von unserer Projektgruppe auf ihre Verwendung als modernes, flexibles Managementsystem für Chemieforschungsdaten untersucht. Die identifizierten Anforderungen an ein System, das den Herausforderungen des zukünftigen Datenmanagements genügt, wurden jedoch nicht erfüllt. Auch steht bisher kein Repositorium im Forschungsbereich Chemie zur Verfügung, welches als Infrastruktur die Speicherung und Bereitstellung von Daten auf einfachem Wege ermöglicht. Forschungsdaten, die für sich alleine genommen keine Publikation rechtfertigen, können bisher in Web-Portalen wie ChemSpider (Kelly & Kidd 2015) (Synthetic Pages) oder SDBS (AIST[8]) abgefragt werden. Allerdings bieten diese Portale meist keine Möglichkeit zur Bereitstellung eigener Daten (SDBS). Wenn dies der Fall ist, erschwert eine zeitaufwändige Eingabe von Forschungsdaten die Pflege der Portale mit aktuellen Daten und lässt eine kontinuierliche Beteiligung der Forscher unattraktiv werden. Weitere Schwierigkeiten bisheriger Repositorien sind die nur begrenzt mögliche Beschreibung von Reak-

1 http://www.cas.org/products/scifinder
2 https://www.reaxys.com/reaxys/secured/search.do
3 https://www.labfolder.com/
4 http://accelrys.com/products/unified-lab-management/biovia-electronic-lab-notebooks/
5 https://github.com/biosistemika/scinote-web
6 https://www.cambridgesoft.com/Ensemble_for_Chemistry/ENotebookforChemistry/
7 https://github.com/ggasoftware/indigo
8 http://www.aist.go.jp

tionen, die geringe Flexibilität der unterstützten Datenformate für experimentelle Daten (meist lediglich Bildformate) und das Fehlen von Schnittstellen zu elektronischen Laborjournalen. Da gerade letzterer Punkt zu einer nur sehr geringen Beteiligung bei der Eingabe von Daten in Repositorien führt, sollte eine moderne Infrastruktur zur Speicherung und Offenlegung von Forschungsinformationen den Transfer von Daten aus einem ELN hin zu einem offenen Repositorium, ebenso wie von einem Repositorium hin zu einem ELN ermöglichen. Die Nutzung einer Kombination aus elektronischem Laborjournal und Repositorium (dargestellt in Abbildung 1) wurde unseres Wissens in den chemischen Wissenschaften noch nicht bereitgestellt.

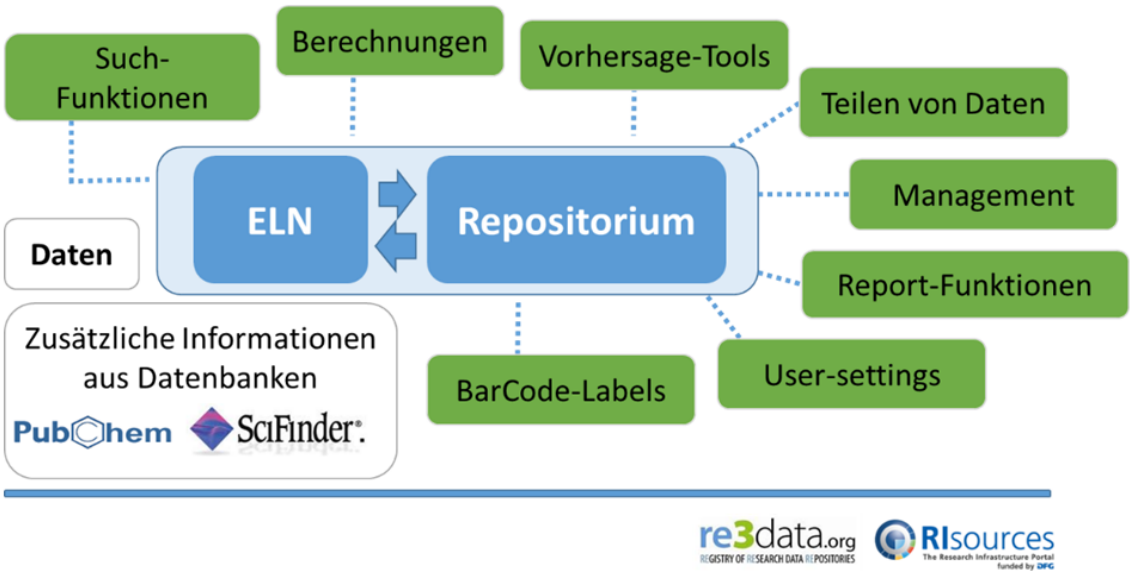


Abbildung 1. Gemeinsame Funktionen von ELN und Repositorium zum Austausch und zur Speicherung von Forschungsdaten.

Chemotion-ELN zur Speicherung digitaler Forschungsdaten

Generische Funktionen und Management

Das ELN der Chemotion-Projektgruppe bietet Grundfunktionen für die forschungsspezifische Arbeit und wird durch generische Werkzeuge für Projekt- und Datenmanagement ergänzt. Diese ermöglichen die Erstellung einer klaren Struktur zur Speicherung und Suche von Forschungsdaten. Das Management von Projekten und die Sortierung einzelner Elemente wird durch die Generierung von sogenannten Collections (oder Sammlungen) unterstützt, die durch einen separaten Organizer generiert, bearbeitet und gelöscht werden können. Die Zuordnung von ausgewählten Elementen zu den gewünschten Sammlungen oder Untersammlungen kann auf einfache und schnelle Weise per Drag & Drop durchgeführt werden. Einmal getroffene Entscheidungen über die Konfiguration der ELN-Struktur können jederzeit problemlos modifiziert werden, sodass das ELN flexibel auf Veränderungen der Forschungsprojekte ausgerichtet werden kann. Der Nutzer des ELNs kann durch das Management-System nicht nur die Daten verwalten, die in eigenen Sammlungen strukturiert sind, es können ebenfalls Projekte verwaltet werden, die durch Zusammenarbeit mit anderen Wissenschaftlern entstehen. Neben der Organisation der Projekte durch

Sammlungen, können einzelne Elemente des ELNs (z.B. Moleküle, Reaktionen, Wellplatten, Experimente oder Vorhaben) in getrennten Listen sortiert und verwaltet werden. Die Einträge innerhalb der Listen sind durch zusätzliche Informationen über die Präsenz der Elemente in den Sammlungen oder aber auch anderen Datenbanken versehen. Bei Auswahl des gewünschten Elements wird eine zweite Informationsebene einsehbar, die durch weitere Details die Informationen der Listen ergänzt. Eine schnelle Organisation der Forschungsdaten wird durch diverse Funktionen wie Drag & Drop, automatisierte Sortierung von Elementen, Farbkennzeichnung für Warn- und Erinnerungsfunktionen erleichtert. Alle Elemente wie Proben und Reaktionen, die bestimmten Sammlungen zugeordnet und in Listen zusammengefasst wurden, können nach ihrer Zuordnung organisiert werden, um eine flexible Gesamtstruktur des ELNs zu ermöglichen. Das ELN wurde als Webanwendung realisiert, so dass die Installation weiterer Software nicht notwendig ist. Zur Nutzung wird lediglich ein aktueller Webbrowser benötigt.

Besondere Qualitätssicherungsmodule

Das Chemotion ELN kann für eine exakte Nachverfolgung von einzelnen Elementen verwendet werden. Dies setzt die systematische und automatische Nummerierung aller Einträge einschließlich einer intuitiven Zuordnung zum jeweiligen Workflow voraus. Entsprechend dieser Bedingung tragen ELN-Einträge, die Teil eines Prozesses sind, Informationen über ihren Ursprung innerhalb des Namens. Einträge, die neu erstellt wurden oder die über eine Kopie erzeugt wurden, erhalten jeweils einen neuen Namen, der aus den Initialen des ELN-Benutzers und einer sequentiellen Nummernfolge besteht. Es ist möglich, Sub-Elemente aus Elementen herzustellen, welche als Child-Item betrachtet werden und durch die Anbringung einer speziellen Chargennummer vom Original unterscheidbar sind. Das ELN ist so konzipiert, dass eine flexible Dokumentation des Forschungsprozesses möglich wird. Gleichzeitig wird jedoch auch eine Manipulation von Werten verhindert. Während alle Parameter einer Reaktion protokolliert und entweder über vordefinierte Felder oder freie Textfelder eingegeben werden können, gibt es weitere Felder, in denen berechnete Daten nur sichtbar, aber nicht editierbar sind. Ein Beispiel für eine solche Begrenzung ist die Eingabe der Ausbeute für Reaktionen. Die Möglichkeit, einen Wert für die Ausbeute einer Reaktion hinzuzufügen, ist für alle Anwendungen deaktiviert, da die Ausbeute jeweils das Ergebnis aus der gewonnenen Menge ist und niemals als Zielwert eingegeben wird. Ein besonderes Merkmal des Chemotion ELN ist die Möglichkeit zur Aufzeichnung von realen Werten parallel zu jenen des ursprünglich geplanten Experiments. Dies ermöglicht die genaue Dokumentation des realen Experiments, während die Information der geplanten Prozedur zusätzlich gespeichert wird.

Einbindung von Informationen aus externen Datenbanken

Informationen aus externen Datenbanken ermöglichen eine objektive und schnelle Einschätzung der eigenen Forschung und wurden daher als wichtiges Element dem ELN beigefügt. In der aktuellen Version unterstützt das Chemotion-ELN die Abfrage von Daten aus der Datenbank SciFinder (kommerziell, Lizenz notwendig) und PubChem. Die Einbindung der SciFinder-Suche wurde durch ein PlugIn realisiert und unterstützt die Suche in der CAS SciFinder Datenbank nach drei verschiedenen Suchmodi. Die Anzahl der Suchergebnisse wird mit einem Link zur Antwort in der

SciFinder-Webanwendung abgerufen. Die direkte Sichtbarkeit der veröffentlichten Strukturen über die ELN ermöglicht einen schnellen Zugriff auf Informationen. Um einen umfassenden Überblick über die Neuheit der Forscherarbeit und die Verfügbarkeit von Forschungsdaten zu geben, wurde zusätzlich ein automatisiertes Verfahren zur Abfrage von gegebenenfalls vorhandenen PubChem-Einträgen implementiert. Wie für die eingebettete SciFinder-Funktion gegeben, sind die passenden Treffer über einen direkten Link zum PubChem Index des identifizierten Elementes zugänglich.

Fachspezifische Funktionen

Das ELN bietet neben der allgemeinen Ausstattung der Software alle notwendigen Funktionen für die Dokumentation und Bearbeitung von chemischen Projekten, einschließlich der Verarbeitung von Molekülen und Reaktionen. Hierzu wurde basierend auf dem als Open Source verfügbaren Struktur Editor Ketcher ein flexibel nutzbarer, fortgeschrittener Strukturedito r erstellt. Eingebettet in das ELN und unterstützt durch eine Verwertung der generierten Information innerhalb des ELNs können zusätzliche Funktionen wie die Eingabe von Templates oder die automatische Berechnung von Strukturmerkmalen eingebunden werden. Die interne Struktur des ELNs folgt strengen Regeln bei der Schaffung neuer Elemente, die die Differenzierung zwischen Molekülen und Einzelproben bewirken. Während Moleküle jeweils mit den Merkmalen der chemischen Struktur und den hieraus errechenbaren Werten assoziiert werden (verfügbar durch z.B. OpenBabel), erlauben Einzelproben ebenfalls die Speicherung zusätzlicher Eigenschaften, die durch die Umgebung oder Anwendungsform (Reinheit, Zusammensetzung, Temperatur) bestimmt werden. Die Registrierung und konsequente Verwendung von Molekülen oder Einzelproben während der Arbeit mit dem ELN ist die Basis für eine gut organisierte und am Ende reproduzierbare synthetische Dokumentation. Während eine solche klare Differenzierung zwischen Molekülen und Proben in den meisten anderen Chemiejournalen nicht angeboten wird, stellt das Chemotion ELN dieses wichtige Merkmal in einer sehr einfachen Weise vor. Die Abbildung einer physikalisch genutzten Substanz oder ihrer Zubereitung umfasst die Zusammenfassung der verfügbaren Daten aus dem verwandten Molekül, die eine schnelle Verfügbarkeit aller Informationen ermöglichen, die für ein schnelles Management des Forschungsprojekts notwendig sind. Die automatisch bereitgestellten Daten sowie die vom Benutzer vorgegebene Eingabe sind in fünf sogenannten Sample-Tabs organisiert. Sie bestehen aus (1) Informationen für eine detaillierte Definition der Eigenschaften, (2) zusätzliche Daten, die an die hochgeladenen Dateien mit Forschungsdaten angehängt werden können, (3) Ergebnissen, die mit der Probe durch einen externen Prozess erhalten wurden, (4) einer automatisierten Online-Anfrage der SciFinder-Datenbank und einer direkten Verbindung zu den Suchergebnissen sowie (5) vorhergesagten NMR-Informationen.

Neben Einzelproben und Molekülen gehören Reaktionen zu den Hauptelementen, die durch das ELN erzeugt und verwaltet werden sollen. Reaktionen können sehr schnell erzeugt werden und durch Hinzufügen von Einzelproben in Funktion von Ausgangsmaterial, Reagenz oder Produkt definiert werden. Das Grundschema für Proben in Reaktionen erlaubt die Zugabe der Menge der Substanzen in verschiedenen Einheiten und die Definition der verwendeten Substanz in Äquivalenten. Aus den gegebenen Werten werden die fehlenden Informationen für eine Reaktion automatisch berechnet. Die Struktur der Reaktions-Benutzeroberfläche ist sehr flexibel, so dass eine Änderung der Zuordnung der Einzelelemente jederzeit möglich ist. Alle Abhängigkeiten werden

kontinuierlich an die Änderungen angepasst. Die Chemikalien, die der Reaktion zugeordnet sind, sind über eine direkte Verbindung zur Detailstufe der Einzelproben zugänglich und alle Daten und Änderungen, die den Proben (z.B. Dichte der Chemikalien) zugeordnet werden fließen sofort in die Berechnung der Reaktion ein. Während das Reaktionsschema und die Reaktionstabelle durch zusätzliche Informationen in Freitext-Form vervollständigt werden können, wird das Hinzufügen einer Reaktionsbeschreibung durch mehrere formatierte Vorlagen unterstützt, die für einen schnellen Bericht über ein chemisches Verfahren in einer standardisierten Weise verwendet werden könnten. Zusätzlich wurden drei weitere Eingabeseiten „Eigenschaften“, „Literatur“ und „Analyse “für die Bereitstellung weiterer Informationen geschaffen.

Austausch und Freigabe von Daten

Um einen möglichst einfachen Austausch von Forschungsinformation zu ermöglichen, werden innerhalb des ELNs Export- und Importfunktionen für einzelne Proben, Reaktionen und Sammlungen erstellt. Verfügbare Dateiformate für diese Anwendungen sind Excel (.xlsx) und SDF (.sdf). Die Details des Exportvorgangs können vom ELN-Benutzer über eine Auswahl der Spalten bestimmt werden, die in der exportierten Datei angegeben werden sollen. Neben dem Import und Export wurden zusätzlich zwei Funktionen („Teilen“ und „Synchronisieren“) implementiert, um Informationen mit anderen Forschern teilen zu können. Die Benutzeroberfläche erlaubt die Organisation einzelner Gruppen nach ihrem Status und nach gewünschten Rechten. Der Benutzer des ELNs und Eigentümer der übermittelten Daten kann dazu eigene Rollen definieren und Gruppen zuweisen oder eine vordefinierte Rolle verwenden. Die Definition von Freigabeberechtigungen umfasst die Auswahl einer Berechtigungsstufe für zulässige Aktionen, die vom Lesen bis zum Besitz und der Ermittlung der verfügbaren Detailstufe für jedes der verfügbaren Elemente reichen. Während die Auswahl der Benutzerrolle und der Rechte für das Teilen- und Synchronisierungswerkzeug gleich sind, unterscheiden sich die beiden Aktionen hinsichtlich der Inhalte der bereitgestellten Forschungsdaten: Die gemeinsame Nutzung von Informationen durch Teilen ermöglicht den Zugriff auf einen gegebenen Satz Forschungsdaten. Der ausgewählte Kollege kann die Informationen z.B. lesen, schreiben, teilen und löschen hat aber keinen Zugriff auf laufende Änderungen der gemeinsamen Sammlung. Dies ist nur über die Erstellung von synchronisierten Sammlungen verfügbar. Synchronisierte Sammlungen werden erstellt, um einen permanenten Zugriff auf andere ELN-Benutzer auf den ausgewählten Satz von Forschungsdaten zu ermöglichen, einschließlich der Sichtbarkeit von Änderungen, die nach der Synchronisation vorgenommen wurden.

Ablage von Messwerten

Neben den unter einer Nutzerkennung eingetragenen Werten fallen in Laboratorien eine Vielzahl digital verfügbarer Daten an, die zur vollständigen Dokumentation des Forschungsprozesses mit im elektronischen Laborbuch erfasst werden müssen. Um eine Nachnutzbarkeit und eine nachvollziehbare Dokumentation zu gewährleisten, müssen diese Daten strukturiert abgelegt werden können. Das heißt, die Daten müssen zu den entsprechenden Analysen zugeordnet werden können. Ist eine Assoziation der Daten zu einem Benutzer nicht gegeben, wie beispielweise bei

der automatischen Speicherung von Daten durch Messgeräte, muss dem Nutzer des elektronischen Laborbuchs eine Möglichkeit gegeben werden, diese Daten den entsprechenden Versuchen zuordnen zu können.

Die Daten können auf dem lokalen Filesystem des Servers gespeichert werden, auf dem auch die elektronische Laborbuch Anwendung installiert ist. Um ein Management größerer Datenmengen gewährleisten zu können, wurde eine Anbindung der Large Scale Data Facility (LSDF) ausgewählt. Da die Daten in einem elektronischen Laborbuch aktiv genutzt werden, das heißt die Daten häufig abgerufen, verändert oder ergänzt werden, sollten die Daten auf dem Datenspeicher in einer einfachen Ordnerstruktur vorliegen, um ein häufiges umkopieren der Daten zu vermeiden. Für die spätere Langzeitarchivierung ist es jedoch notwendig über das Datenmodell strukturierte Datencontainer inklusive der Metadaten (zum Beispiel im BagIt-Format) erzeugen zu können. Mit der Anbindung eines Langzeitarchivs können die Daten dann kostengünstig nachhaltig gesichert werden. Um die Nachhaltigkeit weiter zu adressieren, wird die Integrität der Daten mit ihrer Erzeugung erfasst und bis zur Ablage ins Archiv kontrolliert. Damit ist ein Nachweis über die Unversehrtheit der Daten über den gesamten Forschungszyklus hinweg möglich.

Chemotion-Repositorium zur Bereitstellung von Forschungsdaten

Das Repositorium für chemische Forschungsdaten ist mit den gleichen allgemeinen und fachspezifischen Funktionen ausgestattet, die das Chemotion-ELN bietet. Einige wenige Funktionen wurden hinzugefügt, um eine Datenübermittlung zur Registrierung der Forschungsdaten über DataCite zu ermöglichen und eine Indexierung in der Datenbank PubChem zu initiieren. Die Installation der Software muss auf einem allgemein zugänglichen Server durchgeführt werden, um die Verfügbarkeit der bereitgestellten Daten auch außerhalb der eigenen Forschungseinrichtung zu ermöglichen. Zu diesem Zweck wurde am KIT die Installation auf einer virtuellen Maschine des Rechenzentrums mit öffentlich zugänglicher IP-Adresse gewählt. Während in einer frühen Version des Chemotion-ELNs Daten noch einzeln eingegeben werden mussten, gelingt durch die Verwendung der Freigabe-Funktion des ELNs explizit für die Sammlung Chemotion-Repositorium eine einfache und schnelle Bereitstellung der erhaltenen Forschungsdaten. Zu diesem Zweck können die bereits innerhalb des ELNs zur Verfügung stehenden Module der Rechteverwaltung genutzt und auf den Transfer der Daten und den gewünschten Detaillevel hin eingestellt werden.

Suche nach Forschungsdaten

Eines der Hauptargumente für die Verwaltung von Forschungsdaten mit einem ELN ist die digitale Verfügbarkeit von Informationen und die damit verbundene Möglichkeit, nach Daten und Informationen zu suchen, wenn die Organisation und Pflege des ELNs bzw. des Repositoriums dies in geeigneter Weise unterstützt. Die Chemotion Produkte ermöglichen Text- und Struktursuche innerhalb diverser Inhalte des ELNs. Die Suche nach Textfragmenten oder chemischen Strukturen kann weiter auf unterschiedliche Elemente (Proben, Reaktionen) beschränkt werden, um die Auswertung der Ergebnisse zu erleichtern. Die Textsuche ist dazu bestimmt, nach dem Vorhandensein von Text- oder Formelfragmenten in Proben als Basiseinheit zu durchsuchen. Die meisten der nicht-numerischen Eigenschaften wie z.B. Name, Molekülformel, IUPAC-Name, oder kano-

nischer Smiles-String können als Suchparameter verwendet werden. Der zugehörige Inhalt in Reaktionen wird auf der Grundlage des Suchergebnisses gefiltert. Für Reaktionen sind der Name und die Nummer der Reaktion suchbar.

Die Struktursuche im Gegensatz zur Textsuche kann entweder durch die Suche nach einer Unterstruktur oder einer Ähnlichkeitssuche erfolgen (Bajusz et al. 2015). Diese Suchmethoden sind Fingerprint-basierte Methoden. Innerhalb des ELNs und Repositoriums wurde eine pfadbasierte Fingerabdruckmethode implementiert, die als FP2 (OpenBabel) bezeichnet wird. Dieser Fingerprint ist identisch mit den Daylight Fingerprints (James & Weininger 2006) die als Standard in vielen Publikationen verwendet werden.

Literaturangaben

Winkler-Nees, S. 2013. "Status of Discussion and Current Activities: National Developments". In *Digital Curation of Research Data Experiences of a Baseline Study in Germany*, H. Neuroth, S. Strathmann, A. Oßwald, J. Ludwig (Hrsg.).VWH. Kapitel 2, 18–33.

Coles, S. J., J. G. Frey, C. L. Bird, R. J. Whitby, A. E. Day. 2013. "First steps towards semantic description of electronic laboratory notebook records". *Journal of Cheminformatics* 5 (52).

Rees, I., E. Langley, W. Chiu, S. J. Ludtke. 2013. "EMEN2: an object oriented database and electronic lab notebook". *Microsc. Microanal* 19 (1): 1–10.

Barillari, C., D. S. M Ottoz, J. M. Fuentes-Serna, C. Ramakrishnan, B. Rinn, F. Rudolf. 2016. "openBIS ELN-LIMS: an open-source database for academic laboratories. *Bioinformatics* 32 (4): 638–640.

Rubacha, M., A. K. Rattan, S. C. J. Hosselet. 2011. "A review of electronic laboratory notebooks available in the market today". *Lab. Autom.*16 (1): 90–98.

Day, A. E., S. J. Coles, C. L. Bird, J. G. Frey, R. J. Whitby, V. E. Tkachenko, A. J. Williams. 2015. "ChemTrove. Enabling a Generic ELN To Support Chemistry through the Use of Transferable Plug-ins and Online Data Sources". *Journal of Chemical Information and Modeling* 55: 501–509.

Willoughby, C., C. L. Bird, S. J. Coles, J. G. Frey. 2014. "Creating Context for the Experiment Record. User-Defined Metadata: Investigations into Metadata Usage in the LabTrove ELN". *Journal of Chemical Information and Modeling* 54 (12): 3268–3283.

Milsted, A. J., J. R. Hale, J. G. Frey, J. G. C. Neylon, C. 2013. "LabTrove: A Lightweight, Web Based, Laboratory 'Blog' as a route towards a Marked Up Record of Work in a Bioscience Research Laboratory". *PLOS One* 8: e67460. Online verfügbar unter: https://doi.org/10.1371/journal.pone.0067460. Zuletzt geprüft am 12.08.2017.

Rudolphi, F., L. J. Goossen. 2012. “Electronic laboratory notebook: the academic point of view”. *Journal of Chemical Informations and Modeilng* 52 (2) 293–301.

Kelly, R., R. Kidd. 2015. “Editorial: ChemSpider - a tool for Natural Products research”. *Nat. Prod. Reports* 32: 1163–1164.

Bajusz, D, A. Rácz, K. J. Héberger. 2015. „Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?“. *Journal of Cheminformatics* 7 (20).

James, C.A., D. Weininger. 2006. Daylight theory manual: Daylight Chemical Information Systems.

Chancen und Herausforderungen im Zusammenspiel von Forschungsdatenmanagement und Forschungsinformationssystemen

Dr. Reingis Hauck[1], Dr. Sandra Broll[2]

1,2 Dezernat Forschung und EU-Hochschulbüro und Technologietransfer, Leibniz Universität Hannover

Zusammenfassung. Erfolgreiches Forschungsdatenmanagement (FDM) hängt im großen Maße von der Bereitstellung der entsprechenden IT-Infrastruktur ab, die Wissenschaftlerinnen und Wissenschaftler im Forschungsprozess angemessen unterstützt. Hierbei ist nicht nur auf technische Aspekte zu achten, sondern auch auf wissenschaftsfreundliche Workflows, die darauf Rücksicht nehmen, dass eine Vielfalt von IT-Systemen im Alltag des Hochschullebens zu bedienen ist (Simons et al. 2016). Im Folgenden sollen die Chancen und Herausforderungen im Zusammenspiel von Volltextrepositorium, Datenrepositorium und Forschungsinformationssystem exemplarisch am Beispiel der Leibniz Universität Hannover (LUH) beleuchtet werden.

Die LUH führt zurzeit ein Forschungsinformationssystem (Pure) ein, das Informationen zu den Forschungsaktivitäten ihrer Mitglieder zusammenführt. Diese sogenannten Forschungsinformationen umfassen neben Metadaten zu (Text-)Publikationen und Projekten auch Metadaten zu publizierten Forschungsdatensätzen. Parallel zur Einführung des Forschungsinformationssystems wurde der Aufbau eines institutionellen Volltextrepositoriums umgesetzt. Im Sommer 2016 wurde der Aufbau eines institutionellen Forschungsdatenrepositoriums als Teil des institutionellen Konzepts zum Forschungsdatenmanagement an der LUH beschlossen. In diesem Zusammenhang wird die Frage der Schnittstelle zwischen Daten- und Volltextrepositorium sowie zum Forschungsinformationssystem im Hinblick auf wissenschaftsfreundliche Workflows aufgegriffen. Die Datenpublikation soll nicht als weiterer administrativer Overhead von Wissenschaftlerinnen und Wissenschaftlern verbucht werden müssen. Neben dem Primat der Eingabefreundlichkeit gilt, dass Forschungsinformationssysteme per definitionem das zentrale Berichtstool für das Forschungsmanagement darstellen. Auf institutioneller Ebene besteht das Interesse, das tatsächliche Verhalten hinsichtlich des Forschungsdatenmanagements mit Zahlen zu hinterlegen, u.a. zu Art und Umfang der Daten und Anzahl der Datenpublikationen.

Eine ganzheitliche Systemarchitektur erfordert, sich entwickelnde Anforderungen mitzudenken. Der Blick nach Großbritannien kann hier weiterhelfen: Sowohl hinsichtlich der Umsetzung von Datenmanagementkonzepten als auch in der Nutzung von Forschungsinformationssystemen besteht dort ein erheblicher Erfahrungsvorsprung, der zu nutzen gilt.

Schlagwörter. Forschungsdaten, Forschungsinformationssysteme, Workflow, Repositorium

Ausgangssituation und Systemarchitektur

In 2011 startete an der Leibniz Universität Hannover das Projekt „Aufbau eines Forschungsinformationssystems und einer Dienstleistungsinfrastruktur zum Digitalen Publizieren“. Das Projekt sollte die Infrastrukturdefizite aufgreifen, die im Bereich von Forschungsinformationen und im Bereich des digitalen Publizierens deutlich geworden waren: Weder existierte eine Hochschulbibliographie noch ein institutionelles Volltextrepositorium.

Die Leibniz Universität Hannover hat sich für das Forschungsinformationssystem Pure der Firma Elsevier entschieden, das Informationen zu den Forschungsaktivitäten der Wissenschaftlerinnen und Wissenschaftler der Hochschule zusammenführt. Diese sogenannten Forschungsinformationen umfassen neben Metadaten zu (Text-)Publikationen und Projekten auch Metadaten zu publizierten Forschungsdatensätzen. Das Forschungsinformationssystem synchronisiert soweit wie möglich Daten aus bestehenden Systemen wie SAP HR (SAP Modul Personalwirtschaft) und dem SAP Folders Management (elektronische Drittmittelakte) als auch aus bestehenden Onlinedatenquellen wie z.B. dem GVK, Scopus oder Crossref .

Forschungsinformationen können durch Verknüpfungen miteinander und zu Personen in Bezug gebracht werden. Diese verknüpften Informationen sollen in einem Portal der interessierten Öffentlichkeit präsentiert und zugänglich gemacht werden. Das Portal soll maßgeblich zu einer besseren Außendarstellung der Universität führen. Darüber hinaus wird das Portal für Webcrawler zugänglich sein und so für die Auffindbarkeit der Forschungsinformationen im Netz sorgen. Vergleichbare Portale wurden bereits an anderen Hochschulen umgesetzt. Die Implementierung des Forschungsinformationssystems Pure wird an der Leibniz Universität Hannover voraussichtlich im Herbst 2017 beendet sein. Ein Volltextrepositorium (DSpace) ging Ende 2015 in den Produktivbetrieb.

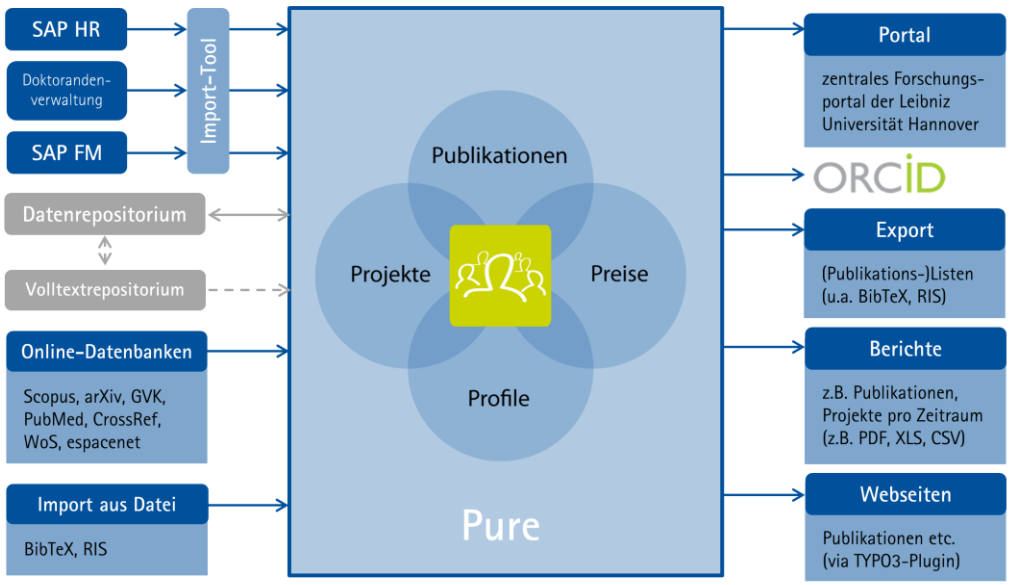


Abbildung 1. Forschungsinformationssystem Systemarchitektur (Eskera 2017)

Ein Projekt zur Konzeptionierung eines Services zum Forschungsdatenmanagement startete in 2015 mit einer universitätsweiten Umfrage zum Umgang mit Forschungsdaten (Hauck et al. 2016). Seit Herbst 2016 wird das erarbeitete institutionelle Konzept zum Forschungsdatenmanagement umgesetzt, das neben dem Ausbau eines integrierten Beratungs- und Schulungsservices zum Thema FDM, den Aufbau eines Datenrepositoriums (CKAN) vorsieht.

Im Rahmen des Aufbaus des Datenrepositoriums wird nicht nur die Frage der Schnittstelle zum Volltextrepositorium sondern auch zum Forschungsinformationssystem im Hinblick auf wissenschaftsfreundliche Workflows aufgegriffen. Die Datenpublikation soll nicht als weiterer administrativer Overhead unter Wissenschaftlerinnen und Wissenschaftlern verbucht werden: Ohne großen Aufwand sollte möglichst sowohl die Datenpublikation veröffentlicht als diese auch mit der Volltextpublikation, dem Projekt, in dessen Rahmen die Daten entstanden ist, und ggf. auch der Infrastruktur (Geräte) verknüpft werden.

Für das weitere Vorgehen erwies es sich als vorteilhaft, dass beide Projekte federführend in einer Abteilung (hier Forschungs- und Transferservice) angesiedelt sind. Das Thema der Systemarchitektur wurde als eigene Teilaufgabe mit weiteren zugeordneten Arbeitspaketen (Workflowanalyse, Schnittstellen, Seamless Integration, administrative Berichtsanforderungen) definiert, die sowohl im Implementierungsprojekt des Forschungsinformationssystems als auch im Umsetzungsprojekt für das institutionelle Forschungsdatenmanagement verankert ist.

Ergebnisse der Workflowanalyse

Im Arbeitspaket Workflowanalysen werden mögliche Workflows im Zusammenhang mit der Datenpublikation aufgenommen und auf Vereinfachungen und Anpassung von Schnittstellen analysiert. Darüber hinaus sollen hieraus auch Ansprüche bezüglich der Gestaltung des Backends der Anwendungen abgeleitet werden. Die Priorisierung der resultierenden Aufgaben soll in eine Roadmap münden.

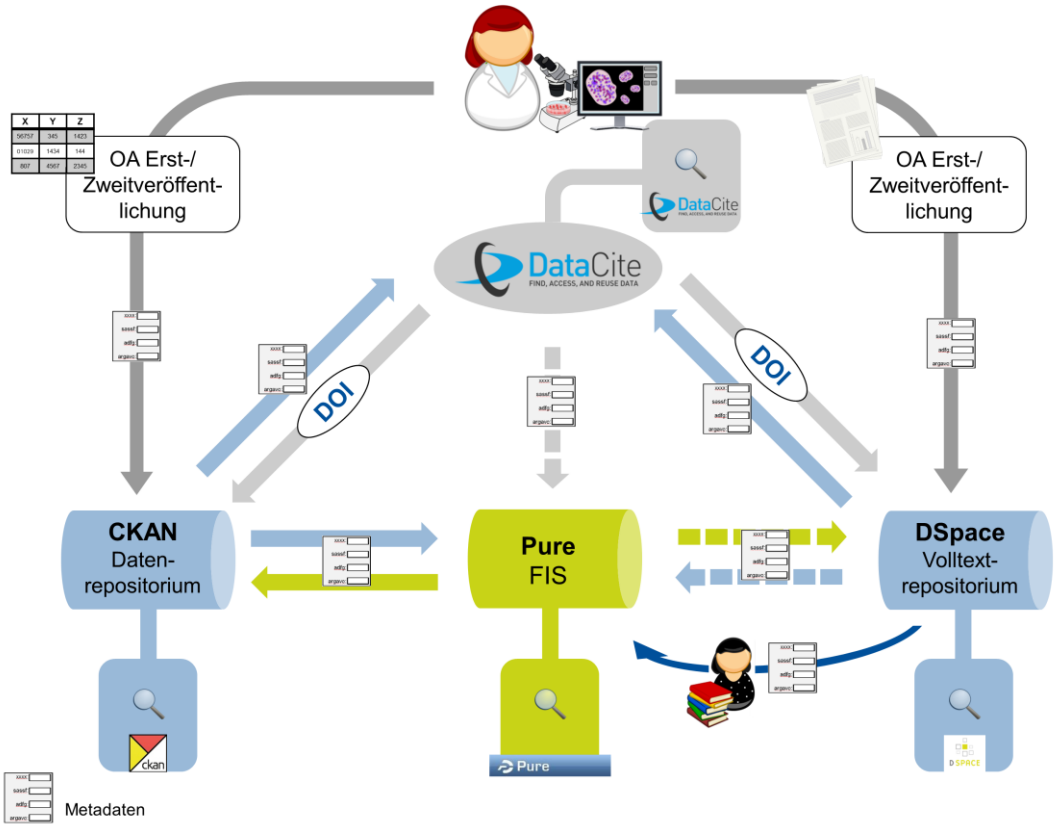


Abbildung 2. Workflowanalyse (Broll 2017)

Abbildung 2 dient zur Veranschaulichung des Workflows der Publikation eines Open Access Volltextes in Verbindung mit der Publikation eines Datensatzes-.und verdeutlicht das komplexe Zusammenspiel zwischen Wissenschaftler/innen, Repository Manager, den Repositories und dem Forschungsinformationssystem. Insgesamt wurden in der Bestandsaufnahme über neun Zwischenschritte im Workflow identifiziert. Im Folgenden werden einige Aspekte näher betrachtet:

Verbindung Forschungsinformationssystem und Volltextrepositorium

Es gibt keine Möglichkeit der automatischen Synchronisation von Metadaten aus dem Volltextrepositorium nach Pure. Der Übertrag der Metadaten nach Pure erfolgt im Hintergrund durch den Repository Manager. Für die weitere Ausgestaltung des Workflows sind zeitliche Verzögerungen zu berücksichtigen, so kann z.B. die Prüfung der rechtlichen Randbedingungen der Zweitveröffentlichung eine Zeitverzögerung darstellen ebenso wie die Qualitätsprüfung des Uploads des Datensatzes.

Verbindung Forschungsinformationssystem und Datenrepositorium

Das Forschungsinformationssystem Pure bietet die Möglichkeit, Metadaten zu Forschungsdaten regelmäßig per xml-Datei zu synchronisieren. Um die Verknüpfung des Datensatzes mit einer Publikation bereits an dieser Stelle zu ermöglichen, soll der Webservice des Forschungsinformationssystems genutzt werden, die in Pure vorhandenen Metadaten zu den eigenen Publikationen in das Datenrepositorium zu übermitteln, um dort dem jeweiligen Autor bzw. Autorin für eine mögliche Verknüpfung angezeigt zu werden.

Die Ausgestaltung der Systemhinweise an die Wissenschaftlerin bzw. den Wissenschaftler im Sinne der Begleitung des Publikationsworkflows muss auf zeitliche Verzögerungen Rücksicht nehmen und die Wiederaufnahme des Workflows vereinfachen bzw. zu geeigneten Zeitpunkten wieder anstoßen.

Aus Abbildung 2 geht hervor, dass die Anbindung von DataCite als Quelle für Metadaten die Befüllung des Forschungsinformationssystem deutlich vereinfachen würde. Wissenschaftlerinnen und Wissenschaftler werden (und sollten), wenn möglich Fachrepositorien nutzen, so dass DataCite als Importquelle für Metadatensätze ins Forschungsinformationssystem von hoher Bedeutung ist. Die Leibniz Universität Hannover wird sich in der Pure User Group entsprechend hierfür engagieren und es gibt bereits erste Hinweise auf den Anschluss dieser Quelle bis Februar 2018.

Ausblick

In der Ausgestaltung der Systemarchitektur müssen außerdem die Anforderungen der administrativen Berichtspflichten berücksichtigt werden. Auf institutioneller Ebene besteht das Interesse das tatsächliche Verhalten im Forschungsdatenmanagement jenseits der qualitativen Aussagen in Befragungen besser zu verstehen und mit Zahlen zu hinterlegen, u.a. zu Art und Umfang der Forschungsdaten, Anzahl der Datenpublikationen, Zitationen, Orte der Publikation, beantragte FDM-relevante Drittmittel, sowie die Anwendung von Datenmanagementplänen in Projekten. Hiermit kann der erhoffte Kulturwandel im Forschungsdatenmanagement bewertet werden, und die Hoch-

schulleitung kann bei Bedarf steuernd in die Entwicklung der Datenmanagementservices eingreifen. Bei der weiteren Gestaltung der Systemarchitektur muss daher überprüft werden, an welcher Stelle diese Daten generiert werden und für die Berichterstattung zusammengezogen werden können (Khokhar et al. 2016). In Großbritannien wurde hierfür die Software DMAonline entwickelt. Auch die Einbindung eines webbasierten Tools zur Datenmanagementplanung wie z B. DMPonline muss in diesem Rahmen auf die Vereinfachung von Workflows bzw. Vermeidung von Mehrfacheingaben überprüft werden.

Eine ganzheitliche Systemarchitektur erfordert Voraussicht, die sich entwickelnde Anforderungen mitdenkt und ihnen gegenüber anschlussfähig bleibt. Der Blick nach Großbritannien kann hier weiterhelfen: Sowohl hinsichtlich der Umsetzung von Datenmanagementkonzepten als auch in der Nutzung von Forschungsinformationssystemen besteht ein erheblicher Erfahrungsvorsprung, der zu nutzen gilt.

Literaturangaben

Hauck, Reingis; Kaps, Reiko; Krojanski, Hans Georg; Meyer, Anneke; Neumann, Janna und Soßna, Volker.2016. Der Umgang mit Forschungsdaten an der Leibniz Universität Hannover : Auswertung einer Umfrage und ergänzender Interviews 2015/16. Hannover: Institutionelles Repositorium der Leibniz Universität Hannover. DOI: https://doi.org/10.15488/265.

Khokhar, Masud, Hardy Schwamm, John Krug and Adrian Albin-Clark. 2016. “UK Data Management Administration Online (DMAOnline)”. Communicating and Measuring Research Responsibly: Profiling, Metrics, Impact, Interoperability": Proceedings of the 13th International Conference on Current Research Information Systems (2016): Procedia Computer Science (2016, In Press). http://hdl.handle.net/11366/499.

Simons, Ed, Mijke Jetten, Marnix van Berchum, Maaike Messelink, Hans Schoonbrood und Marion Wittenberg. 2016. “The important role of CRIS’s for registration and archiving of research data”. Communicating and Measuring Research Responsibly: Profiling, Metrics, Impact, Interoperability": Proceedings of the 13th International Conference on Current Research Information Systems (2016): Procedia Computer Science (2016, In Press). http://hdl.handle.net/11366/524.

Preserving Containers

Klaus Rechert[1], Thomas Liebetraut[2], Stefan Kombrink[3], Dennis Wehrle[4], Susanne Mocken[5], Maximilian Rohland[6]

1,2,4,5,6 University of Freiburg
3 Ulm University

Abstract. Container technology has been quickly adopted as a tool to encapsulate and share complex software setups, e.g. in the domain of computational science. With growing significance of this class of complex digital objects their longevity is also of growing importance. This paper provides a detailed analysis of a container's long-term preservation risks. Based on this analysis, we propose an emulation-based preservation strategy to maintain access to software-based research methods by converting them into a generic archival representation for containers and providing a generic runtime environment.

Keywords. containers, long-term preservation, emulation

Introduction

Modern science is almost inconceivable without computer technology and sophisticated software, which is leading to novel research methodology. In particular, software is able to encapsulate a significant part of research, and thus, can be a crucial element of a research project's outcome by accompanying published articles and published data sets.

Computational science communities have already recognized the need for reproducible compute-based research results to improve scientific practice and to create sustainable results.[1] While publication and citation of research data has made progress recently (Callaghan 2014, Altman et al. 2015), management of software-based research methods remains an open challenge. "Software is a critical part of modern research and yet there is little support across the scholarly ecosystem for its acknowledgement and citation." (Katz, Niemeyer and Smith 2016). Concepts and practice of software citation are currently discussed [2] [3], but software also needs to be available, i.e. retrievable from a dedicated software-archive. [4] [5] Currently, several projects are working on software preservation concepts and services.

However, if software reproducibility is defined as "the ability for someone to replicate a computational experiment that was done by someone else, using the same software and data, and

1 Supercomputing 2017 Reproducibility Initiative, http://sc17.supercomputing.org/2017/02/07/submitting-a-technical-paper-to-sc17-participate-in-the-sc17-reproducibility-initiative/(online, version of Feb 22, 2017)
2 National Institutes of Health (NIH), https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-015.html
3 https://danielskatzblog.wordpress.com/2017/02/22/creation-publication-and-citation-issues-in-software-citation-versus-paper-and-data-citation/
4 Software Preservation Network http://www.softwarepreservationnetwork.org
5 UNESCO PERSIST Initiative, https://www.unesco.nl/digital-sustainability

then to be able to change part of it (the software and/or the data) to better understand the experiment and its bounds[.]"[6] then just availability of software is usually not sufficient. A software-based research process may contain multiple individual software components. And even with detailed documentation - if available at all - manually rebuilding a complex software setup is a laborsome and error-prone process, in particular because (implicit) operational knowledge is lost over time. Additionally, all of the preserved software's dependencies also need to remain available, e.g. operating system, libraries, build environment and a suitable hardware platform are necessary to run the whole software setup. Hence, reproducible computational science requires a portable, self-contained software setup and additionally, a defined runtime environment.

Driven by the demand for reproducible computational science, virtualization and container technologies have been adopted quickly by researchers (Meng et al. 2015, Boettinger 2015). Containers and virtual machines (VMs) are able to encapsulate a software environment, for instance, a complex software tool-chain including application-specific settings, into a single portable entity. They allow researchers to develop, prepare and test a complex software setup locally and to deploy this setup without additional effort in Cloud or HPC facilities. Hence, the configured environment can be shared and (re-)used independently of its original creation environment. Compared to VMs, low computational overhead is one of the main reasons for the rising popularity of containers.

Hence, containers already offer a set of features convenient for the preservation of (complex) software setups, e.g. portability. However, popular container implementations have been criticized for their poor backward compatibility.[7] To make scientific methods accessible, usable and citable in the long-term, the longevity of containers themselves needs to be ensured. This paper provides a detailed analysis on the archivability constraints of containers. Based on this analysis, we propose a preservation strategy to maintain access to software-based research methods by converting them into a generic archival representation for containers and providing a generic runtime environment.

Containers - A New Class of Digital Objects with Preservation Risks?

Container technology (also called operating system virtualization) is promoted as a lightweight alternative to virtual machines, requiring less resources while maintaining portability. While virtual machines provide a virtual "partitioning" of physical resources, containers leverage the abstraction of the underlying operating system to provide exclusive environments for individual software setups. Thus, container technology virtualizes operating system features like starting and running applications, filesystem access or network connectivity, i.e. virtualizing their corresponding programming interfaces (operating system APIs).

For Linux-based container implementations, the virtualization of programming interfaces is built on a specific feature set of the Linux kernel that provide isolated environments - *namespaces* (Biederman 2006) and (process) control groups (*cgroups*) (Menage 2007). *Namespaces* allow users to create multiple isolated views on the operating system's interfaces. Anything running inside a namespace is separated from other *namespaces*, e.g. other containers or the underlying

6 https://danielskatzblog.wordpress.com/2017/02/07/is-software-reproducibility-possible-and-practical/
7 https://thehftguy.com/2016/11/01/docker-in-production-an-history-of-failure/

host system. *cgroups* are able to control and limit access to system resources shared with the host and among other containers.

Dependency Analysis

Due to the strict isolation, containers need to be self-contained, i.e. all software dependencies (libraries, applications etc.) have to be included within the container. Therefore, the main component of a container is a self-contained filesystem with installed and configured software components. The only remaining external or unresolved software dependency is the underlying operating system, i.e. the operating system's application binary interface (ABI).

The second container component is its runtime-configuration defining, for instance, filesystem mappings, e.g. shared folders between container and its host system, and the definition of an entry-point within the container (i.e. a script or program).

A container's *technical runtime* is composed of two components: a hardware component (the computer) and a software component. In general, the software component represents typically a basic Linux installation with an installed and configured (vendor specific) container runtime.

Furthermore, the isolation of a container simplifies the determination of (hardware) dependencies. For instance, applications running within a container usually have only a very abstract view on available hardware of the system, i.e. having only access to individual files instead of a raw hard disk. Direct hardware access has to be explicitly allowed for by *mapping* entries of the */dev*-subsystem into a container's namespace. Fig. 1 illustrates the full container software stack.

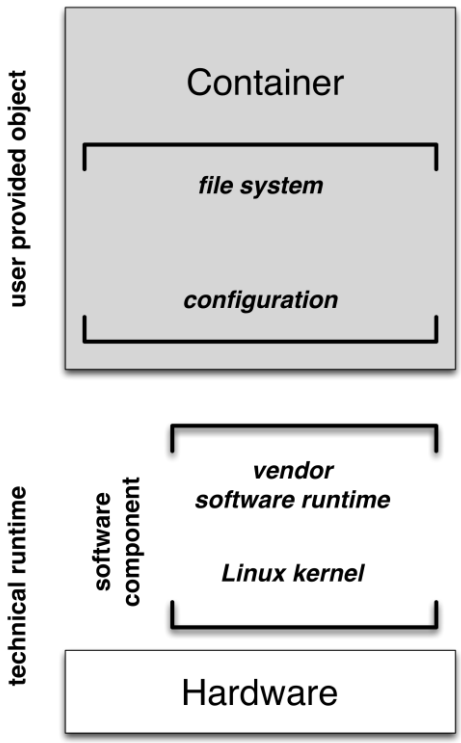


Figure 1. Container components and its technical runtime.

Unfortunately, in addition to these explicit dependencies, there are two further classes of (implicit) dependencies to be investigated. The first class is defined by specific dependencies of the com-

piled software binary. This includes dependencies on specific kernel features, because the software calls certain functions which may only be present from a certain kernel release. Similarly, software binaries depend on a CPU instruction set and highly optimized scientific software may require a very specific CPU version. In a long-term perspective, however, these risks seem manageable, as long as newer kernel versions and newer CPU generations remain available as an emulation environment. A second class of (implicit) dependencies are the containers' *expectations* on an external (technical) environment, e.g. the availability of network services such as licensing servers or data to be available at a certain remote server. This kind of (implicit) external dependencies pose not only the highest preservation risk for an individual container, but require specific effort and infrastructure to be identified and are expensive to monitor. However, every explicit or implicit dependency not only reduces the chances of successful long-term preservation but also limit the portability of a container. Hence, it should be also in the research communities' best interest to maintain external dependencies.

Preservation Risks

The aforementioned properties are not only useful to share software-based research methods but also provide a solid foundation for long-term preservation. Portability, defined/limited external dependencies and a common runtime environment provide the basis for developing a preservation strategy. Furthermore, preservation planning of containers and their technical runtime can be decoupled: while preservation planning of individual containers focuses on determining and monitoring external dependencies (license servers, data dependencies, etc.), preservation planning of the technical runtime focuses on keeping the Linux container runtime usable by replacing outdated hardware with virtual hardware.

Even though containers build on the same technical foundations, currently popular container implementations such as *Docker*[8], *Singularity*[9] or *Shifter*[10] use different container representations and configuration formats. These representations require a specific technical runtime to be used. But ideally, only a small number of software runtimes are necessary to run a large number of containers, hence a common container representation and especially configurations reduces the preservation risk to the availability of a common software runtime. If we assume that the operating system interface rarely changes in incompatible ways, only a few variations of such runtimes are necessary.

With regards to long-term archival, we can assume that a software runtime instance is fixed and achievable. Hence, the long-term preservation risk of containers can be reduced to the availability of hardware required to host the software runtime, i.e. providing necessary hardware supported by the Linux kernel version used as software runtime. When compatible physical hardware becomes unavailable, virtual hardware can be used by means of emulation. In this case, specific properties of containers come in handy again. Typically, software running within containers has no access at all or only a very limited one to the underlying hardware (see discussion above). Furthermore, Linux operating systems support a wide range of hardware. Both properties ease the

8 https://www.docker.com/
9 http://singularity.lbl.gov/
10 http://www.nersc.gov/research-and-development/user-defined-images/

process of replacing physical hardware by emulators. A generic emulation-based preservation strategy is described by Rechert, Falcao and Emson 2016.

Implementation and Preliminary Results

In order to be able to reduce the containers' preservation risks to a generic emulation strategy, container and their runtime require a common archive format, such that for an archive's perspective all container "objects" share the same technical properties, i.e. technical dependencies. For this it is necessary to

1. define a common archive representation of container, as well as provide tool support to ingest and convert container "flavours" into this common format;
2. develop a common runtime for the common container archive format and keeping this runtime long-term available and usable.

Our development builds on the existing Emulation as a Service (EaaS) framework (Liebetraut et al. 2014, Rechert et al. 2012), which provides abstract emulation-based services, the foundation for an emulation-based preservation strategy, and is able to allocate and utilize compute resources in public or private Cloud services. Within the EaaS framework we distinguish between a rendering environment and a digital object to be rendered. A rendering environment is typically represented by a hardware emulator and a software environment (operating system and installed applications) and is able to render a certain class of digital objects. Digital objects are stored and maintained independently and are assigned either manually or in an automated way to a suitable emulation environment (Rechert et al. 2015). To make use of this model, in a first step, a suitable EaaS rendering environment for container has to be defined and implemented.

Integrating a Container Runtime

The Open Container Initiative (OCI)[11] is a governance body formed to create an industry standard for container technology. As part of their work a runtime specification (*runtime-spec*[12]), an image specification (*image-spec*[13]) and a reference runtime implementation (*runC*[14]) have been released.

For preservation purposes, we focus on the runtime-spec and the reference implementation *runC* as a generic container runtime. We have implemented an EaaS rendering environment for containers based on *runC*.

In a second step we have implemented an object *ingest* process for *Docker* and *Singularity* containers. The result of this process is a representation of the the container's file system and a runtime configuration. In OCI terms this is called *filesystem bundle*[15]. Of course, through converting a specific container format into a generic one, information is typically lost, for instance the container creation history represented by *Docker's* image layers. However, the resulting flat con-

11 Open Container Initiative (OCI), http://www.opencontainers.org
12 https://github.com/opencontainers/runtime-spec
13 https://github.com/opencontainers/image-spec
14 https://github.com/opencontainers/runc
15 https://github.com/opencontainers/runtime-spec/blob/master/bundle.md

tainer state is identical with the state of a specific container at the moment of execution, i.e. the container's filesystem is fully assembled, unpacked and usable. Extracting and archiving additional meta-data, e.g. the original `Dockerfile`, which documents the creation process, has to be addressed separately.

A basic runtime configuration (`default.json`) has been pre-configured and will be enriched by the conversion process. Furthermore, defaults can be altered to satisfy the needs of different host environments. For instance, specific namespaces can be set up or specific devices can be mounted.

Example Docker

The conversion process for *Docker* uses the `docker export` command to extract the container's file system. The conversion process further uses `docker inspect` to determine the container's entry point and environment variables, specified during the creation of the container. For instance, the official *MariaDB* image from the *Docker Hub* provides a `docker-entrypoint.sh` as entry point and adds `mysqld` as command that can be altered during the creation of a container, resulting in the final command line "`docker-entrypoint.sh mysqld`". If no entry point is defined, the default entry point is set to `/bin/sh`. The `docker export` command only exports the root file system of a container, mounted volumes are not included in the generated archive. Therefore, it is necessary to archive volumes separately. Changes that are made during the runtime of the *filesystem bundle* are stored in the root file system by default. If a volume does contain data needed for the execution of the bundle, e.g. configuration or research data, it has to be mounted manually in the `config.json`. Currently, the conversion process is not able to recreate the network environment of converted containers. The basic configuration does not utilize network namespaces, such that contained applications share the network configuration with the host machine. If it is desired to isolate the network, appropriate namespaces can be set up. This shortcoming might be an issue especially if the container to be archived is part of a multi-container environment where some containers are linked with others, e.g. an application with a database backend. These links are not part of the Open Container Specification and are set up by the *Docker* Daemon. To recreate the functionality of such containers it is necessary to rebuild the needed network environment and provide the *runC* bundle with a suitable /etc/hosts file. Alternatively, the application configuration has to be changed, which would imply a content-related analysis of the container.

Example Singularity

Similar to the *Docker* conversion process, the configuration is generated based on the same default.json, which is altered to satisfy the needs of different host environments. The command for the *runC* configuration is *Singularity's* `/.run.sh` script, which not only contains the command to be executed but also sets the environment variables defined by the author. If the script is not present, a shell `(/bin/sh)` is used as command.
Because of *Singularity's* focus on portability images created with it already contain all information needed for a successful execution. As described above, the run script sets up the environ-

ment of the contained application. Setups consisting of multiple, linked containers are not intended, since Singularity does not utilize network namespaces. Because of this, a correctly authored *Singularity* image should pose no problem. Successful migration from an bare-metal HPC setup into a cloud-computing environment has been shown using a singularity container of *GROMACS*. The container was originally designed for the JUSTUS HPC Compute Cluster[16].

Access

A container prepared in the aforementioned way can then be accessed using the EaaS framework. Containers map directly to rendering environments in the EaaS terminology and can be started just like any other environment using the provided EaaS REST API by posting a JSON request similar to the following:

```
HTTP POST "http://emulation.solutions/api/components"
{
   "containerId": "gromacs"
}
```

Because the EaaS framework's core component is an abstract rendering environment, containers fall into the same category. The selected environment in this case is the *GROMACS* container which automatically starts its computation when it is invoked and prints results to the standard console output. The framework will return a session ID, which enables allows the user to control the session and send further related requests.

In order to provide different types of access, each component in the EaaS framework has a set of so-called control URLs. Once a session is initiated and a session ID has been retrieved, a control URL can be requested for this session. Depending on the session type the user is either able to interact with the running instance (HTML5) and/or is able to retrieve the container's output.

Since containers usually provide no interactive user interface, we have implemented a *headless* access method, which invokes the container's entrypoint and retrieves output from standard- and error-out output (*stdout* and *stderr*). Both outputs are available as zipped text files. Consequently, in order to retrieve the container's result, the component's control URL. This URL can be downloaded e.g. using a regular web browser at any time. Data transfer does not start until the container's computation is finished. Once downloaded, the ZIP archive contains the mentioned output text files. This mechanism can later be extended to also put a certain directory into the archive where the contained application is known to store its results. A network-based live-interaction mode is currently work in progress. Furthermore, research data needs to be integrated, ideally directly via DOI references.

The presented REST API can be used completely automated, which can also serve as a verification tool to assert that a container ingested into the EaaS system produces the same results as the original container.

16 JUSTUS Compute Cluster, https://www.uni-ulm.de/einrichtungen/kiz/service-katalog/high-performance-computing/justus/

Discussion & Outlook

Containers are and will be interesting research topic in the domain of digital preservation and in particular, reproducible science, not only because of their widespread use but also because of their technical characteristics. This initial study suggests, that if the long-term preservation of containers focuses on their technical runtime, e.g. by structured archival of the software runtime, the preservation risks can be reduced to a generic emulation strategy, which seem to be manageable in a (cost) efficient way (compared to the number of objects to be preserved). However, preserving containers is only a complementary strategy to software archiving and not a substitute. Containers are able to capture a complex software-based research method with built-in quality control (completeness), but for individual re-use of software components an independent software archive strategy is necessary.

While this work provides the necessary conceptual and technical groundwork for preserving containers, a main success factor lies still within the scientific community. Research communities need to agree on usage and access pattern to support interoperability of their research methods. Common, documented access to external services, creates awareness of external preservation risks. This should also include maintaining external dependencies.

References

Altman, Micah, Christine Borgman, Mercè Crosas, and Maryann Matone. 2015 "An introduction to the joint principles for data citation." *Bulletin of the American Society for Information Science and Technology* 41 (3): 43-45.

Callaghan, Sarah. 2014. "Joint declaration of data citation principles." *Data Citation Synthesis Group: FORCE11*.

Smith, Arfon M., Daniel S. Katz, and Kyle E. Niemeyer. 2016. "Software citation principles." *PeerJ Computer Science* 2: e86.

Katz, Daniel S., Kyle Niemeyer, and Arfon M. Smith. 2016. "Strategies for biomedical software management, sharing, and citation." *PeerJ Preprints* 4: e2640v1.

Meng, Haiyan, Rupa Kommineni, Quan Pham, Robert Gardner, Tanu Malik, and Douglas Thain. 2015. "An invariant framework for conducting reproducible computational science." *Journal of Computational Science* 9: 137-142.

Boettiger, Carl. 2015. "An introduction to Docker for reproducible research." *ACM SIGOPS Operating Systems Review* 49 (1): 71-79.

Biederman, Eric W., and Linux Networx. 2006. "Multiple instances of the global linux namespaces." In *Proceedings of the Linux Symposium*, vol. 1: 101-112.

Menage, Paul B. 2007. "Adding generic process containers to the linux kernel." In *Proceedings of the Linux Symposium*, vol. 2: 45-57.

Goecks, Jeremy, Anton Nekrutenko, and James Taylor. 2010. "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences." *Genome biology* 11 (8): R86.

Rechert, Klaus, Patricia Falcao, and Tom Emson. 2016. “Towards a Risk Model for Emulation-based Preservation Strategies: A Case Study from the Software-based Art Domain.” In *Digital Preservation (iPRES) 2016 13th International Conference on*: 139 - 148.

Liebetraut, Thomas, Klaus Rechert, Isgandar Valizada, Konrad Meier, and Dirk Von Suchodoletz. 2014. "Emulation-as-a-Service-The Past in the Cloud." In *Cloud Computing (CLOUD), 2014 IEEE 7th International Conference on*: 906-913. IEEE.

Rechert, Klaus, Isgandar Valizada, Dirk von Suchodoletz, and Johann Latocha. 2012. "bwFLA – A functional approach to digital preservation." *PIK-Praxis der Informationsverarbeitung und Kommunikation* 35 (4): 259.

Rechert, Klaus, Thomas Liebetraut, Oleg Stobbe, Isgandar Valizada, and Tobias Steinke. 2015. "Characterization of CD-ROMs for Emulation-based Access." *Digital Preservation (iPRES) 2015 12th International Conference on* (2015): 144.

Skalierbare und flexible Arbeitsumgebungen für Data-Driven Sciences

Dennis Schridde[1], Martin Baumann[2], Vincent Heuveline[3]

1,2,3 Universitätsrechenzentrum, Universität Heidelberg

Abstract. In vielen wissenschaftlichen Fachdisziplinen stellt die IT-gestützte Analyse und Exploration erzeugter Daten einen wesentlichen Schritt zum Erkenntnisgewinn dar. Aufgrund des technischen Fortschritts, z.B. in Bezug auf Messungen wie auch auf computergestützte Simulationen, nimmt die auszuwertende Datenkapazität rasant zu. Gleichzeitig steigt die Komplexität der eingesetzten Analysemethoden. Beides zusammen führt zu einer enormen Herausforderung, die durch die IT-Systeme erfüllt werden muss.

In der Praxis werden häufig Daten auf zentralen IT-Systemen erzeugt und gespeichert, jedoch auf lokalen Workstations analysiert und visualisiert. Die Übertragung der umfangreichen Daten führt zu langen Wartezeiten. Außerdem steht lokal meist nicht die gleiche Rechenleistung wie in zentralen IT-Systemen an Rechenzentren zur Verfügung. Andererseits haben zentrale Rechensysteme oft eine hohe Einstiegshürde.

Der im Projekt bwVisu verfolgte Ansatz setzt auf Virtualisierungstechniken, die es Nutzern ermöglicht, die individuell benötigten Werkzeuge zur interaktiven Visualisierung auf einer leistungsfähigen und skalierbaren Infrastruktur zu betreiben. Nutzer können Community-spezifische Arbeitsumgebungen einsetzen, weiterentwickeln und diese wiederum mit anderen teilen. Die Bedienung der Infrastruktur erfolgt ebenfalls über eine graphische Nutzeroberfläche. Die vorgestellte Lösung ist an einen lokalen Speicherdienst für wissenschaftliche Daten angebunden, sodass Forscher in ihrer individuellen Arbeitsumgebung sehr schnell auf die Daten ihrer Projekte zugreifen können und diese sehr performant analysieren und visualisieren können.

Schlagwörter. Visualisierung, Remote-Visualisierung, Landesdienst

Einführung

Motivation und Bedarf

In vielen Fachbereichen stellt die Erhebung und Auswertung großer wissenschaftlicher Daten einen essentiellen Bestandteil im Forschungsvorgang dar. Auf Grund des technischen Fortschritts können immer schneller immer größere Datenmengen erzeugt und gespeichert werden, beispielsweise durch verbesserte Sensorik (z.B. Mikroskope oder 3D-Scanner) oder schnellere Hochleistungsrechner.

Für die Analyse solcher Daten ist häufig eine visuelle Darstellung mit interaktiver Steuerung erforderlich, die flexibel auf die jeweiligen Datensätze und Fragestellungen angepasst werden kann. Beispielsweise können komplexe dreidimensionale Strömungsvorgänge mit geeigneten Visualisierungstechniken sehr einfach erfasst und gewisse Phänomene (z.B. Wirbelstrukturen)

lokalisiert werden. Die Verarbeitung erfolgt meist in mehreren Arbeitsschritten und Phasen, in denen quantitative und qualitative Analysen durchgeführt werden (Upson et al. 1989).

Allerdings stellt die interaktive Arbeit mit solchen sehr umfangreichen Daten eine Herausforderung für die Forscher dar, da deren Übertragung auf die Workstations der Nutzer aufgrund der Größe nicht praktikabel ist: Die Länge des Kopiervorgangs steht oft in ungünstigen Verhältnis zur späteren Betrachtungsdauer oder die lokale Speicherkapazität ist nicht ausreichend. Andererseits ist die Rechenleistung von Arbeitsplatzcomputern im Vergleich zu jener von HPC-Rechenknoten häufig deutlich geringer, sodass die verarbeitbare Datenmenge stark eingeschränkt ist und die Berechnungsgeschwindigkeit für eine komfortable Nutzung nicht ausreicht (Stegmaier et al. 2003, Jomier et al. 2011, Mouton, Sons, and Grimstead 2011, Shu and Hsu 2015).

Während es zwar an einigen Universitätsrechenzentren Angebote zum interaktiven Arbeiten auf den HPC-Systemen gibt[1], ist deren Nutzung häufig für die Forscher nicht zufriedenstellend, wie sich in Befragungen ergab. Zum einen hat nicht jeder Forscher Zugriff auf HPC-Ressourcen, zum anderen sind diese nicht immer zeitnah verfügbar. Gleichzeitig verfügt nicht jeder Wissenschaftler über ausreichend Erfahrung im Umgang mit HPC-Systemen und könnte diese Systeme als zu kompliziert wahrnehmen, bzw. die Anforderungen an die Arbeitsumgebung sind schwer innerhalb der Nutzungskonzepte bzw. Betriebsmodelle der IT-Systeme umsetzbar. Das typische Betriebsmodell von Hochleistungsrechnern bietet eine vordefinierte Arbeitsumgebung, bestehend aus Betriebssystem und Vorauswahl an Software und Tools, die auf Leistung optimiert wurden. Nutzer können i.d.R. zusätzliche Software für die eigene Verwendung ergänzen, manuell oder mittels eines Modul-Systems, was jedoch gewisse Linux-Kenntnisse erfordert. Die Individualisierung der Arbeitsumgebung erfolgt dabei immer nutzer- und systemspezifisch, d.h. sie muss durch jeden Nutzer auf jedem verwendeten System erneut erfolgen (Mauch et al. 2014). Außerdem sind die Nutzungszeiten bei Hochleistungsrechnern mit entsprechenden Ressource-Management-Systemen auf hohe Auslastung der Systeme über größere Zeiträume optimiert, was eine spontane, interaktive Nutzung erschwert.

Wenn die Netzwerkbandbreite zwischen Datenquelle und dem lokalen Rechner des Forschers nicht genügt oder die Möglichkeiten der lokal am Arbeitsplatz vorhandenen IT-Ressourcen nicht ausreichend sind, werden die zu analysierenden Daten meist vereinfacht oder verringert (z.B. durch Reduktion der Auflösung). Dadurch kann eine Analyse zwar durchgeführt werden, allerdings kann die Datenreduktion eine große Einschränkung darstellen, da beispielsweise wichtige Merkmale nicht mehr erkennbar sein können.

Es liegen also große Datenmengen vor, die analysiert werden sollen. Gleichzeitig ist jedoch der Erkenntnisgewinn daraus aufgrund einer schwer durchführbaren Visualisierung und Analyse teilweise stark eingeschränkt. Eine Arbeitsumgebung für datenintensive Forschungsfelder sollte daher mehreren Anforderungen genügen:

1. **Interaktive 3D-Remote-Visualisierung**: Nur die Visualisierung der Daten (statt der Daten selbst) wird zum Nutzer übertragen. Eine lange Übertragungszeit der Daten kann vermieden werden. Performante Rechenressourcen (CPU und GPU) in der Nähe des Datenstandortes sind erforderlich, um z.B. aufwendige 3D-Visualisierungen zu ermöglichen.

1 Vgl. z.B. https://wickie.hlrs.de/platforms/index.php/Graphic_Environment oder https://www.bwhpc-c5.de/wiki/index.php/Start_vnc_desktop

2. **Nutzerfreundlichkeit**: Einfache Bedienung der Arbeitsumgebung mittels graphischer Oberfläche sollte gewährleistet sein.
3. **Nutzerspezifische Arbeitsumgebungen**: Nutzer können eigene Software in eine Arbeitsumgebung integrieren. Arbeitsumgebungen können innerhalb einer Gruppe weitergeben und gemeinsam weiterentwickelt werden.
4. **Leistung und Skalierbarkeit**: Das Potential der Hardware soll ausgeschöpft werden. Ressourcenintensive Anwendungen sollen auf mehrere Knoten verteilt werden können (horizontale Skalierbarkeit). Außerdem sollte es möglich sein, mehrere Anwendungen mit geringen Ressourcenanforderungen auf einem Knoten auszuführen (vertikale Skalierbarkeit nach unten).

Existierende Lösungen und verwandte Arbeiten

Es gibt verschiedene Projekte, in denen mittels Virtualisierung bzw. Containerisierung flexible wissenschaftliche Arbeitsumgebungen entwickelt werden oder die eine komfortable Steuerung eines Hochleistungsrechners mittels einer Web-Oberfläche ermöglichen. Einige dieser Vorhaben werden nachfolgend vorgestellt.

In der Biologie und den Life Sciences werden teilweise bereits Container-basierte Lösungen eingesetzt, um die Verteilung von Software-Paketen und die Reproduzierbarkeit der Ergebnisse zu unterstützen (siehe z.B. „BioContainers“[2]). Es wurde beispielsweise bereits 2008 die iPlant Collaborative gegründet, in welcher eine Infrastruktur für die Pflanzenanalyse geschaffen wurde, um virtuelle Maschinen und später auch Container auszuführen. Diese wurde 2015 unter dem Namen „CyVerse“[3] einer breiteren Öffentlichkeit zugänglich gemacht. Ziel ist es, den Zugang zu leistungsfähigen IT-Ressourcen und die anschließende Analyse zu vereinfachen, wozu u.a. 2011 eine Cloud-Computing-Plattform namens „Agave“[4] geschaffen wurde. Das primäre Augenmerk liegt darauf, die Prozesse beginnend mit der Speicherung der Primärdaten, über deren Auswertung mittels virtualisierter Software in reproduzierbaren Umgebungen, bis hin zur Visualisierung, in einer einfach zu bedienenden Web-Oberfläche abzubilden. Während dieser Ansatz zwar die Anforderungen 2) bis 4) erfüllt, steht die interaktive 3D-Remote-Visualisierung nicht im Vordergrund. Des weiteren ist diese Plattform stark auf die Forschungsgebiete Life Sciences und Biologie zugeschnitten. Hinzu kommt, dass viele der verwendeten Komponenten Eigenentwicklungen darstellen, für die es inzwischen etablierte Lösungen gibt.

Auch im kommerziellen HPC-Kontext gibt es Bestrebungen, die Job-Verwaltung und Ansteuerung des jeweiligen Cluster-Scheduling-Systems zu vereinfachen und mit Remote-Visualisierungs-Software zu verknüpfen. So integriert die Firma NICE mit ihrem Produkt „EnginFrame“[5] seit einigen Jahren die eigene Visualisierungslösung „DCV“, sowie die Lösungen anderer Hersteller (z.B. HP „RGS“, Citrix „XenDesktop“) sowie Technologien wie „VNC“ und „VirtualGL“ mit verschiedenen Job-Schedulern (Adaptive Computing „Moab“, IBM „LSF“, Univa „GridEngine“, Altair „PBS/Pro“ und „SLURM“) in einer Web-Oberfläche. Auch die Firma

2 https://biocontainers.pro/
3 http://www.cyverse.org/
4 https://agaveapi.co/
5 http://www.nice-software.com/products/enginframe

Adaptive Computing bietet seit ca. 2016 mit ihrem Produkt „Viewpoint“[6] eine Web-Oberfläche an, die die Umsetzung allgemeiner Workflows und Visualisierungsaufgaben auf einem mit dem Adaptive Computing „Moab“ Scheduling-System betriebenen traditionellen HPC-Cluster für den Nutzer vereinfachen soll. Diese beiden Ansätze bauen auf traditionellen HPC-Systemen und entsprechenden Betriebsmodellen auf. Die Flexibilität im Hinblick auf nutzerspezifische Umgebungen ist somit hierdurch festgelegt. Beispielsweise ist die Portabilität einer Arbeitsumgebung auf ein anderes System daran geknüpft, dass alle beteiligten HPC-Cluster bzgl. der zur Verfügung gestellten Software (z.B. Bibliotheken, Betriebssystem, Software-Module, etc.) homogen aufgebaut sind, sodass Nutzersoftware mit diesen HPC-Clustern kompatibel ist.

Im eng mit der wissenschaftlichen Visualisierung verknüpften Scientific-Computing-Kontext gibt es Bestrebungen, verschiedene Cloud-Technologien für wissenschaftliche Berechnungen nutzbar zu machen: So beschäftigt sich das Projekt „Singularity“ mit der Integration von Container-basierter Virtualisierung in traditionelle HPC-Umgebungen (Kurtzer, Sochat und Bauer 2017), bzw. NERSC „Shifter“ mit der Nutzung von Container-Image-basierten Distributionstechniken in HPC-Clustern (Jacobsen and Canon 2015, Bahls 2016). Hierzu muss ein gewisser Entwicklungsaufwand betrieben werden und es ist eine Abweichung von in der Industrie etablierten Techniken nötig. Auf diese Weise sollen die Vorteile der einfachen Verteilung von Software-Paketen auch hier nutzbar werden. Diese beiden Projekte konzentrieren sich auf den Aspekt der nutzerspezifischen Arbeitsumgebung (Anforderung 3).

Ebenso gibt es Bestrebungen in traditionellen HPC-Systemen verwendete Hardware, wie z.B. InfiniBand Netzwerke, in virtuellen Maschinen nutzbar zu machen (Mauch, Kunze und Hillenbrand 2013) und auch im Allgemeinen die Diskrepanz zwischen traditionellen HPC-Systemen und virtuellen Maschinen vor allem in Hinblick auf I/O-Leistung zu reduzieren. So setzte sich das „MIKELANGELO“[7] Projekt (Laufzeit 2015 – 2017) zum Ziel, auf diese Weise eine Nutzung von Plattformvirtualisierungstechniken mit seinen verschiedenen Vorteilen im wissenschaftlichen Kontext zu ermöglichen. Bei diesen Projekten werden sehr spezifische technische Fragestellungen adressiert, die auf die optimale Ausnutzung der Hardware-Leistungsfähigkeit im Kontext der Virtualisierung abzielen (Anforderung 4).

Gleichzeitig entwickelten sich aus Richtung der Cloud-Anbieter, angestoßen durch die Erfahrungen, die Google mit ihrem internen, „Borg“ genannten System sammelte (Verma et al. 2015), verschiedene Umgebungen zur Ausführung allgemeiner Anwendungs-Container. So existieren z.B. seit 2015 das „Datacenter Operating System“ („DCOS“) der Firma Mesosphere oder seit 2014 die „Kubernetes“ genannte Container-Orchestration-Engine der „Cloud Native Computing Foundation“ („CNCF“). Ihr Fokus ist die zeitnahe, horizontal sowie vertikal skalierbare Ausführung beliebiger Anwendungs-Container. Auch diese Projekte konzentrieren sich auf die Anforderungen 3 und 4, bieten selbst jedoch noch keine Lösungen für die Anforderungen 1 und 2.

Unser Ansatz

Keines der zuvor genannten Projekte erfüllt alle der im Kapitel „Motivation und Bedarf“ genannten Anforderungen 1) bis 4), die aus Sicht der Autoren für eine Arbeitsumgebung für datenintensive Forschungsfelder notwendig sind. Daher besteht der Bedarf einer neuen Lösung, die in Form

6 http://www.adaptivecomputing.com/products/hpc-products/viewpoint/
7 https://www.mikelangelo-project.eu/

des Projekts bwVisu adressiert wird. Der im Rahmen dieses Projekts entwickelte Dienst setzt systematisch auf Remote-Visualisierung, wodurch das Kopieren von teilweise sehr umfangreichen wissenschaftlichen Daten vermieden werden kann. Nutzerfreundlichkeit soll durch eine sehr einfache Web-Oberfläche erreicht werden. Der Einsatz von Virtualisierungstechnologien ermöglicht nutzerspezifische Arbeitsumgebungen und Skalierbarkeit. Um die Hardware-Leistung für den Anwender nutzbar zu machen, wird in dem Projekt auf Container-basierte Operating-System-Level-Virtualisierung gesetzt.

In dieser Arbeit stellen wir den Dienst bwVisu vor, erläutern das Konzept und skizzieren die technische Umsetzung. Hierzu gehen wir insbesondere auf die weiter oben genannten Anforderungen an eine Arbeitsumgebung für datenintensive Forschungsfelder ein. Abschließend geben wir einen Ausblick auf mögliche zukünftig Erweiterungen und Verbesserungen.

bwVisu – ein flexibler und skalierbarer Remote-Visualisierungsdienst

Übersicht

Der Visualisierungsdienst bwVisu wird im Rahmen des gleichnamigen Projekts entwickelt, das über die Laufzeit von 2014 bis 2017 vom Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg gefördert wird. Es wurde ein Remote-Visualisierungsdienst für die Darstellung und Analyse von wissenschaftlichen Daten entwickelt, in dessen Mittelpunkt eine private Cloud-Umgebung, ausgestattet mit performanten Servern, Graphikkarten und einem Hochleistungsnetzwerk, steht. Diese wurde am Universitätsrechenzentrum Heidelberg aufgebaut, wo der Dienst auch betrieben wird. Für die Nutzer wurde ein Portal bwVisu-Web zur Ausführung individueller Arbeitsumgebungen entwickelt. Die Entwicklung und der Einsatz von Community-spezifischen Arbeitsumgebungen unterstützt Fortschritte im Bereich der „Data-Driven Sciences“ und fördert den Gedanken von „Open Data“ und „Open Source“. Die Projektpartner untersuchen gemeinsam Methoden zur verteilten wissenschaftlichen Visualisierung und entwickeln entsprechende Softwarepakete weiter, die im Rahmen von bwVisu besonders vielversprechend erscheinen. Geförderte Einrichtungen sind das Universitätsrechenzentrum Freiburg, das Universitätsrechenzentrum Heidelberg, der Lehrstuhl für Computergrafik am Karlsruher Institut für Technologie und das Höchstleistungsrechenzentrum Stuttgart. Also assoziierter Partner fungiert das Heidelberger Institut für Theoretische Studien. Das Vorhaben wird durch das Universitätsrechenzentrum Heidelberg koordiniert.

1. Interaktive 3D-Remote-Visualisierung

Der Ansatz der Remote-Visualisierung sieht vor, dass die Visualisierung von Daten durch Hardware bzw. Systeme erfolgt, die einen sehr schnellen Zugriff auf die Daten ermöglichen. Die Anzeige und interaktive Steuerung der Visualisierung erfolgt jedoch von einem entfernten Rechner (also „remote“), der nur über eine vergleichsweise geringe Bandbreite mit dem Visualisierungssystem verbunden ist. Das Konzept der Remote-Visualisierung wird in vielen Werkzeugen und Umgebungen eingesetzt und ist weit verbreitet (Stegmaier et al. 2003, Jomier et al. 2011, Mouton, Sons, and Grimstead 2011, Shu and Hsu 2015). Durch diesen Ansatz können Personen auch dann

sehr umfangreiche Daten betrachten und analysieren, wenn die verfügbaren lokalen Ressourcen der Arbeitsplatzrechner oder die vorhandene Netzwerkbandbreite eher gering sind. Ein langwieriges Kopieren der zu visualisierenden Daten auf den Arbeitsplatzrechner ist nicht erforderlich. Wesentlich für diese Technologie ist, dass eine leistungsfähige Visualisierungs-Hardware vorhanden ist und diese sehr schnell auf die zu visualisierenden Daten zugreifen kann.

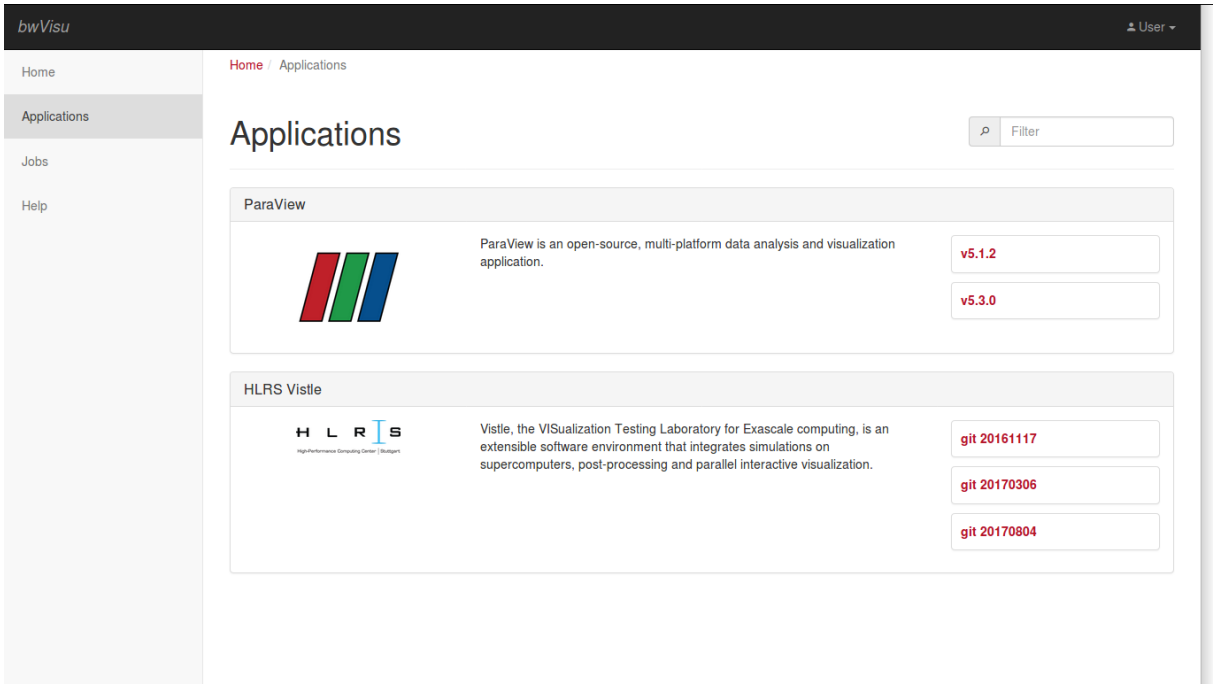


Abbildung 1. bwVisu-Web bietet dem Nutzer eine graphische Auswahl der vorhandenen Anwendungs-Software in verschiedenen Versionen mit Beschreibungstext.

Insbesondere große wissenschaftliche Daten liegen häufig bereits auf zentralen Datenspeichern vor, z.B. auf dem Speicherdienst „SDS@hd - Scientific Data Storage“[8] oder auf einem der Hochleistungsrechner „bwForCluster“[9] oder „bwUniCluster“[10]. Gemäß dem zuvor beschriebenen Konzept der Remote-Visualisierung werden schnelle Netzwerkverbindungen genutzt, um von den bwVisu-Systemen aus auf die Daten zuzugreifen und diese zu visualisieren. Hierbei werden möglichst sichere Zugriffswege verwendet, etwa NFSv4 mit Kerberos Authentifizierung und Verschlüsselung im Betriebsmodus „krb5p“. Durch die exzellente Anbindung an die Netzwerkinfrastruktur in Heidelberg und das Netzwerk der wissenschaftlichen Einrichtungen in Baden-Württemberg („BelWü“) sind hohe Bandbreiten zu den genannten zentralen Speicherdiensten und Speichersystemen der HPC-Cluster verfügbar. Insbesondere können die in Heidelberg lokal verfügbaren Systeme („SDS@hd“, „bwForCluster MLS&WISO“, „heiCLOUD“, etc.) mit zusätzlichen Netzwerkverbindungen mit besonders hohen Bandbreiten und niedrigen Latenzen angebunden werden.

Aus dem dadurch möglichen Direktzugriff auf die Daten resultiert für Forscher die Möglichkeit noch effizienter zur arbeiten: Sie verbinden sich von ihrem Arbeitsplatz zu den Anwendungen

8 http://sds-hd.urz.uni-heidelberg.de/
9 https://www.bwhpc-c5.de/wiki/index.php/BwForCluster_User_Access
10 https://www.bwhpc-c5.de/wiki/index.php/Category:BwUniCluster

auf der bwVisu-Hardware, von wo aus sie über die bandbreitenstarken Verbindungen des Rechenzentrums direkt auf die Daten der Speichersysteme zugreifen können.

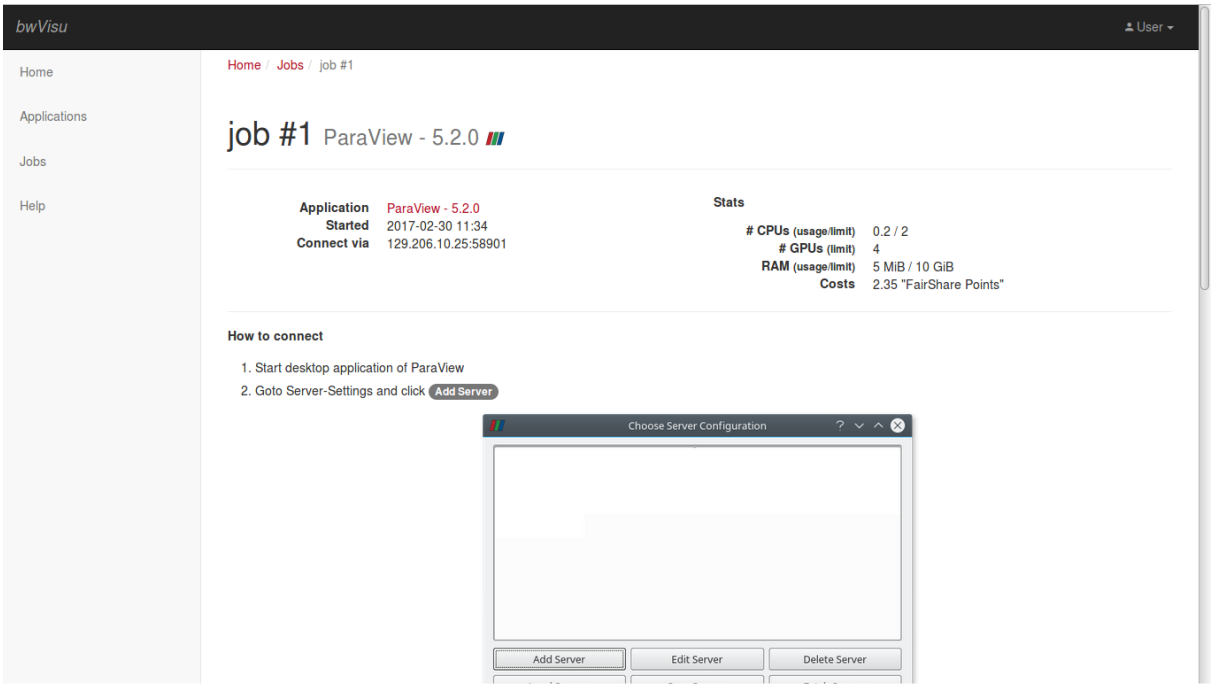


Abbildung 2. Nach dem Start einer Anwendung leitet bwVisu-Web den Nutzer durch den Verbindungsvorgang und liefert nützliche Informationen.

2. Nutzerfreundlichkeit

Ein besonderes Augenmerk wurde auf verschiedene Aspekte der Nutzerfreundlichkeit gelegt. So sollen auch weniger technisch versierte Forscher durch eine einfache, graphische Nutzeroberfläche die Gelegenheit erhalten, das bwVisu-Remote-Visualisierungssystem zu nutzen. Insbesondere ist, im Vergleich zu vielen anderen Systemen, die Verwendung der Linux-Kommandozeile im Regelfall nicht notwendig.

Zu diesem Zweck wurde eine Web-Oberfläche (siehe 1. Abbildung) entwickelt, die es dem Nutzer ermöglicht, die gewünschte Anwendung anhand von Namen oder Beschreibungstext auszuwählen und die erforderliche Version zu starten. Nach einem weiteren Klick und Angabe weniger Informationen, wie erforderlicher Ressourcenmenge (RAM, Anzahl CPUs, Anzahl GPUs), wird die Anwendung vom bwVisu-System für den Nutzer gestartet und steht nach wenigen Sekunden zur Verfügung, siehe im Kapitel zum technischen Aufbau.

In einer Befragung ergab sich, dass Nutzer primär an einer sofortigen Nutzung eines Remote-Visualisierungsdienstes interessiert sind und eher geringen Bedarf an Reservierungen für zukünftige Nutzungen sehen. Hierauf wurde der Dienst bwVisu ausgerichtet: Nutzer verwenden freie Ressourcen und blockieren diese dadurch für andere. Die freien Ressourcen und die aktuelle Auslastung des Systems werden kontinuierlich angezeigt, sodass die Forscher ihre Ressourcen-Anforderungen ggf. hierauf anpassen können.

Sobald die Anwendung gestartet ist, erhält der Nutzer alle nötigen Verbindungsinformationen, sowie Hilfe bei der Einrichtung des evtl. nötigen Client-Programms (2. Abbildung). Für Pro-

gramme, die eine Remote-Display-Software wie „Xpra“[11] nutzen, kann der Zugriff zukünftig sogar ohne die gewohnte graphische Umgebung zu verlassen, direkt über die Web-Oberfläche erfolgen, da der Client direkt in die Web-Anwendung integriert sein wird. So soll erreicht werden, dass nicht nur erfahrene Nutzer, sondern sehr viele Wissenschaftler den Dienst bwVisu ohne besondere Lernphase nutzen können. Direkte Links auf die Web- und Hilfeseiten der Anwendung stehen dem Nutzer während der gesamten Nutzung zur Verfügung.

Bei dem Dienst bwVisu handelt es sich um einen Landesdienst, der u.a. von allen Forscher baden-württembergischer Hochschulen genutzt werden kann. Wie üblich bei Landesdiensten, muss der Nutzer nur über einen Account verfügen, der im Rahmen des föderierten Identitätsmanagements „bwIDM“[12] zur Authentifizierung und Autorisierung eingesetzt wird. Dies ermöglicht einen einfachen, einheitlichen und sicheren Zugang zu diesem Landesdienst.

3. Nutzerspezifische Umgebungen

Ein Hauptmerkmal von bwVisu besteht darin, dass Nutzer angepasste Arbeitsumgebungen verwenden können, die im Hinblick auf die spezifischen Anforderungen des Nutzers hin optimiert werden können. Durch die Nutzung von Container-Technik ist es für bwVisu möglich, seinen Nutzern sowohl übliche, weit verbreitete Anwendungen anzubieten, als auch äußerst fachspezifische Anforderungen zu bedienen (Higgins et al. 2015, Jacobson and Canon 2015, Bahls 2016, Priedhorsky and Randles 2016, Hale et al. 2017). Für typische Anforderungen werden durch die Projektpartner Container entwickelt und für die Nutzer des Dienstes bereit gestellt, die anschließend mit wenigen Klicks genutzt werden können. Hierzu gehören beispielsweise die weit verbreitete Open-Source-Anwendung KitWare „ParaView“[13] für die parallele Analyse mehrdimensionaler und zeitabhängiger Daten, sowie die am Höchstleistungsrechenzentrum Stuttgart (HLRS) entwickelte Software „VISTLE“[14] (VISualization Testing Laboratory for Exascale computing), aber auch die mittels Remote-Display-Software „Xpra“[15] zugänglichen Anwendungen UIUC „VMD“[16] (Moleküldynamikvisualisierung) und ISTI-CNR „Meshlab“[17] (3D-Dreiecksgitterbearbeitung). Diese im bwVisu-Projekt entwickelten Container-Images können durch Nutzer weiter angepasst und spezialisiert werden. Die öffentliche Weitergabe von Container-Images kann optional entfallen, z.B. um restriktiven Software-Lizenzen zu genügen.

Neben der Verwendung offizieller bwVisu-Container, besteht für Nutzer auch die Möglichkeit, eine neue Umgebung am eigenen Arbeitsplatz zu erstellen und daraufhin in bwVisu zu verwenden (3. Abbildung). Nach dem Erzeugen des Container-Images wird dieses in ein Container-Repository (Teil einer sogenannten Container-Registry) geladen. Schließlich wird die Anwendung unter Angabe einiger Metadaten zum Anwendungstyp, Autor, usw. in die bwVisu-Anwendungsdatenbank eingepflegt und steht dann zur Verwendung bereit. Nachdem dieser Vorgang einmal durchgeführt wurde, ist das Starten der Anwendung für Nutzer ohne Entwicklungserfahrung mittels der graphischen Web-Oberfläche einfach möglich.

11 https://xpra.org/
12 https://www.bwidm.de/
13 http://www.paraview.org/
14 http://www.hlrs.de/vistle
15 https://xpra.org/
16 http://www.ks.uiuc.edu/Research/vmd/
17 http://www.meshlab.net/

Der genannte Vorgang des Erstellens eines Container-Images ist vergleichbar mit der Installation oder Kompilierung von Anwendungen auf einer Linux-Kommandozeile und erfordert gewisse Kenntnisse im Bereich Linux. Um die Einstiegshürden zu senken, haben die bwVisu-Projektpartner bereits verwendbare Grundbausteine erstellt, etwa zur Anbindung an Nutzerverzeichnisse oder Nutzung von Remote-Display-Software. Darüber hinaus soll eine Dokumentation mit Empfehlungen und Richtlinien erstellt werden, aus der hervorgeht, wie eine Anwendung mit der bwVisu-Steuerung kommunizieren kann oder wie Graphikbeschleunigung nutzbar wird. So können sich Nutzer voll auf die Installation und Konfiguration ihrer Arbeitsumgebungen und Anwendungen konzentrieren.

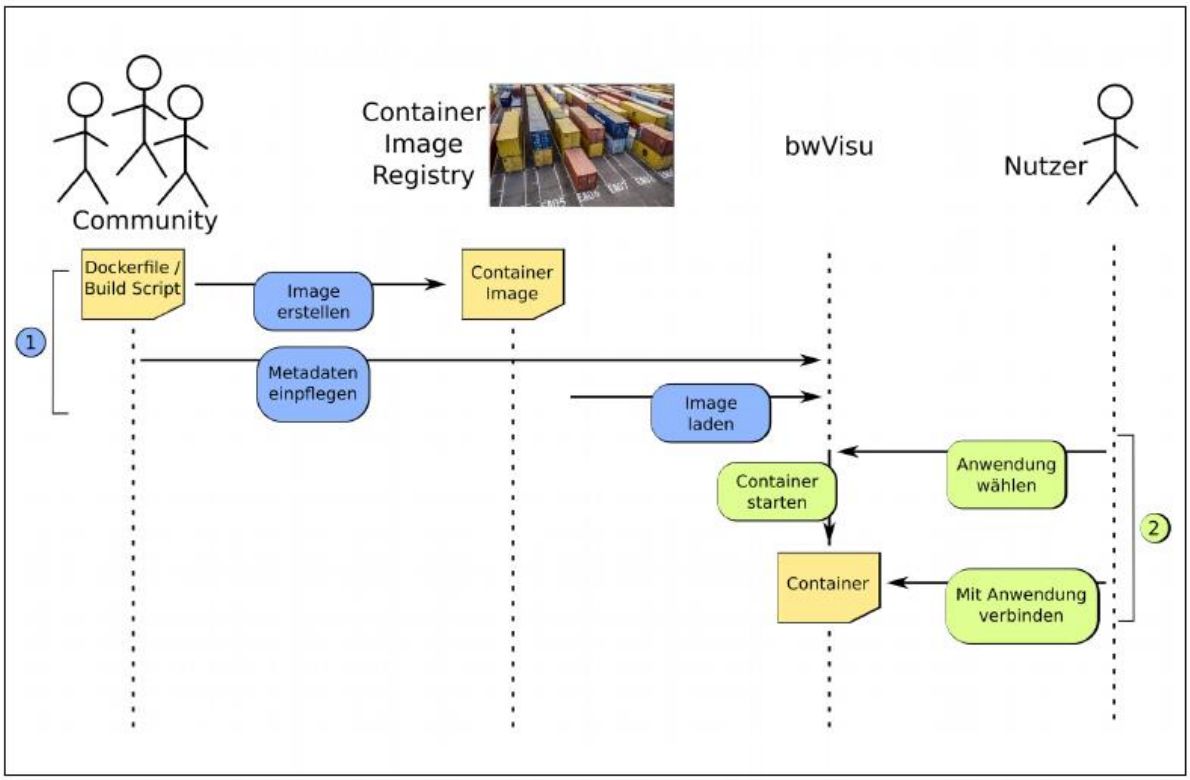


Abbildung 3. Nachdem die Community ein Container Image erstellt hat (1), kann sich der Nutzer nach wenigen Schritten mit seiner Anwendungsinstanz verbinden (2). (Graphik unter Verwendung von Bildmaterial unter CC BY-ND 2.0 Lizenz © 2014 Tristan Taussac: https://www.flickr.com/photos/tristantaussac/15145365916)

Um redundante Arbeit zu vermeiden, ist eine gemeinsame Weiterentwicklung und Pflege dieser Arbeitsumgebungen im Sinne des Free-and-Open-Source-Software-Gedankens in der jeweiligen Arbeitsgruppe oder der weltweiten Forschungsgemeinschaft empfehlenswert: Da ein Container-Image nach Erstellung ohnehin auf einer Container-Registry verfügbar ist, kann es dort auf Wunsch auch mit anderen Forschern geteilt oder auf anderen Rechen-Clustern oder Visualisierungssystemen verwendet werden. Nutzer können sozusagen ihre eigene Software zu bwVisu „mitbringen“ und diese auch von dort auf andere Systeme „mitnehmen“.

4. Skalierbarkeit und Leistung

Für den Dienst bwVisu wurden am Standort Heidelberg leistungsstarke Rechenknoten mit großem Arbeitsspeicher und performanten Graphikkarten beschafft. Eine gute Auslastung der vorhandenen Hardware wird durch den Einsatz von Virtualisierungstechnik und insbesondere durch vertikale Skalierbarkeit nach unten ermöglicht: Sofern Ressourcenanforderungen der Nutzer und Auslastung der Maschinen es zulassen, können mehrere Anwendungen auf einem Rechenknoten ausgeführt werden. Dies wird durch das bwVisu Ressource-Management organisiert, ohne dass der Nutzer selbst Maßnahmen treffen muss. Da Container-Technik nahezu keinen Overhead gegenüber der nicht-virtualisierten Anwendung verursacht, besteht – abgesehen von der Beeinflussung der Anwendungen untereinander – kein Nachteil gegenüber einer Exklusivnutzung von Rechenknoten (Xavier et al. 2013, Felter et al. 2015). Für die Forscher besteht weiterhin die Möglichkeit, ihre Arbeitsumgebung horizontal zu skalieren, also auf mehrere Rechenknoten verteilt auszuführen, sodass eine besonders große Rechenleistung von einer einzelnen Anwendung genutzt werden kann und besonders große Datenmengen verarbeitet werden können, sofern die Anwendung dies unterstützt. Somit kann eine hohe Skalierbarkeit und Performance bei gleichzeitiger optimaler Ausnutzung der Hardware erreicht werden.

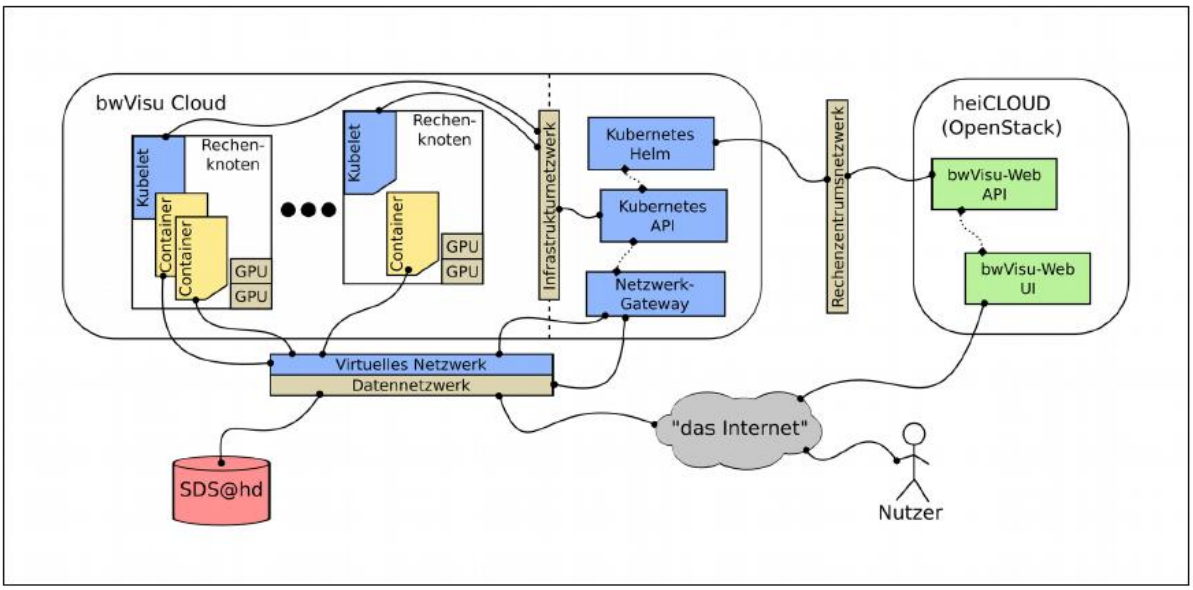


Abbildung 4. Aufbau der bwVisu-Cloud. Die Steuerung der virtuellen Funktionen und Netzwerke (blau) wird durch „CNCF Kubernetes“ übernommen, welches die Container mit den Arbeitsumgebungen der Forscher (gelb) verwaltet. Der Zugriff auf die in SDS@hd gespeicherten Forschungsdaten erfolgt über ein Datennetzwerk, welches den Forschern gleichzeitig den Remote-Zugang zu den in Containern laufenden Anwendungen ermöglicht. Start und Stop derselben steuert der Nutzer über die bwVisu Web-Oberfläche (grün).Physikalisch existierende Einheiten sind braun dargestellt.

Technischer Aufbau der bwVisu-Cloud

Wie zuvor bereits erwähnt, kommt in bwVisu Container-Technologie zum Einsatz. Dadurch werden kurze Startzeiten der Anwendungen (s. Kapitel „Nutzerfreundlichkeit“) ermöglicht, denn der für Anwendungs-Container-Images erforderliche Speicherplatz ist im Vergleich zu Images virtueller Maschinen, die ein vollständiges Betriebssystem enthalten, sehr gering (Xavier et al. 2013,

Felter et al. 2015, Higgins et al. 2015, Jacobson and Canon 2015, Hale et al. 2017), was die Download-Zeit im Hintergrund stark verringert. Gleichzeitig muss hierdurch nur die Anwendung selbst, nicht jedoch ein vollständiges Betriebssystem gestartet werden, was die Wartezeit bis zur Nutzung einer Anwendung weiter reduziert. Für häufig genutzte Software kann die zum Start benötigte Zeit auf oft weniger als eine Sekunde gesenkt werden.

Der schematische Aufbau der bwVisu-Cloud wird in der 4. Abbildung dargestellt. Im geplanten Ausbau besteht sie aus 10 Rechenknoten mit je 128 GiB RAM und 2 CPUs à 14 Kernen. Die Hälfte dieser Knoten verfügt über je 2 Graphikkarten vom Typ AMD „FirePro S7150 X2“, wogegen die andere Hälfte mit je 2 Graphikkarten vom Typ Nvidia „Tesla M10“ ausgestattet ist. Diese Maschinen sind jeweils an einen 10 GbE Switch angebunden, der wiederum an BelWü und das Datennetz des des Speicherdienstes „SDS@hd“-Dienst angebunden ist. Entsprechend sind hohe Zugriffsgeschwindigkeiten der Anwendungen auf die Daten gewährleistet. Zusätzlich verfügt jeder Visualisierung-Server über eine Anbindung an ein bwVisu-internes InfiniBand-FDR-Netzwerk, über das zukünftig anwendungsinterne Kommunikation latenzarm mit hoher Bandbreite (56 Gbps) transportiert werden soll. Alle beteiligten Maschinen werden einheitlich mit CoreOS „Container Linux“ betrieben, einem minimalen, auf die Ausführung von Anwendungs-Containern ausgelegtem Betriebssystem.

Gesteuert werden diese Rechenknoten von mehreren redundant ausgelegten Infrastrukturknoten, auf denen u.a. die Container-Verwaltung CNCF „Kubernetes“[18], die Netzwerksteuerung Tigera „Calico“[19], das Monitoring-System CNCF „Prometheus“[20] ausgeführt werden. Diese Steuerkomponenten können zur Erhöhung der Verfügbarkeit auf der Heidelberger „OpenStack“-basierten Private-Cloud-Lösung „heiCLOUD“[21] betrieben werden, da die Kommunikation der Steuerungs- mit den Rechenknoten über HTTPS verschlüsselt und authentifiziert wird. Die genannten Infrastrukturdienste werden genauso wie die Remote-Visualisierunganwendungen in Containern betrieben.

Die Steuerung des „Kubernetes“-Systems wird wiederum durch den bwVisu-API-Dienst vorgenommen, auf den die Weboberfläche bwVisu-Web zugreift. Durch diese zweistufige Architektur (Trennung von Frontend und Backend) können leicht Vereinfachungen, erweiterte Kontrollmechanismen oder zusätzliche bwVisu-spezifische Funktionen implementiert werden, da somit eine zentrale Stelle existiert, die von sämtlichen Nutzeranfragen durchlaufen wird. Dieser API-Dienst ist es auch, welcher in Verbindung mit Microsoft „Helm“[22] die einfache Anforderung des Nutzers nach einer bestimmten Anwendungsumgebung in Kubernetes-Begriffe wie virtuelle Netzwerke, Datenanbindungen an „SDS@hd“ und Container übersetzt und gleichzeitig die Authentifizierung des Datenzugriffs durch Bereitstellung von Kerberos-Tickets ermöglicht.

Der Remote-Zugriff des Nutzers auf seine laufenden, virtuellen Arbeitsumgebungen erfolgt über Netzwerk-Gateways, die die internen, virtuellen Netzwerke mit den externen (z.B. „das Internet“) verbinden, was z.B. durch für SDNs (Software Defined Network) ausgelegte Hardware übernommen werden kann.

18 https://kubernetes.io/
19 https://www.projectcalico.org/
20 https://prometheus.io/
21 http://heicloud.uni-heidelberg.de/
22 https://helm.sh/

Zusammenfassung und Ausblick

Da in den unterschiedlichsten wissenschaftlichen Fachdisziplinen sehr umfangreiche Daten anfallen, die visualisiert und verarbeitet werden müssen, sind bestimmte Anforderungen an die entsprechenden Werkzeuge und Arbeitsumgebungen zu stellen: Neben rein technischen Anforderungen (Leistung und Skalierbarkeit) sind bedarfsgerecht auch die Nutzerfreundlichkeit und der Wunsch nach Community-spezifischen Arbeitsumgebungen zu betrachten. In dieser Arbeit werden diese Anforderungen vorgestellt und Projekte verschiedener Gruppen genannt, die manche dieser Anforderungen adressieren und Konzepte und technische Lösungen für Teilaspekte bieten.

Im Hauptteil der Arbeit wird der dem Projekt bwVisu zugrunde liegende Ansatz für einen Remote-Visualisierungsdienst vorgestellt. Durch den Einsatz von Containern und einem gut abgestimmten Betriebs- und Nutzungsmodell, kann die vorhandenen Rechenressourcen über eine Web-Oberfläche für die Datenanalyse und -visualisierung angeboten werden. Der Zugriff auf lokale Datenspeicher, sowie die Anbindung von entfernten Speichersystemen über existierende breitbandige Netzwerke wird im Projekt ebenfalls untersucht. Der gleichnamige Dienst bwVisu ist noch in der Entwicklung und verspricht eine bedarfsorientierte Lösung zu werden.

Im weiteren Verlauf des Projekts sollen weitere Nutzungsszenarien und Anwendungen in bwVisu integriert und die Nutzerdokumentation erweitert werden. Diese Maßnahmen sollen dazu dienen, dass neue Nutzer den Dienst besonders einfach für ihre jeweiligen Forschungsvorhaben einsetzen können. Sofern der Bedarf dies erfordert, können Warteschlangen und Reservierungen für zukünftige Ressourcen-Nutzung eingeführt werden.

Die Erstellung neuer Container-Images, sowie deren Einbindung in bwVisu, soll durch verschiedene Maßnahmen in der Zukunft noch weiter vereinfacht werden. Derzeit erfordert das lokale Testen eines Container-Images auf dem Arbeitsplatzrechner viel Vorwissen. Eine stärkere Unterstützung der Container-Erstellung durch Automatismen würde den Entwicklungsprozess vereinfachen und beschleunigen. Hierzu wären z.B. Scripte nötig, die automatisiert auf dem lokalen Arbeitsplatzcomputer eine Umgebung ähnlich der bwVisu-Cloud schaffen können, sodass der Anwender sein Image mit wenigen Befehlen lokal testen kann.

Zur Zeit werden neue Container-Images noch manuell durch Administratoren in die Anwendungsdatenbank von bwVisu eingepflegt. Dieser Vorgang könnte zukünftig graphisch unterstützt werden, sodass Nutzer diese Eintragungen selbstständig vornehmen können, und auch die Nutzungsrechte in der Arbeitsgruppe und der Forschungsgemeinde selbst verwalten können. Zusätzlich wären Automatismen hilfreich, welche neu in die Container-Repositories geladene Images automatisiert zur Verfügung stellen, oder bei der Erkennung der Anwendungsart (z.B. "MPI" oder "Xpra") unterstützen.

Zusätzlich dürfte es für Nutzer hilfreich sein, direkt und live Zugriff auf die Monitoring-Daten des bwVisu-Systems zu erhalten. Dies kann z.B. helfen, die Ressourcen-Anforderungen an die tatsächliche Nutzung anzupassen oder das Verhalten ihrer Anwendung unter Gesichtspunkten der Leistung (z.B. CPU- oder I/O-Auslastung auf Prozessebene) besser zu verstehen. Langfristig wäre auch die Möglichkeit zusätzlicher Dienstleistungen, wie integrierte Debugger oder Performance-Monitoring-Werkzeuge auf noch feingranularerer Ebene, interessant.

Datenzugriffe sind stets limitiert durch die Netzwerkverbindung zwischen der bwVisu-Cloud und dem System, auf dem sich die zu analysierenden Daten befinden. Bisher wurde der lokale Speicherdienst „SDS@hd“ an bwVisu angebunden. Da in Baden-Württemberg durch „BelWü“ eine Netzwerkverbindung von hoher Bandbreite zwischen den wissenschaftlichen Einrichtungen

bereitgestellt wird, könnten standortübergreifende Datenzugriffe im Rahmen der Analyse in Betracht gezogen werden. Beispielsweise ist im Rahmen des Projekts geplant die bwVisu-Cloud mit dem Hochleistungsrechner „bwForCluster NEMO“ in Freiburg zu verbinden, um auf die dort befindlichen Daten zugreifen zu können. Durch diese Schritte könnte es bwVisu mehr Nutzern erlauben, ortsunabhängig, insbesondere unabhängig vom Ort der Datenspeicherung, zu arbeiten.

Die Entwicklung und Erprobung des Dienstes bwVisu geschieht direkt auf dedizierter Hardware, ohne zusätzliche Virtualisierungsschichten („bare-metal“), da dies die beste Leistung bietet (Xavier et al. 2013, Felter et al. 2015, Higgins et al. 2015). Eine Integration in traditionelle HPC-Systeme wäre aus Leistungssicht ebenfalls gut motiviert, jedoch müsste aus Sicht der Infrastruktur eine Cloud-artige Flexibilität vorhanden sein. Entsprechende Möglichkeiten sollten zukünftig untersucht werden. Weiterhin könnte auch ein Betriebsmodell auf einer oder mehreren Cloud-Plattformen untersucht werden, welches voraussichtlich Automatisierungs- und Administrationsvorteile mit sich bringen könnte. Es sollte zukünftig geprüft werden, welche Nachteile bzgl. der Ausnutzung der Leistungsfähigkeit der Hardware sich daraus ergeben würden, und wie sich diese z.B. mit den Erkenntnissen aus o.g. „MIKELANGELO“ Projekt für bestimmte Nutzungsszenarien verringern lassen. Die Integration von bwVisu in vorhandene HPC- oder Cloud-Systeme birgt potentielle Synergie-Effekte bzgl. der Hardware-Auslastung. Die erforderlichen Anpassungen und Weiterentwicklungen könnten in einem Folgeprojekt adressiert werden.

Danksagung

Das Projektvorhaben bwVisu wurde durch das Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg (MWK) gefördert.

Literaturangaben

Bahls, Donald. "Evaluating Shifter for HPC Applications" In *Cray User Group Conference Proceedings*. 2016.

Felter, Wes, Alexandre Ferreira, Ram Rajamony, and Juan Rubio. "An updated performance comparison of virtual machines and linux containers" In *Performance Analysis of Systems and Software (ISPASS), 2015 IEEE International Symposium On*, pp. 171-172. IEEE, 2015.

Hale, Jack, Lizao Li, Chris Richardson, and Garth Wells. "Containers for portable, productive and performant scientific computing" *Computing in Science & Engineering* (2017).

Higgins, Joshua, Violeta Holmes, and Colin Venters. "Orchestrating docker containers in the HPC environment" In International Conference on High Performance Computing, pp. 506-513. Springer, Cham, 2015.

Jacobsen, Douglas M., and Richard Shane Canon. "Contain this, unleashing docker for HPC" *Proceedings of the Cray User Group* (2015).

Jomier, Julien, Sebastien Jourdain, Utkarsh Ayachit, and Charles Marion. "Remote visualization of large datasets with MIDAS and ParaViewWeb." In *Proceedings of the 16th International Conference on 3D Web Technology*, pp. 147-150. ACM, 2011.

Kurtzer, Gregory M., Vanessa Sochat, and Michael W. Bauer. "Singularity: Scientific containers for mobility of compute." *PloS one* 12, no. 5 (2017): e0177459.

Mauch, Viktor, M. Bonn, S. Chilingaryan, A. Kopmann, W. Mexner, and D. Ressmann. "OpenGL-based data analysis in virtualized self-service environments." *Proc. PCaPAC2014, http://jacow. org* (2014).

Mauch, Viktor, Marcel Kunze, and Marius Hillenbrand. "High performance cloud computing." *Future Generation Computer Systems* 29, no. 6 (2013): 1408-1416.

Mouton, Christophe, Kristian Sons, and Ian Grimstead. "Collaborative visualization: current systems and future trends." In *Proceedings of the 16th International Conference on 3D Web Technology*, pp. 101-110. ACM, 2011.

Priedhorsky, Reid, and Timothy C. Randles. "Charliecloud: Unprivileged containers for user-defined software stacks." (2016).

Shi, Shu, and Cheng-Hsin Hsu. "A survey of interactive remote rendering systems." *ACM Computing Surveys (CSUR)* 47, no. 4 (2015): 57.

Stegmaier, Simon, Joachim Diepstraten, Manfred Weiler, and Thomas Ertl. "Widening the remote visualization bottleneck." In *Image and Signal Processing and Analysis, 2003. ISPA 2003. Proceedings of the 3rd International Symposium on*, vol. 1, pp. 174-179. IEEE, 2003.

Upson, Craig, T. A. Faulhaber, David Kamins, David Laidlaw, David Schlegel, Jefrey Vroom, Robert Gurwitz, and Andries Van Dam. "The application visualization system: A computational environment for scientific visualization." *IEEE Computer Graphics and Applications* 9, no. 4 (1989): 30-42.

Verma, Abhishek, Luis Pedrosa, Madhukar Korupolu, David Oppenheimer, Eric Tune, and John Wilkes. "Large-scale cluster management at Google with Borg." In *Proceedings of the Tenth European Conference on Computer Systems*, p. 18. ACM, 2015.

Xavier, Miguel G., Marcelo V. Neves, Fabio D. Rossi, Tiago C. Ferreto, Timoteo Lange, and Cesar AF De Rose. "Performance evaluation of container-based virtualization for high performance computing environments" In *Parallel, Distributed and Network-Based Processing (PDP), 2013 21st Euromicro International Conference on*, pp. 233-240. IEEE, 2013.

Open Data? Zum Umgang mit Forschungsdaten in den ethnologischen Fächern

Sabine Imeri[1]

1 Fachinformationsdienst Sozial- und Kulturanthropologie, Universitätsbibliothek, Humboldt-Universität zu Berlin

Zusammenfassung. In den ethnologischen Fächern sind Fragen von Forschungsdatenmanagement bisher kaum diskutiert. Der Beitrag berichtet vor dem Hintergrund der Spezifik ethnografischer Forschung und auf der Grundlage einer Online-Umfrage wie EthnologInnen Chancen und Probleme von Langzeitarchivierung und Data-Sharing einschätzen und welche Bedingungen für die Überführung ethnografischer Materialien in digitale Langzeitarchive sich daraus ergeben. Wie offen können Forschungsdaten überhaupt gemacht werden?

Einleitung

Gefragt, ob er zu einer Publikation eine ethnologische Deutung von drei Interviews beitragen könne, die ein im Schnittfeld zur Kulturwissenschaft forschender Psychologe mit Überlebenden und Zeitzeugen eines Lawinenunglücks im österreichischen Galtür geführt hatte, hatte Michael Simon, Professor für Europäische Ethnologie in Mainz Bedenken: Als „äußerst kitzlig“ empfand er die Aufgabe, denn es „handelt sich um Fremddaten, über deren Zustandekommen mir an und für sich wenig bekannt ist. […] Bei seinen Erhebungen in Galtür bin ich nicht dabei gewesen, konnte nicht die Personen kennenlernen, mit denen er gesprochen hat, habe nicht ihre Nähe gespürt und einen Eindruck von den Örtlichkeiten gewinnen können [...].“ Während die ebenfalls um eine Interpretation gebetenen Psychologen und Psychotherapeutinnen sich zu diesem Umstand überhaupt nicht äußerten, schreibt Simon: „Diesem Ansinnen nachzukommen, kostet Überwindung. […] Dem Ethnologen bereiten solche Formen des Austauschs Schwierigkeiten […].“ (Simon 2015: 93f)

Die Episode verweist mindestens auf zweierlei: Zum einen ist die Sekundärnutzung von Daten anderer ForscherInnen in den ethnologischen Fächern offenbar nicht üblich, Strategien und Formen der Nachnutzung wenig etabliert. Zum zweiten macht sie auf methodologische und erkenntnistheoretische Dimensionen der Nachnutzung von Daten aufmerksam, die eine disziplinäre Spezifik aufweisen.

Die DFG hat im Rahmen des Fachinformationsdienstes (FID) „Sozial- und Kulturanthropologie“ an der Universitätsbibliothek der Humboldt-Universität zu Berlin auch ein Teilprojekt mit explorativem Charakter zum Forschungsdatenmanagement in den ethnologischen Fächern bewil-

ligt (2016-2018).[1] Die Beantragung erfolgte dabei wesentlich auf Anregung und mit Unterstützung aus den Fachcommunities, die Arbeit des FID wird – wie schon die des vorausgegangenen Sondersammelgebiets Volks- und Völkerkunde – von einem wissenschaftlichen Beirat begleitet.[2] In einem ersten Schritt werden fortlaufend der aktuelle Stand des Umgangs mit Forschungsdaten sowie Einstellungen zu und Erfahrungen mit dem Thema Forschungsdatenmanagement erhoben und Möglichkeiten der Langzeitarchivierung sondiert, um in einem zweiten Schritt Nutzungsszenarien und einen Muster-Workflow der forschungsbegleitenden Datensicherung zu entwickeln und zu erproben. Dieser Beitrag beruht wesentlich auf den Ergebnissen einer Online-Umfrage, die der FID zwischen Dezember 2016 und Februar 2017 durchgeführt hat. Erfragt haben wir unter anderem Speicherroutinen, Datenformate und Datenmengen.[3] Besonders wichtig war aber auch, mit der Umfrage in eine Diskussion über Chancen und Probleme von Langzeitarchivierung und Data-Sharing in den ethnologischen Fächern einzutreten und Anforderungen an Forschungsdatenmanagement und Infrastrukturen zu eruieren.

Ethnografische Forschungspraxis

EthnologInnen betreiben nicht nur, aber in vielen Forschungsszenarien *teilnehmende Beobachtung*, eine Methode der Datengewinnung, die auf „Begegnung, Interaktion und der sozialen Teilnahme am Alltagsleben unterschiedlicher Menschen“ beruht. (Knecht 2013: 83) Umfassender bezeichnet *Feldforschung* den gesamten Forschungsprozess, in dem teilnehmende Beobachtung mit Interviewmethoden, Kartierungen und anderen Formen der Dokumentation, aber auch Archivstudien kombiniert wird. Mit der Feldforschung favorisieren EthnologInnen einen dezidierten Methodenpluralismus und generieren dabei vorwiegend qualitatives, oft sehr diverses, sich meist gegenseitig kommentierendes Datenmaterial (Abb. 1), das geeignet ist, die Komplexität sozialer, sozio-technischer und sozio-natürlicher Ordnungen angemessen zu beschreiben (Knecht 2013), ganz gleich, ob sie sich für Migration und Mobilität interessieren, für Globalisierungsprozesse, für Formen und Netzwerke von Religiosität, Geschlechterverhältnisse und Sexualität oder für Phänomene der Medialisierung und Technisierung von Alltag.[4] Auch deshalb kann es zum Problem werden, wenn man – wie im eingangs geschilderten Fall – für die Interpretation sozialer Wirklichkeiten „nur“ über Interviews verfügt.

1 Unter „ethnologische Fächer“ sind hier die Fachtraditionen der Völkerkunde/Ethnologie und der Volkskunde/Europäischen Ethnologie im deutschsprachigen Raum subsumiert, deren Institutionen und Einrichtungen auch unter anderen Bezeichnungen wie Empirische Kulturwissenschaft, Populäre Kulturen oder Sozial- und Kulturanthropologie, teils in Kombination zu finden sind.

2 Der Beirat setzt sich aus VertreterInnen relevanter Institutionengruppen und Fachgesellschaften zusammen. https://www.ub.hu-berlin.de/de/literatur-suchen/fachinformationsdienste/ssg-volks-und-voelkerkunde [14.03.2017]

3 Erwartungsgemäß weichen die Antworten hier kaum von ähnlichen Umfragen ab, z. B. werden Daten häufig auf externen Festplatten oder USB-Sticks gesichert.

4 Zu den Themenfeldern, die derzeit in den ethnologischen Fächern schwerpunktmäßig bearbeitet werden vgl. die Kommissionen und Arbeitsgruppen der Fachgesellschaften Deutsche Gesellschaft für Volkskunde http://www.d-g-v.org/kommissionen und Deutsche Gesellschaft für Völkerkunde https://www.dgv-net.de/arbeitsgruppen/

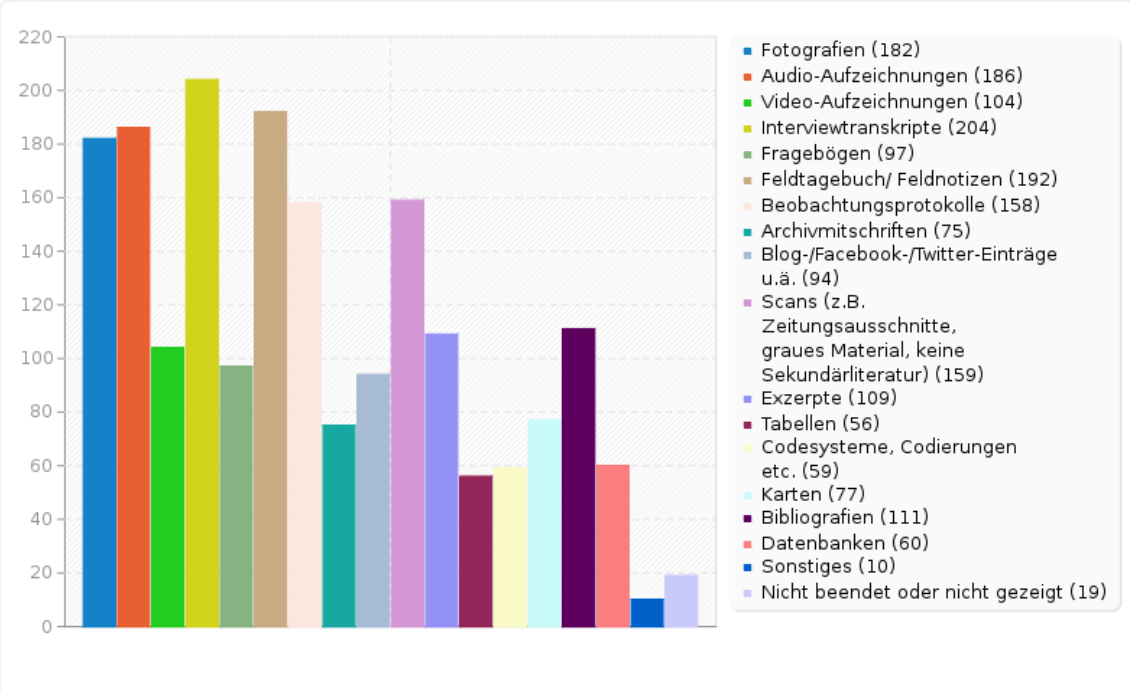


Abbildung 1. Was sind in Ihrer Forschung „Forschungsdaten"? (Mehrfachantworten), Umfrage des FID Sozial- und Kulturanthropologie zum Forschungsdatenmanagement in den ethnologischen Fächern (2017)

Feldforschungen sind konzipiert als nicht replizierbare, weil situations- und beobachterabhängige, offene und flexible, oft unvorhersehbar verlaufende Prozesse. Feldforschung bedeutet in der Regel, dass Forschende die Lebensräume der ProtagonistInnen ihrer Forschung persönlich aufsuchen, konstitutiv ist dabei „ein gründliches, aufwändiges Sich-Einlassen auf *real-world-situations*". (Knecht 2013: 91) Für einen großen Teil ethnografischer Daten gilt entsprechend, dass sie in einem Prozess entstehen, in dem Forschende und Akteure im Forschungsfeld mehr oder weniger langfristige Beziehungen aufbauen.[5] In jedem Fall sind Forschende auf die Kooperation und Zustimmung der Akteure im Feld angewiesen, anders als das etwa bei großflächigen Datenerhebungen mit Behördenunterstützung wie beim *Nationalen Bildungspanel* oder beim *Zensus* der Fall ist. Feldforschungen sind in vielen Fällen Einzelforschungen, die wesentlich davon leben, dass ein Feldzugang überhaupt erarbeitet werden kann. Entsprechend sind solche Forschungsbeziehungen „fragile Gebilde" (Breidenstein et al ^{2}2015: 62) und Vertrauensverhältnisse, in denen mehrfach sensible Daten produziert werden: EthnografInnen erhalten intensive Einblicke in die Alltage von Personen und Gruppen, erfahren Details zum Beispiel über religiöse Praktiken, sexuelle Orientierungen, ethnische Zugehörigkeiten oder auch illegale Praktiken. Jedenfalls kann ihr Wissen, können ihre Daten zu Risiken für die beforschten ProtagonistInnen werden, etwa in eskalierenden Konflikten. Sensibel sind ethnografische Daten aber auch mit Blick auf die Forschenden selbst, weil zum Beispiel die Emotionen der Forschenden – im Sinne einer zentralen Erfahrungsdimension – eine wichtige Rolle im Erkenntnisprozess spielen können. (Stodulka 2014) Die Teilnahme der Forschenden am Geschehen berührt daher deren ganze Person, die „keine unabhängige Größe darstellt, sondern bei der Abrechnung der Forschungsleistungen mit all

5 Das gilt für kontinuierliche, stationäre Langzeitforschung ebenso wie für temporalisierte, zeitlich diskontinuierliche Feldaufenthalte. Letztere gehen in der Regel einher mit der durchaus verbreiteten Nachnutzung eigener Forschungsdaten, teils zu einem deutlich späteren Zeitpunkt. (Welz 2013)

ihren menschlichen Stärken und Schwächen voll zu Buche schlägt“. (Simon 2015: 93) Und das bedeutet eben auch, dass Forschende selbst in Forschungsdaten als Personen erkennbar werden.

Ergebnis einer seit mehreren Jahrzehnten anhaltenden und im Grunde unabgeschlossenen „disziplinären Selbstaufklärung“ (Knecht 2009) über die Bedingungen, die historischen und politischen Kontexte und Rahmungen ethnografischer Wissensproduktion ist überdies eine besonders intensive Auseinandersetzung mit forschungsethischen Fragen.[6] Die permanente systematische Reflexion des Verhältnisses von Forschenden und ihren Gegenübern und der machtvollen Dynamiken, die Forschungsbeziehungen und -situationen mit strukturieren, ist Teil jeder ethnografischen Feldforschung, aus ihr resultiert auch eine besondere Verantwortung der Forschenden für ihr Feld, für ihre Daten und die Verwendung dieser Daten auch nach dem Abschluss einer Forschung.

Umfrageergebnisse

Die Umfrage selbst wurde mit der freien Applikation *LimeSurvey* erstellt, die Konzeption basierte wesentlich auf früheren Umfragen und Materialien verschiedener Infrastruktureinrichtungen[7] sowie auf ersten Gesprächen mit Expertinnen und Experten. Insgesamt sind 270 Fragebogen von Forschenden aller Statusgruppen aus Universitäten, außeruniversitären Forschungseinrichtungen und Museen ausgefüllt worden und in die Auswertung eingeflossen.[8] Ein Drittel der TeilnehmerInnen hat die Umfrage vorzeitig abgebrochen, so dass 181 bis zur letzten Frage ausgefüllte Bögen vorliegen.

Etwas überraschend ist, dass das Thema Forschungsdatenmanagement offensichtlich bekannter ist, als erwartet. (Abb. 2) Konkrete Anwendungskenntnisse und vertieftes Wissen sind gleichwohl wenig verbreitet: Lediglich zehn Prozent der TeilnehmerInnen geben zum Beispiel an, für ihr aktuelles Forschungsprojekt einen Datenmanagementplan erstellt zu haben, weitere rund zehn Prozent haben die Absicht, dies zu tun. 48 % haben keinen und rund 20 % Prozent haben keine Vorstellung, was ein Datenmanagementplan ist.

6 Manche Fachgesellschaften haben eigene Erklärungen und Richtlinien zur Forschungsethik erarbeitet, vgl. z.B. American Anthropological Association http://www.americananthro.org/ParticipateAndAdvocate/Content.aspx?ItemNumber=1895
Deutsche Gesellschaft für Völkerkunde https://www.dgv-net.de/dgv/ethik/

7 Heinrich and Schäfer 2013, Simukovic, Kindling and Schirmbacher 2013, Opitz and Mauer 2005.

8 Zum Vergleich: Eine Umfrage des Projektes IANUS erreichte 2013 240 AltertumswissenschaftlerInnen. (Heinrich, Jahn and Schäfer 2014)

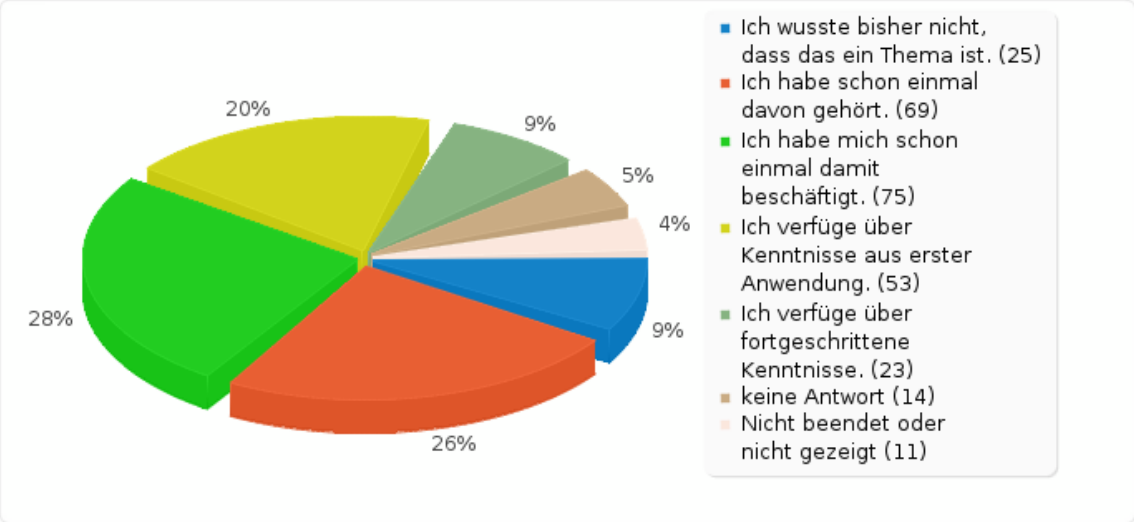


Abbildung 2. Wie gut kennen Sie sich im Thema Forschungsdatenmanagement aus? Umfrage des FID Sozial- und Kulturanthropologie zum Forschungsdatenmanagement in den ethnologischen Fächern (2017)

Insgesamt, das zeigen auch die Gespräche, die wir geführt haben, ist das Spektrum der Einschätzungen zu Fragen von Forschungsdatenmanagement enorm breit: Stark ablehnende Haltungen begegnen hier ebenso wie Nachfragen nach konkreten infrastrukturellen Angeboten zur Datenarchivierung. Das wird auch in der allgemeinen Einschätzung zum wissenschaftlichen Wert nachnutzbarer Daten erkennbar, wobei sich die ProfessorInnen hier etwas skeptischer geäußert haben, als das gesamte Sample. (Abb. 3)[9] Gleichwohl wird hier auch eine grundsätzliche Zustimmung zu Formen der Recherchierbarkeit und Nachnutzbarkeit von Forschungsdaten sichtbar.

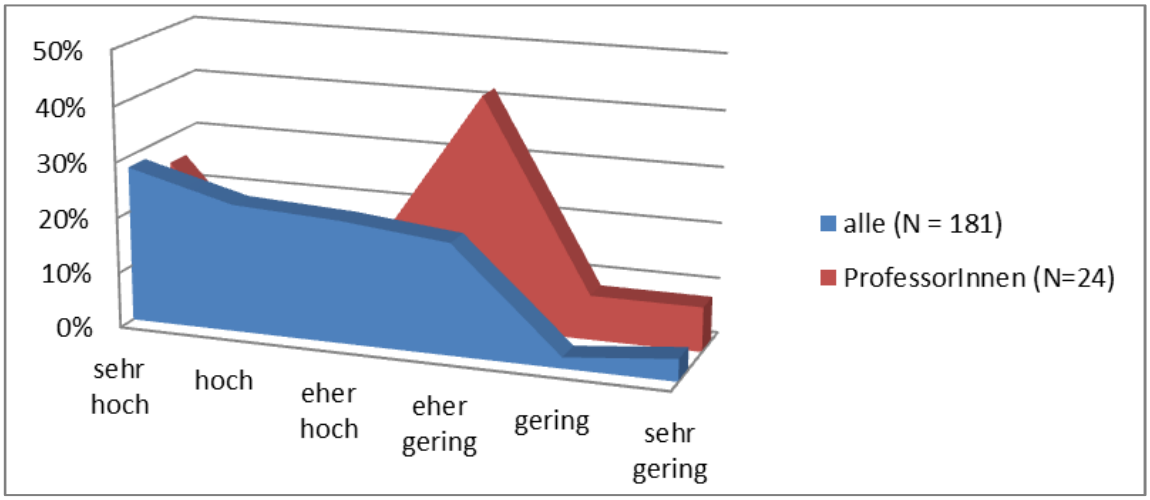


Abbildung 3. Wie hoch schätzen Sie allgemein den wissenschaftlichen Mehrwert recherchierbarer und nachnutzbarer ethnografischer Forschungsdaten ein? Umfrage des FID Sozial- und Kulturanthropologie zum Forschungsdatenmanagement in den ethnologischen Fächern (2017)

Weil in vorbereitenden Gesprächen deutlich wurde, dass es Bedenken und auch grundsätzliche Zweifel an der Langzeitarchivierung und besonders an der Sekundärnutzung ethnografischer Daten geben würde, hatten wir auch dezidiert nach skeptischen Einschätzungen und Kritik gefragt:

9 Wie Abb. 3 zeigt, war die Gruppe der ProfessorInnen mit 24 Personen allerdings recht klein.

Neben Einwänden die auch aus anderen Umfragen bekannt sind, etwa hinsichtlich des Arbeitsaufwands oder der Finanzierung der Datenaufbereitung (vgl. z. B. Droß 2015), lässt sich die Skepsis nach einer ersten Sichtung im Wesentlichen auf vier Aspekte oder Problemkreise verdichten, die sich vor allem auf die methodischen Besonderheiten ethnografischer Feldforschung beziehen. Die hier verwendeten Zitate geben exemplarisch Einschätzungen wieder, die sich ähnlich in zahlreichen Kommentaren finden.

Kontexualisierung

> „Die Besonderheit der Feldforschung ist eben genau die Intersubjektivität der Forschungssituation, die sich anhand von Tonaufnahmen oder Aufzeichnungen weder ‚überprüfen' noch rekonstruieren lässt."
> „Ein Interviewtranskript sagt oft wenig (oder führt gar in die Irre), wenn Hintergrundinformationen [...] fehlen. Solche Kontextinformationen, die vielleicht aufgeschrieben wurden, vielleicht auch aus forschungsökonomischen Gründen nirgendwo in den Daten explizit auftauchen, können bei der Bewertung durch die Forschenden aber entscheidend sind."
> Wenn ich Daten anderer nachnutzen würde, bräuchte ich eine „ausführliche Kontextualisierung des Forschungsprozesses, genaue Informationen über die verwendeten Methoden, die Samplingstrategien, die Datenaufbereitung etc."

„Datasets don't speak for themselves." (Lederman 2016: 261) Ob und wie Feldforschungsbeziehungen und die komplexen Kontexte der Datenerhebung angemessen so dokumentiert werden können, dass sie für Dritte – vielleicht sogar fachfremde Forschende – verständlich und damit überhaupt sinnvoll nutzbar werden, ist eine offene Frage, die auch im Fall der Interviews zum Lawinenunglück bereits anklang. Angemessenheit hat hier nicht nur eine methodische Dimension, sondern auch eine forschungsökonomische, weil, wie oben geschildert, Datenvielfalt und Datenmengen in ethnografischen Projekten erheblich sein können. Diskutiert werden müssen demnach fachliche Standards für angemessene (forschungsbegleitende) Kontextualisierung ethnografischer Daten, die methodische Zugänge und Feldsituationen möglichst transparent und nachvollziehbar machen. Vielfach wird die Notwendigkeit, Kontakt zu den Primärforschenden aufnehmen zu können, in Betracht gezogen. Metadatenstandards müssen flexibel angepasst werden können, weil sich Forschungsfelder dynamisch entwickeln.

Methodenentwicklung

> „Das Bewusstsein, dass die Daten später nachgenutzt werden können, könnte die Art des ethnografischen Aufschreibens und Beschreibens mitunter deutlich verändern. Auch bei Interviews gilt: das Bewusstsein um eine mögliche Nachnutzung könnte auch dazu führen, dass die/der Interviewende die eigene Art zu fragen und sich selbst in das Gespräch einzubringen an der Vorzeigbarkeit des eigenen Verhaltens ausrichtet."
> „Die Entkopplung von ‚Daten' und den dazugehörigen ‚Emotionen' in der Feldforschung, die nur der/dem tatsächlich Forschenden gegenwärtig sind, birgt Probleme in der zukünftigen Auswertung."
> „Die Speicherung [...] bedeutet für den gesamten Forschungsprozess eine erhebliche Veränderung gegenüber früherer Praxis. Das betrifft sowohl die Frage, wie die Erforschten ausreichend informiert werden, als auch die Notwendigkeit, die Möglichkeit der Nachnutzung frühzeitig in der Auswertung und Dokumentation zu berücksichtigen [...]."

Die Frage nach der Kontextualisierung ethnografischer Daten ist eng verbunden mit Fragen der Methodenentwicklung. Bedenken, dass die für die Feldforschung charakteristische Offenheit des Forschungsprozesses zu Gunsten von Überprüfbarkeit eingeschränkt und die Involviertheit der „ganzen Person" der Forschenden als zentrales Moment der Erkenntnisprozesses in Frage gestellt werden könnten, zeigen, das sowohl Rückwirkungen auf den gesamten Forschungsprozess als auch neue Strategien der forschungsbegleitenden Dokumentation intensiv reflektiert werden müssen. Die Überlegung, Langzeitarchivierung könnte künftig vor allem „Meta-Aggregation statt Interpretation" bedeuten, weist aber auch darauf hin, dass der Status, den Material aus Datenarchiven gegenüber selbst erhobenen Felddaten haben kann, ungeklärt ist. (Vgl. Lederman 2016) Im Fall der Interviews zum Lawinenunglück in Galtür entschied sich der Autor zum Beispiel, die Transkripte „quasi als historische Dokumente aus dem Jahre 2008 [zu] betrachten". (Simon 2015: 96)

Auswahl

> „Feldtagebücher können sehr intime Quellen bzw. ein geschützter Raum sein, die eben nicht zur Veröffentlichung gedacht sind, sondern Zweifel, Emotionen, Irrwege etc. beinhalten sollen und müssen."
> „Teils gehören zu ethnographischer Forschung ja auch persönliche Beobachtungen und Anmerkungen etc. dazu. Es stellt sich somit je nach Daten auch die Frage nach einem gewissen Datenschutz der forschenden Person."
> „Umgekehrt ließe sich fragen: Wie vermeidet man als Feldforscherin die ‚Schere im Kopf', da man diese oder jene schnelle Formulierung (z. B. eine negativ wertende) als Erinnerungsstütze hilfreich findet, sie aber nicht teilen möchte?"

Auf die geschilderte Erkennbarkeit der Forschenden im Material wird in den Kommentaren häufig Bezug genommen. Es geht dabei jedoch keineswegs nur um Selbstschutz – der durchaus ein legitimes Anliegen ist – oder Vorbehalte gegenüber Überprüfung oder Kontrolle von Forschungser-

gebnissen, sondern im Grunde um die Frage der Auswahl und Aufbereitung von zur Nachnutzung geeignetem Material. Die folgende Abbildung zeigt, dass die Bereitschaft, Daten in ein Datenarchiv zu überführen abnimmt, je mehr das Material tatsächlich oder vermeintlich solches ist, in dem die Person der Forscherin erkennbar wird. (Abb. 4) Besonders mit Blick auf das Feldtagebuch bzw. Feldnotizen, die als hybrides Material in der Regel nicht nur „Rohdaten" aus Beobachtungen, sondern bereits Interpretationen, aber auch Erfahrungen, Befindlichkeiten, Aversionen enthalten können (Emerson 2011), ist die Skepsis besonders groß.

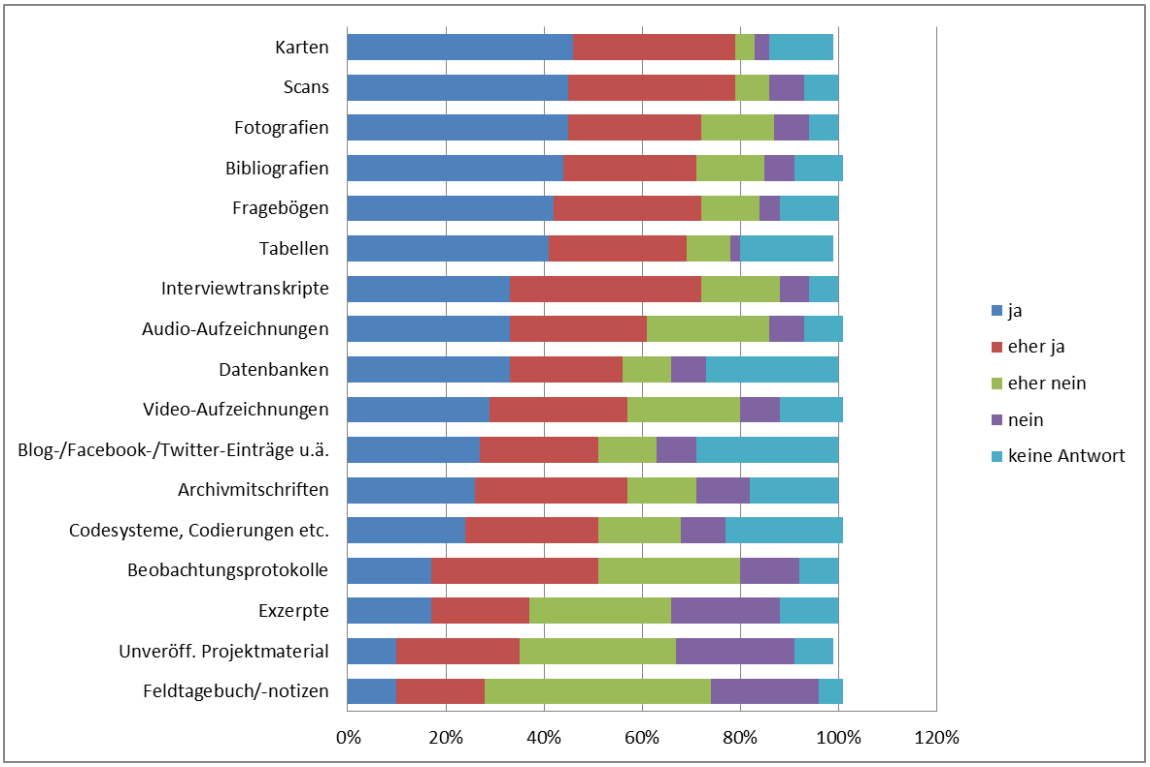


Abbildung 4. Würden Sie die folgenden Datentypen in einem Datenarchiv ablegen? Umfrage des FID Sozial- und Kulturanthropologie zum Forschungsdatenmanagement in den ethnologischen Fächern (2017)

Zur Entwicklung fachspezifischer Kriterien sind weitere Diskussionen notwendig, etwa der Frage, welches Material bei welchem Stand der Aufbereitung es „wert" ist, dauerhaft archiviert zu werden oder welches Material geeignet ist für welche Nutzungsszenarien. Feldtagebücher werden möglicherweise nicht oder nur unter spezifischen Bedingungen dazu gehören. (Vgl. Lederman 2016)

Vertraulichkeit/ Persönlichkeitsschutz

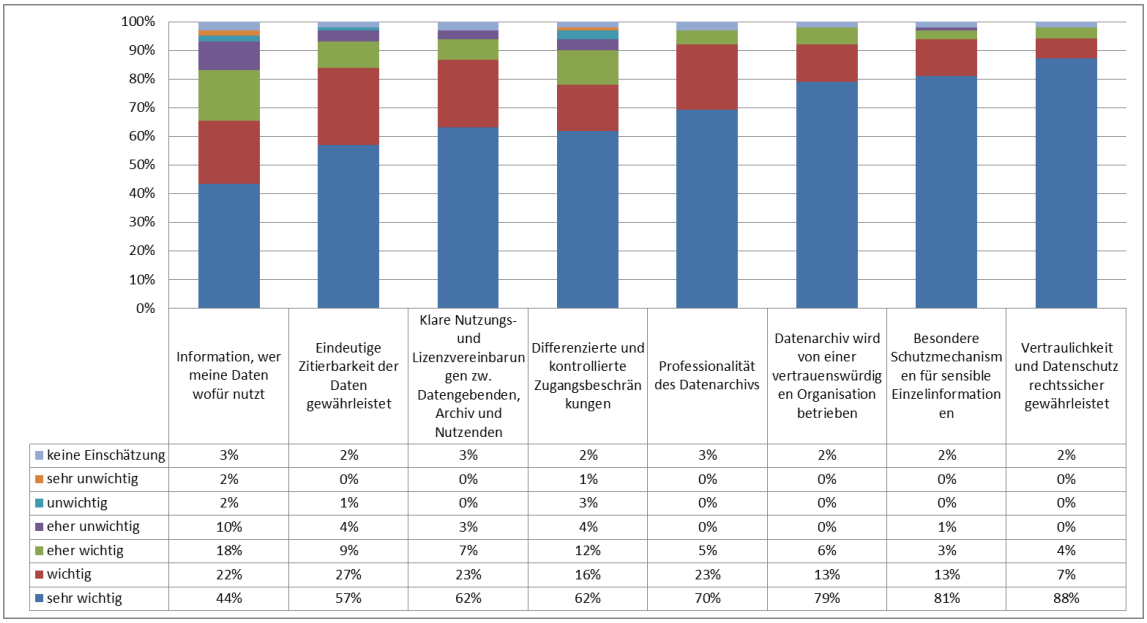
> „In vielen ethnologischen Forschungskontexten bedarf es einer Vertrauensbildung […]. Wenn Transkripte aus solchen Kontexten danach frei zugänglich gemacht werden, unterläuft das systematisch diese Basis."
> „Ein großes ethisches Problem sehe ich darin, dass EthnografInnen aufgrund der [...] teilnehmenden Beobachtung und ihrer Involviertheit in lokale Lebenswelten,

> Dinge erfahren und Informationen erhalten, die [...] auf einem persönlichen Vertrauensverhältnis basieren. Hiermit besteht auch die Verpflichtung, die erhaltenen Informationen sorgfältig zu verwalten und nicht zu externalisieren."
> „Aus der Fachgeschichte sind Fälle bekannt, in denen EthnografInnen Forschungsdaten zurückgehalten und auch vernichtet haben, um sie dem Zugriff anderer, auch dem staatlicher Stellen, zu entziehen, und damit das Vertrauen, das ForschungspartnerInnen in den/die ForscherIn gesetzt haben zu schützen."

Das mit Blick auf Langzeitarchivierung und Data-Sharing wichtigste und kontroverseste Thema ist die Wahrung der Vertraulichkeit in Verbindung mit Fragen von Daten- und Persönlichkeitsschutz sowie Schnittfelder zu forschungsethischen Fragen. Strittig ist zum Beispiel, ob und wie in offenen Feldsituationen Einwilligungserklärungen eingeholt werden können, die dann auch auf Nachnutzungsszenarien ausgedehnt werden müssen. Lösungen im Spannungsfeld von Anonymisierung komplexen ethnografischen Materials und dem Erhalt von dessen Interpretierbarkeit wird hier eine zentrale Rolle zukommen. Zumal sich Konzepte von „Privatheit" oder „Sensibilität" von Daten dynamisch entwickeln und bei Datenübergabe kaum abzusehen ist, auf welchen Wegen offene oder ungenügend anonymisierte Daten künftig Schaden anrichten können. (Vgl. Cliggett 2016: 245) Auch werden vereinzelt Forderungen nach regelrechtem Quellenschutz erhoben, vor allem dann, wenn EthnografInnen in Konfliktfeldern wie z. B. Land- und Ressourcenkonflikten oder politischen Protestbewegungen forschen. Bestenfalls andiskutiert sind überdies Fragen des Urheberrechts, etwa weil die Daten als in der Forschungsbeziehung von Forschenden und Beforschten ko-produziert angesehen werden. (Vgl. grundlegend Fabian 1983)

Anforderungen an Datenarchive

Trotz dieser vielen Unklarheiten, dem großen Diskussions- und Regelungsbedarf und der immer wieder geäußerten Skepsis zeigt sich in den Umfrageergebnissen gleichwohl eine überwiegend positive Grundhaltung gegenüber Datenarchivierung und -nachnutzung genauso wie die Bereitschaft, sich mit dem Thema weiter zu befassen: Rund 65 % derjenigen, die die Umfrage bis zum Schluss bearbeitet haben (N=181), können sich vorstellen, Daten anderer sekundär für die eigene Forschung zu nutzen. Und mehr als 70 % sind *unter bestimmten Bedingungen* bereit, künftig Daten in Repositorien zugänglich zu machen, weitere 15 % sind bereit, genauer darüber nachzudenken. Die Bedingungen für Zugänglichmachung eigener Daten sind hier jedoch entscheidend. Wie im Grunde überwältigend wichtig Fragen nach Persönlichkeits- und Datenschutz sind, zeigt die Abbildung 5.



	Information, wer meine Daten wofür nutzt	Eindeutige Zitierbarkeit der Daten gewährleistet	Klare Nutzungs- und Lizenzvereinbarungen zw. Datengebenden, Archiv und Nutzenden	Differenzierte und kontrollierte Zugangsbeschränkungen	Professionalität des Datenarchivs	Datenarchiv wird von einer vertrauenswürdigen Organisation betrieben	Besondere Schutzmechanismen für sensible Einzelinformationen	Vertraulichkeit und Datenschutz rechtssicher gewährleistet
keine Einschätzung	3%	2%	3%	2%	3%	2%	2%	2%
sehr unwichtig	2%	0%	0%	1%	0%	0%	0%	0%
unwichtig	2%	1%	0%	3%	0%	0%	0%	0%
eher unwichtig	10%	4%	3%	4%	0%	0%	1%	0%
eher wichtig	18%	9%	7%	12%	5%	6%	3%	4%
wichtig	22%	27%	23%	16%	23%	13%	13%	7%
sehr wichtig	44%	57%	62%	62%	70%	79%	81%	88%

Abbildung 5. Wenn Sie Daten in einem Datenarchiv archivieren würden, wie wichtig wären Ihnen folgende Faktoren? Umfrage des FID Sozial- und Kulturanthropologie zum Forschungsdatenmanagement in den ethnologischen Fächern (2017)

Ob Ethnologinnen und Ethnologen ihre Daten in ein Repositorium geben werden, steht und fällt also mit den Antworten, die auf diese Fragen gegeben werden und den Lösungen, die Datenarchive anbieten können. Denn es wird auch befürchtet, dass es sonst schwieriger werden könnte, überhaupt noch GesprächspartnerInnen zu finden. Dass es dabei nicht nur darum geht, deutsches oder EU-Datenschutzrecht einzuhalten, ist deutlich geworden. Mit Blick auf ethnografische Daten hat „Open Data“ Grenzen, Datenpublikationen im eigentlichen Sinn werden nur in wenigen Fällen möglich sein. Deutlich ist auch, dass dies keine Verweigerungshaltung ist, sondern dass es dafür gute forschungsimmanente Gründe gibt. Entsprechend wird seitens eines Datenarchivs dauerhaft ein umfassendes Rechtemanagement benötigt, das nicht intendierte Verwertungen von Daten verhindert, gestufte Zugangsrechte bis hin zur On-Site-Nutzung einschließt und vor allem Zugangskontrolle zuverlässig gewährleistet, also insgesamt einen hohen Grad an Professionalisierung aufweist.[10] Auch Exit-Strategien bzw. Möglichkeiten, archivierte Daten wieder zurückzuziehen, müssen diskutiert werden.

Viele der Fragen die hier aufgeworfen werden sind solche, die in den Fachcommunities diskutiert und bearbeitet werden müssen, die ihrerseits auch oft aufgefordert werden, das zu tun. (Vgl. z. B. RatSWD 2015) Wie Klärungen herbeigeführt bzw. wie Rückkopplungen an die Forschenden in Fachcommunities realisiert werden können, die sich bisher kaum mit Forschungsdatenmanagement befasst haben, scheint jedoch bisher weniger diskutiert.[11] Der FID Sozial- und Kulturanthropologie sieht sich im Moment in der Rolle, Diskussionsprozesse anzustoßen, zu befördern, wo möglich auch zu bündeln und positioniert sich als Ansprechpartner. Der wissenschaftliche Beirat des FID ist hier ein wichtiges beratendes Gremium, einbezogen sind überdies die relevanten Fachgesellschaften sowie fachlich einschlägige Forschungsprojekte. In dieser Zusam-

10 Ähnlich, wie es etwa Qualiservice an der Universität Bremen praktiziert und weiter erprobt. http://www.qualiservice.org/, vgl. Kretzer 2013.

11 Gleichwohl arbeiten Projekte an der Entwicklung und Erprobung entsprechender Strategien, vgl. etwa Helbig and Aust 2017.

menarbeit soll entlang konkreter Forschungspraxis ein Musterworkflow ebenso entwickelt und abgestimmt werden wie fachspezifische Empfehlungen zum Umgang mit Forschungsdaten.

Literaturangaben

Breidenstein, Georg, Stefan Hirschauer, Herbert Kalthoff and Boris Nieswand. 2015 *Ethnografie. Die Praxis der Feldforschung*. Konstanz u.a.: UVK Verlagsgesellschaft mbH.

Cliggett, Lisa, 2016. "Preservation, Sharing and Technological Challenges of Longitudinal Research in the Digital Age." In *eFieldnotes. The makings of anthropology in the digital world*. ed. by Roger Sanjek and Susan W. Tratner, 231-250. Philadelphia: University of Pennsylvania Press.

Droß, Patrick. 2015. *Kurzstudie: Anforderungen an die Archivierung sozial- und wirtschaftswissenschaftlicher Forschungsdaten*. WZB Berlin. https://sowidatanet.de/images/pdfs/Meldungen/Kurzstudie_Qualitative_Interviews.pdf [05.04.2017]

Emerson, Robert M., Rachel I. Fretz and Linda L. Shaw. 2011. *Writing Ethnographic Fieldnotes*. Chicago. London: University of Chicago Press.

Fabian, Johannes. 1983. *Time and the Other. How Anthropology Makes Its Object.* New York: Columbia University Press.

Heinrich, Maurice, Sabine Jahn and Felix Schaefer. 2014. *Stakeholderanalyse 2013 zu Forschungsdaten in den Altertumswissenschaften. Teil 1: Ergebnisse. Rohdaten* [Version 1.0] ed. by IANUS. doi:10.13149/000.jah37w-q

Heinrich, Maurice and Felix Schäfer. 2013. *Fragebogen zur Stakeholderanalyse 2013 – zu Forschungsdaten in den Altertumswissenschaften.* [Version 1.0] ed. by IANUS. doi: 10.13149/000.jah37w-q

Helbig, Kerstin and Pamela Aust. 2017. „Kein Königsweg - die Vermittlung von Forschungsdatenkompetenz auf allen universitären Ebenen." In *o-bib. Das offene Bibliotheksjournal*, 108-116. http://dx.doi.org/10.5282/o-bib/2017H1S108-116

IANUS ed. 2017. *IT-Empfehlungen für den nachhaltigen Umgang mit digitalen Daten in den Altertumswissenschaften* [Version 1.0.0.0] doi: 10.13149/000.111000-a

Knecht, Michi. 2009. "Contemporary Uses of Ethnography. Zur Politik, Spezifik und gegenwartskulturellen Relevanz ethnographischer Texte." In *Bilder, Bücher, Bytes, 36. Kongress der Deutschen Gesellschaft für Volkskunde 2007* ed. by Michael Simon, Timo Heimerdinger and Thomas Hengartner, 148-15, Münster: Waxmann Verlag.

Knecht, Michi. 2013. “Nach Writing Culture, mit Actor-Network: Ethnographie/ Praxeographie im Feld der Wissenschafts-, Medizin- und Technikanthropologie.“ In *Europäisch-ethnologisches Forschen. Neue Methoden und Konzepte*, ed. by Sabine Hess, Johannes Moser and Maria Schwertl, 79-106, Berlin: Reimer.

Kretzer, Susanne. 2013. “Infrastruktur für qualitative Forschungsprimärdaten - Zum Stand des Aufbaus eines Datenmanagementsystems von Qualiservice.“ In *Forschungsinfrastrukturen für die qualitative Sozialforschung* ed. by. Denis Huschka et.al., 91-107, Berlin: SCIVERO Verlag, https://www.ratswd.de/dl/downloads/forschungsinfrastrukturen_qualitative_sozialforschung.pdf [02.04.2017]

Lederman, Rena. 2016. “Archiving Fieldnotes? Placing ‘Anthropological Records’ Among Plural Digital Worlds.” In *eFieldnotes. The makings of anthropology in the digital world*. ed. by Roger Sanjek and Susan W. Tratner, 251-271. Philadelphia: University of Pennsylvania Press.

Opitz, Diane and Reiner Mauer. 2005. „Erfahrungen mit der Sekundärnutzung von qualitativem Datenmaterial – Erste Ergebnisse einer schriftlichen Befragung im Rahmen der Machbarkeitsstudie zur Archivierung und Sekundärnutzung qualitativer Interviewdaten“ [50 Absätze]. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 6 (1), Art. 43, urn:nbn:de:0114-fqs0501431

RatSWD. 2015. *Archivierung und Sekundärnutzungen von qualitativen Daten. Eine Stellungnahme des RatSWD*. Output Series 1/ 2015. https://www.ratswd.de/publikationen/output [06.04.2017]

Simon, Michael. 2015. “Ethnologische Anmerkungen zu Bernd Riekens ‚Gesprächen mit Einheimischen‘ in Galtür.“ In *Wie bewältigt man das Unfassbare? Interdisziplinäre Zugänge am Beispiel der Lawinenkatastrophe von Galtür*, ed. by Bernd Rieken, 93-105. Münster u.a.: Waxmann.

Simukovic, Elena, Maxi Kindling and Peter Schirmbacher 2013. *Umfrage zum Umgang mit digitalen Forschungsdaten an der Humboldt-Universität zu Berlin. Umfragebericht* [Version 1.0] urn:nbn:de:kobv:11-100213001

Stodulka, Thomas. 2014. „Feldforschung als Begegnung – Zur pragmatischen Dimension ethnographischer Daten.“ In *Sociologus* vol. 64, Issue 2, 179-206.

Welz, Gisela. 2013 “Die Pragmatik ethnografischer Temporalisierung. Neue Formen der Zeitorganisation in der Feldforschung.“ In *Europäisch-ethnologisches Forschen. Neue Methoden und Konzepte*, ed. by Sabine Hess, Johannes Moser and Maria Schwertl, 39-54. Berlin: Reimer.

Projekt DataWiz: Entwicklung eines Assistenzsystems zum Management psychologischer Forschungsdaten

Martin Kerwer[1], Ronny Bölter[2], Ina Dehnhard[3], Armin Günther[4], Erich Weichselgartner[5]

1,2,3,4,5 Leibniz-Zentrum für Psychologische Information und Dokumentation, ZPID, Trier

Zusammenfassung. Über Disziplinen hinweg werden Nullbefunde kaum veröffentlicht und Daten selten geteilt. Selbst auf Anfrage sind Daten, die Publikationen zugrunde liegen, häufig nicht mehr verfügbar. Neben den damit verbundenen wissenschaftstheoretischen Problemen führt dies zu einem ineffizienten Einsatz öffentlicher Gelder, weswegen Forschungsförderer und Fachorganisationen verstärkt für einen besseren Umgang mit Forschungsdaten und deren freie Verfügbarkeit eintreten. Die (deutsche) Psychologie nimmt hierbei mit ihren neuen Empfehlungen zum Umgang mit Forschungsdaten eine Vorreiterrolle ein. Um den Anforderungen, die sich durch ein hochwertiges Forschungsdatenmanagement stellen, gerecht werden zu können, benötigen Forschende aber auch Werkzeuge, die ihnen die Bewältigung dieser Anforderungen mit möglichst wenig Zusatzaufwand ermöglichen. Die Entwicklung und Bereitstellung solcher Werkzeuge stellt eine disziplinübergreifende Herausforderung dar. Mit dem Projekt DataWiz wird die Entwicklung eines Forschungsdatenmanagement-Werkzeugs, das diese Herausforderung in der Psychologie angeht, als Fallstudie vorgestellt.

Das derzeit am ZPID entwickelte computergestützte Assistenzsystem DataWiz unterstützt Psychologen/innen in ihrem täglichen Forschungsdatenmanagement wissensbasiert und prozedural. Wissensbasierte Unterstützung findet innerhalb der Anwendung in Form einer durchgängigen, kontextsensitiven Bereitstellung von Informationen zu forschungsdatenrelevanten Themenbereichen (Datenschutz, Urheberrecht, Qualitätssicherung, usw.) statt.

Gleichzeitig wird Forschenden mit DataWiz aber auch ein Werkzeug zur Verfügung gestellt, das ihnen bei der Bewältigung typischer Aufgaben des Forschungsdatenmanagements assistiert. Die Webanwendung verbindet grundlegendende Funktionen, die das kollaborative Arbeiten in einer virtuellen Forschungsumgebung ermöglichen, wie einer abgestuften Nutzerrechtevergabe und ein entsprechend abgestufter Zugang zu abgelegten Dateien, mit prozeduralen Funktionen, die Nutzer/innen im Forschungsprozess begleiten. Dazu zählen eine angeleitete Planung des Forschungsdatenmanagements, die fachspezifische Dokumentation der Datenerhebung und der erhobenen Forschungsdaten, sowie die Aufbereitung der Forschungsdaten und ihre Übergabe an fachlich fundierte und etablierte Datenarchive. DataWiz zeichnet sich hierbei unter anderem durch den einfachen Import von Datenmatrizen und zugehöriger Metadaten aus disziplinüblicher Software, einer Verknüpfung von Datenmanagementplanung, Codebucherstellung und Datenerhebungsdokumentation, sowie die Unterstützung des Forschenden in der Datenaufbereitung und Versionierung aus.

Neben einer frei zugänglichen, am ZPID gehosteten Version, wird Nutzer/innen die Möglichkeit gegeben, eigene Versionen der Anwendung an ihrer Institution zu betreiben. Um auch das interdisziplinäre Nachnutzungspotenzial der Daten zu erhöhen, werden disziplinübergreifende Metadatenstandards unterstützt. Die Übergabe der Daten an Forschungsdatenarchive verwandter Disziplinen wird damit stark vereinfacht. Darüber hinaus wird der Quellcode des Programms der Community offen zur Verfügung gestellt, wodurch DataWiz eine potentielle Basis fächerübergreifender Entwicklungen darstellt.

Schlagwörter. Forschungsdaten, Data Sharing, Data Management

Theoretischer Hintergrund

Forderung nach Data Sharing

Die Forschung der letzten Jahre hat wiederholt gezeigt, dass die Replizierbarkeit wissenschaftlicher Befunde zu wünschen übriglässt (z.B. Open Science Collaboration, 2015). Eine Ursache hierfür ist der als File-Drawer-Problem bekannte Sachverhalt, dass Nullbefunde häufig nicht berichtet werden. Dieser führt zu Publication Bias - einer Verzerrung der in der Literatur berichteten Ergebnisse von Studien. Diese Verzerrung ist aus methodischer Sicht höchst problematisch, da die Ergebnisse von Sekundäranalysen nicht zuverlässig für Publication Bias korrigiert werden können (Scargle, 2000). Ein weiteres Problem liegt darin, dass häufig auch nicht die kompletten Ergebnisse einer Studie berichtet werden. Stattdessen findet vermehrt eine auszugsweise Berichterstattung signifikanter Ergebnisse statt (sog. P-Hacking, Simonsohn, Nelson, Simmons, 2014). Diese verzerrte Berichterstattung innerhalb von Studien wird auch als Reporting Bias bezeichnet.

Beide Problemstellungen können dadurch gelöst werden, dass die Hypothesen einer Studie präregistriert und die kompletten Rohdaten, die der Publikation zugrunde liegen, veröffentlicht werden. Abweichungen vom ursprünglichen Analyseplan werden damit transparent gemacht und alternative Spezifikationen der Analysen können auf Grundlage dieser Daten getestet werden. Das Potenzial der Bereitstellung von Forschungsdaten – des Data Sharing – reicht aber weit über diesen Anwendungshorizont hinaus, wenn Daten nicht nur zur Qualitätssicherung verfügbar gemacht werden, sondern auch zur weiteren Nutzung bereitgestellt werden. Die Vorteile des Data Sharing umfassen in diesem Fall unter anderem eine verbesserte Nutzung der Ressourcen, die für Forschungsvorhaben aufgebracht werden, die Vermeidung einer unnötigen Beanspruchung von Versuchsteilnehmern und die Erhaltung einmaliger, nicht replizierbarer Datenbestände.

Die breite Wahrnehmung der Replikationskrise und des potenziellen Nutzens des Data Sharing hat dazu geführt, dass die freie Verfügbarkeit von Forschungsdaten in den letzten Jahren immer stärker ins Blickfeld wissenschaftlicher Akteure gerückt ist. Zu diesen Akteuren zählen unter anderem nationale und internationale Förderorganisationen, wie die Deutsche Forschungsgemeinschaft (Richtlinien zum Umgang mit Forschungsdaten, Deutsche Forschungsgemeinschaft, 2015) und die Europäische Union (Ausweitung des Open Data Pilot des Horizon 2020 Programms, European Commission, 2016, Art 29.3.), Publisher (z.B. PLOS[1], Society for Judgment and Decision Making[2]) und wissenschaftliche Fachgesellschaften. Die Deutsche Gesellschaft für Psychologie (DGPs) beispielsweise verabschiedete fachspezifische Empfehlungen zum Datenmanagement in der Psychologie (Schönbrodt, Gollwitzer, & Abele-Brehm, 2016), die sich für eine freie Verfügbarkeit der Daten, die Publikationen zugrunde liegen, einsetzen. Vor dem Hintergrund, dass die Bereitschaft Daten zu teilen in der Psychologie unverändert niedrig ist (Dehnhard, Weichselgartner und Krampen, 2013, Vanpaemel, Vermogen, Deriemaecker und Storms, 2015, Wicherts, Borsboom, Kats und Molenaar, 2006), sollte die Bedeutung dieser Empfehlungen nicht unterschätzt werden.

1 http://journals.plos.org/plosone/s/data-availability (abgerufen am 08.02.17)
2 http://journal.sjdm.org/(abgerufen am 08.02.17)

Forderung nach Datenmanagement

Zunehmend wird außerdem erkannt, dass die reine technische Verfügbarkeit von Forschungsdaten nicht genügt, um ihre Nachnutzbarkeit zu ermöglichen. Neben der technischen Verfügbarkeit der Daten muss nämlich hierzu auch die langfristige Verständlichkeit der Daten sichergestellt werden. Für die Erhaltung der Interpretierbarkeit bedarf es zusätzlicher Informationen zu den Daten, sogenannter Metadaten. Diese können zum Teil genereller und damit fächerübergreifender Natur sein, in den meisten Disziplinen werden aber auch fachspezifische Informationen benötigt. Die Forderung nach einem fachgerechten Datenmanagement wird deswegen immer lauter (z.B. Council of the European Science, 2016, Schönbrodt, Gollwitzer, & Abele-Brehm, 2016). Fachspezifische Besonderheiten in der Psychologie, die spezifisches Wissen und Kompetenzen für das Datenmanagement verlangen, sind zum Beispiel der Einsatz einer großen Bandbreite an (häufig kaum dokumentierten) Datenerhebungsverfahren, ein hoher Anteil an Daten aus Studien ohne institutionalisiertes Forschungsdatenmanagement, oder das Vorliegen personenbezogener Daten, die besondere datenschutzrechtliche Kenntnisse erfordern (Weichselgartner, 2011). Das entsprechende Wissen muss Forschenden so einfach wie möglich zur Verfügung gestellt werden. Fraglich ist aber auch, ob Forscher/innen ohne prozessorientierte Unterstützung in der Lage sind, datenmanagementbezogene Aufgaben fachgerecht zu bewältigen. So konnten Vines et al. (2014) zeigen, dass die Verfügbarkeit von Daten, die Publikationen zugrunde liegen, mit zunehmendem Alter der Publikation stark abnimmt. Dies galt unabhängig davon, ob Forscher gewillt waren ihre Daten zu teilen oder nicht. Die Kompetenz von Forschern, ihre Forschungsdaten (auch für die eigene Nutzung) zu erhalten, muss damit angezweifelt werden.

Wir müssen uns also fragen, wie wir Forschende dazu befähigen können den Anforderungen, die ein fachgerechtes Datenmanagement an sie stellt, gerecht zu werden. Dabei ist der (wahrgenommene) Arbeitsaufwand, der hierfür notwendig ist, eine kritische Größe. Denn auch wenn Data Sharing mittel- und langfristig gesehen helfen kann, Ressourcen einzusparen, bindet ein sachgemäßes Forschungsdatenmanagement zunächst einmal Ressourcen. Die Fragmentierung datenmanagementbezogener Informationen und unterstützender Angebote (z.B. zur Planung des Forschungsdatenmanagements, zur Dokumentation und Aufbereitung der Daten selbst oder zur Übergabe der Daten an ein Archiv) trägt sicher zum wahrgenommenen Arbeitsaufwand bei. Eine Reduktion dieses Aufwands kann dadurch erreicht werden, dass Datenmanagement frühzeitig in den Forschungsprozess integriert wird (Gutmann et. al., 2009). Die frühzeitige Integration in den Forschungsprozess erlaubt es, Informationen dort abzugreifen, wo sie entstehen. Auf diese Weise wird die mühsame Rekonstruktion von Informationen am Ende eines Forschungsprojekts vermieden.

Darüber hinaus muss berücksichtigt werden, dass Forscher ein berechtigtes Interesse daran haben, zunächst selbst in ihrer Arbeitsgruppe die Kontrolle über ihre Forschungsdaten zu behalten. Diese Rückmeldung aus der Fachgemeinschaft hat das ZPID im Betrieb des Forschungsdatenzentrums PsychData immer wieder erhalten und auch die Teilnehmer des DataWiz-KickOff-Workshops haben sich entsprechend geäußert. Wichtig ist deshalb, dass Forschende die Bedingungen des Datenzugangs selbst festlegen können (Simukovic, Kindling und Schirmbacher, 2013).

Anforderungen an Forschungsdatenmanagement-Tools

Der Druck auf Forschende, hochwertige Daten bereitzustellen, muss also von Angeboten begleitet werden, die Forschende bei der Bewältigung dieser neuen Anforderungen unterstützen. Forschungsdatenmanagement-Werkzeuge können dies leisten, wenn sie empirisch arbeitende Forscher und Forscherinnen in die Lage versetzen, selbstständig ein hochwertiges Forschungsdatenmanagement durchzuführen. Folgenden Anforderungen an die Entwicklung solcher „Assistenzsysteme“ lassen sich aus unseren vorangegangen Überlegungen ableiten:

- Forschungsdatenmanagement-Werkzeuge müssen nicht nur dabei helfen, die Verfügbarkeit der Daten, sondern auch ihre Interpretierbarkeit sicherzustellen.
- Fachspezifische Erfordernisse müssen berücksichtigt werden.
- Forschende brauchen einen einfachen Zugriff (single-point access) auf relevante inhaltliche Informationen zum Datenmanagement (wissensbasierte Unterstützung).
- Forschende brauchen Unterstützung bei der Bewältigung zentraler Aufgaben des Datenmanagements (prozedurale Unterstützung).
- Forschende müssen frühzeitig und über den gesamten Forschungsprozess hinweg unterstützt werden.
- Forschende müssen die Kontrolle über ihre Daten (zunächst) behalten. Die Freigabe der Daten für andere (Data Sharing) darf nur optional sein.
- Insgesamt ist der Aufwand, den Forschende in ein adäquates Forschungsdatenmanagement investieren müssen, zu minimieren.

Aus Perspektive der Forschungsdateninfrastruktur (insbesondere Forschungsdatenzentren, Repositorien) hätte ein durch entsprechende Tools assistiertes Forschungsdatenmanagement gegenüber einem nicht-assistierten Datenmanagement den wesentlichen Vorteil, dass der erforderliche Archivierungsaufwand bei der Übernahme der Daten durch eine höhere Qualität der Einreichungen reduziert werden würde.

Das Assistenzsystem DataWiz

Am Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID) wird im DFG-geförderten Projekt DataWiz derzeit ein gleichnamiges webbasiertes Assistenzsystem entwickelt, das sich an diesen Anforderungen orientiert und zum Ziel hat, Psychologinnen und Psychologen zur selbstständigen Umsetzung eines fundierten Forschungsdatenmanagements im Forschungsprozess zu befähigen. Im Folgenden werden wir auf den Programmaufbau, technologische Grundlagen, sowie Nachnutzungspotenziale und Perspektiven, die sich aus der Entwicklung ergeben, eingehen.

Programmaufbau

DataWiz gliedert sich in ein System zur wissensbasierten Unterstützung, ein System zur prozeduralen Unterstützung und übergreifende Systemfunktionen, die das kollaborative Arbeiten im Team ermöglichen. Die prozedurale Unterstützung entlang des Forschungsdatenlebenszyklus wird

so implementiert, dass Redundanzen (Mehrfacheingabe gleicher Informationen) möglichst komplett vermieden werden, während gleichzeitig die wissensbasierte Unterstützung kontextsensitiv in die prozedurale Unterstützung integriert wird.

Wissensbasierte Unterstützung

Zur inhaltlichen Unterstützung der Forschenden wurde eine auf die Bedürfnisse der Psychologie ausgerichtete Wissensbasis zum Datenmanagement erstellt, die den Nutzern grundlegende Informationen zu forschungsdatenrelevanten Themenbereichen (Datenschutz, Urheberrecht, Qualitätssicherung, usw.) in der Psychologie bereitstellt und als kontextsensitives Hilfesystem fungiert. Dies führt zu einer erheblichen Aufwandsreduktion für Forschende, da die Zusammenstellung und Bewertung dieser Informationen sonst erhebliche Ressourcen erfordert. Außerdem werden Fehlentscheidungen vermieden, die das spätere Data Sharing möglicherweise erschweren oder sogar verhindern (z.B. aufgrund der Verwendung ungeeigneter, überrestriktiver Einwilligungserklärungen bei der Erhebung psychologischer Daten).

Prozessorientierte Unterstützung

Die prozessorientierte Unterstützung der Forschenden bei der Bewältigung grundlegender Aufgaben des Forschungsdatenmanagements beinhaltet Funktionalitäten zur Planung des Forschungsdatenmanagements, Datenablage, Datendokumentation, Qualitätskontrolle und zum Data Sharing sowie die Integration dieser Funktionalitäten in ein Gesamtsystem.

Datenmanagementplanung

Zu Beginn ihres Projekts unterstützt DataWiz Forscher und Forscherinnen in der Planung ihres Forschungsdatenmanagements. Informationen der Wissensbasis fungieren hierbei als Hilfestellung, während die in der Planungsphase abgefragten Informationen sich an den Auflagen großer Förderorganisation (DFG, EU, BMBF) orientieren.

Datenablage

Nach der Datenerhebung können Forschende ihre Forschungsdaten in DataWiz ablegen. DataWiz unterstützt die herstellerunabhängige Verwaltung von Forschungsdaten, die als rechteckige Datenmatrizen vorliegen, in offenen Formaten. Um den Aufwand des Imports und Exports von Datenmatrizen zu reduzieren und die Integration in den psychologischen Forschungsprozess voranzutreiben, wurden Import- und Exportroutinen für die psychologieübliche Software SPSS sowie Dateien mit festem Trennzeichen (delimiter-separated values) entwickelt.

Datendokumentation

Ein Codebucheditor erlaubt die Dokumentation der hochgeladenen Forschungsdaten durch die Erfassung und Bearbeitung von Metadaten auf Variablenebene. Importroutinen aus disziplinspezifischer Software vereinfachen diesen Prozess erheblich.

Die Studiendokumentation stellt sicher, dass der Kontext der Datenerhebung verstanden wird und Daten damit bewertet oder nachgenutzt werden können. In der Entwicklung wurde hierbei auf eine größtmögliche Kompatibilität zum Metadatenschema des durch das ZPID-betriebenen Forschungsdatenzentrums PsychData und einschlägigen fachspezifischen Berichtsstandards geachtet. Interoperabilität zu verwandten Disziplinen und Maschinenlesbarkeit der Metadaten wird durch Kodierung nach dem Data Documentation Initiative (DDI)-Metadatenstandard[3] gefördert.

Qualitätskontrolle

Zur Unterstützung der Qualitätssicherung werden bei Eingabe von Metadaten in Codebücher automatisiert Konsistenzprüfungen durchgeführt, wodurch Inkonsistenzen vermieden werden. Gleichzeitig sorgt ein automatisiertes System zur Versionierung von Forschungsdaten und Codebüchern dafür, dass die Entstehung des Datensatzes nachvollzogen werden kann.

Data Sharing

Die Weitergabe der Daten und Metadaten im Forschungsprojekt wird durch die übergreifenden Systemfunktionen zum kollaborativen Arbeiten unterstützt. Exportroutinen zur Datenübergabe an das psychologische Forschungsdatenzentrum PsychData sind bereits im Design des Assistenzsystems angelegt, da das Metadatenschema des Assistenzsystems auf einem Ausbau des PsychData-Metadatenschemas basiert. Für die Übergabe von Forschungsdaten und Metadaten aus DataWiz an andere Archive werden Exportfunktionen entwickelt, die Daten und Metadaten in offenen Formaten zusammenstellen und den Arbeitsaufwand für Forscher damit reduzieren.

Integration des Workflows

Der Arbeitsaufwand für Forschende wird durch eine Integration des Forschungsdatenmanagement-Workflows weiter reduziert, da keine Redundanzen zwischen den verschiedenen in DataWiz angelegten prozeduralen Funktionalitäten anfallen. So können Nutzer beispielsweise in der Planung ihres Forschungsdatenmanagements angeben, welche Konstrukte (nicht direkt beobachtbare Personenmerkmale) in ihrem Projekt untersucht werden sollen. Im weiteren Forschungsprozess kann dann ausgewählt werden, in welcher Teil-Untersuchung welches Konstrukt erhoben wurde. Schließlich wird bei der Erstellung eines Codebuchs dokumentiert, welche Variable eines hochgeladenen Datensatzes sich welchem Konstrukt zuordnen lässt. Gleichzeitig erhält der Forschende über die Wissensbasis des Programm Informationen dazu, welche Vokabulare und Ontologie ihm zur Verfügung stehen, um das Konstrukt in einer nachnutzbaren Form zu beschreiben.

3 http://www.ddialliance.org/ (abgerufen am 08.02.17)

Übergreifende Systemfunktionen

Natürlich fallen im Forschungsprozess neben den eigentlichen Forschungsdaten Projektoutputs an, die für das Verständnis von Daten oder Analysen bzw. für die Zusammenarbeit im Projekt notwendig sind. DataWiz verfügt deshalb über ein System zur Verwaltung beliebiger digitaler Zusatzmaterialien. Um als kollaborative Arbeitsplattform nutzbar zu sein, wurde zudem eine abgestufte Zugriffs- und Rechtekontrolle implementiert, die die Vergabe differenzierter Zugriffsrechte auf Forschungsdaten, Metadaten und Dateien an andere Nutzer erlaubt.

Technologie

Die Implementierung von DataWiz erfolgt in Java (Enterprise Edition) unter Hinzunahme von weiteren, gängigen Technologien. Hierzu zählt im Frontend Bereich neben den üblichen Standards wie HTML5, CSS3 und JavaScript vor allem die Hinzunahme von Bootstrap, jQuery und einigen kleineren jQuery Libraries. Bei der Entwicklung des Backend Bereiches wurde neben Java/Java EE vor allem weite Teile des Spring MVC Frameworks verwendet. Zusätzlich kommt die Spring Security Erweiterung zum Einsatz, um das System vor Angriffen abzusichern. Zur Speicherung hochgeladener Datensätze in Dateiform (und anderer Zusatzmaterialien) wurde auf das Minio Cloud Storage System zurückgegriffen. DataWiz legt nicht nur die Originaldatensätze in diesem System ab, sondern führt während des Importprozesses auch Konsistenzchecks auf den importierten Datenmatrizen und Codebüchern aus, die anschließend in einer MySQL Datenbank abgelegt werden. Diese redundante Speicherung hat den Vorteil, dass die Originaldatei erhalten bleibt und dem Datengeber kontinuierlich zur Verfügung steht. Gleichzeitig liegen die Daten in einer aufbereiteten Form vor, welche für den späteren Export in Langzeitarchive genutzt werden kann. Durch dieses Vorgehen ist es möglich, Änderungen an den Datensätzen von DataWiz nachzuvollziehen (Data Provenance). Die Wissensbasis wird im Content-Management-System Wordpress entwickelt. Prozedurale Funktionalitäten sind auf Deutsch und Englisch implementiert, während die Wissensbasis in der internationalen Psychologie-Fachsprache Englisch bereitgestellt wird.

Entwicklungsstand und Release

Ab Anfang 2017 ist eine erste Fassung der Wissensbasis unter www.datawiz.de zugänglich. Im Laufe des Frühjahrs 2017 sollen Import- und Export-Prozeduren für SPSS-Dateien als eigenständige Komponente auf GitHub bereitgestellt werden. Das Release der finalen Version und die freie Bereitstellung des Source Codes ist für Ende 2017 geplant.

Resümee

Der psychologischen Community wird DataWiz am Projektende durch eine frei zugängliche am ZPID gehostete Version zur Verfügung gestellt. DataWiz bietet aber darüber hinaus erhebliches Nachnutzungspotenzial gerade im Hinblick auf die zukünftige Entwicklung und Bereitstellung von Forschungsdatenmanagement-Werkzeugen im interdisziplinären Kontext.

Nachnutzungspotenziale

Durch die externe Bereitstellung des Source Code ist das DataWiz System zur Nachnutzung zugänglich. Nachnutzungspotenzial besteht hierbei (1) im Betrieb lokaler Installationen von DataWiz an der eigenen Institution, (2) in der Adaption von DataWiz (als prototypische Lösung) an andere Disziplinen und der Anpassung des Programms gemäß den Besonderheiten dieser Disziplinen und (3) in der Nachnutzung bestimmter Module des Source Codes, wie den Import- und Exportfunktionalitäten von SPSS-Dateien, unabhängig vom Gesamtkontext des Projekts.

Des Weiteren wird das interdisziplinäre Nachnutzungspotenzial der durch Psychologen in DataWiz erzeugten Forschungsdaten gestärkt, da aus DataWiz exportierte Metadaten und Daten in nicht-proprietären Formaten zugänglich sind. Die Interoperabilität der in DataWiz dokumentierten Daten wird darüber hinaus durch Unterstützung des sozialwissenschaftlichen DDI-Metadatenstandards erheblich verbessert.

Literaturangaben

Council of the European Science. 2016. *Council conclusions on the transition towards an open science system*. Online verfügbar unter http://data.consilium.europa.eu/doc/document/ST-9526-2016-INIT/en/pdf. Zuletzt geprüft am 08.02.2017.

Dehnhard, I., E. Weichselgartner, und G. Krampen. 2013. “Researcher’s willingness to submit data for data sharing: A case study on a data archive for psychology”. *Data Science Journal* 12:172-180. doi: 10.2481/dsj.12-037.

Deutsche Forschungsgemeinschaft. 2015. *Leitlinien zum Umgang mit Forschungsdaten.* Online verfügbar unter http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf. Zuletzt geprüft am 08.02.2017.

European Commission. 2016. H2020 Programme: AGA – annotated model grant agreement (Version 2.2). Online verfügbar unter http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf. Zuletzt geprüft am 08.02.2017.

Gutmann, M. P., M. Abrahamson, M. O. Adams, M. Altman, C. Arms, K. Bollen, … C. H. Young 2009. “From preserving the past to preserving the future: The Data-PASS project and the challenges of preserving digital social science data”. *Library Trends* 57 (3): 315-337. doi:10.1353/lib.0.0039

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349 (6251): aac4716–1 - aac4716-8. http://doi.org/10.1126/science.aac4716.

Scargle, J. D. 2000. “Publication bias: The "File Drawer" problem in scientific inference”. *Journal of Scientific Exploration* 14: 91-106.

Schönbrodt, F., M. Gollwitze, und A. Abele-Brehm. 2016. *Data management in psychological science: Specification of the DFG guidelines*. Online verfügbar unter http://www.dgps.de/fileadmin/documents/Empfehlungen/Data_Management_in_Psychological_Science_20160928.pdf. Zuletzt geprüft am 08.02.2017.

Simonsohn, U., L. D. Nelson, und J. P. Simmons. 2014. „P-curve: a key to the file-drawer". *Journal of Experimental Psychology: General* 143 (2): 534.

Simukovic, E., M. Kindling, und P. Schirmbacher. 2013. *Forschungsdaten an der Humboldt-Universität zu Berlin: Bericht über die Ergebnisse der Umfrage zum Umgang mit digitalen Forschungsdaten an der Humboldt-Universität zu Berlin* (Umfragebericht, Version 1.0). Online verfügbar unter http://nbn-resolving.de/urn:nbn:de:kobv:11-100213001. Zuletzt geprüft am 08.02.2017.

Vanpaemel, W., M. Vermorgen, L. Deriemaecker, und G. Storms. 2015. „Are we wasting a good crisis? The availability of psychological research data after the storm". *Collabra* 1 (1): 1–5. http://doi.org/10.1525/collabra.13.

Vines, T. H., A .Y. K. Albert, R. L. Andrew, F. Débarre, D. G. Bock, M. T. Franklin, … D. J. Rennison. 2014. The availability of research data declines rapidly with article age. *Current Biology* 24(1): 94–97.

Weichselgartner, E. 2011. *Disziplinenspezifische Aspekte des Archivierens von Forschungsdaten am Beispiel der Psychologie (RatSWD Working Paper Series, Nr. 179)*. Berlin: Rat für Sozial- und Wirtschaftsdaten.

Wicherts, J. M., D. Borsboom, J. Kats, und D. Molenaar. 2006. „The poor availability of psychological research data for reanalysis". *The American Psychologist* 61 (7): 726-728. doi:10.1037/0003- 066X.61.7.726

eDissPlus – Optionen für die Langzeitarchivierung dissertationsbezogener Forschungsdaten aus Sicht von Bibliotheken und Forschenden

Dirk Weisbrod[1], Ben Kaden[2], Michael Kleineberg[3]

1 Deutsche Nationalbibliothek
2,3 Universitätsbibliothek, Humboldt-Universität zu Berlin

Zusammenfassung. Im Rahmen des DFG-Projektes "Elektronische Dissertationen Plus" (eDissPlus) beschäftigen sich die Universitätsbibliothek der Humboldt-Universität zu Berlin (HU) und die Deutsche Nationalbibliothekbibliothek (DNB) mit konzeptionellen und technischen Anforderungen für eine zeitgemäße Archivierung und Publikation von Forschungsdaten, die im Zusammenhang mit Promotionsvorhaben entstehen. Der Beitrag diskutiert zunächst die Verantwortlichkeiten für die Veröffentlichung und Langzeitarchivierung dissertationsbezogener Forschungsdaten auf nationaler Ebene und über Community-Grenzen hinweg. Dabei stellen sich insbesondere die Fragen, welche Forschungsdaten überhaupt zu einer Dissertation gehören und unter welchen Umständen die DNB für deren Langzeitarchivierung verantwortlich sein sollte. Anschließend werden Zwischenergebnisse der Anforderungsanalyse vorgestellt, die auf Interviews mit Promovierenden, Post-Docs und Gutachtern aus unterschiedlichen Wissenschaftsbereichen an der Humboldt-Universität zu Berlin basieren.

Schlagwörter. Elektronische Dissertationen, Forschungsdaten, Langzeitarchivierung, Forschungsdatenpublikation, Bibliotheken

Einleitung

Entsprechend der sehr ausdifferenzierten fachspezifischen Datenerhebung bzw. -verarbeitung erfolgt die Veröffentlichung und Archivierung von Forschungsdaten in der Regel in Community- oder fachspezifischen Repositorien und nicht Community-übergreifend.[1] Es gibt aber auch Ausnahmen. So ergibt sich aus den gesetzlichen Vorgaben der Deutschen Nationalbibliothek (DNB) der Auftrag, Forschungsdaten, die Teil einer Publikation sind, auf nationaler Ebene zu sammeln. Eine Lösung für diesen Auftrag erarbeitet die DNB derzeit zusammen mit der Humboldt-Universität zu Berlin (HU) im DFG-Projekt „Elektronische Dissertationen Plus" (eDissPlus), das sich auf einen Teilbereich dieses Komplexes konzentriert, nämlich auf Forschungsdaten, die Promovierende im Rahmen ihres Dissertationsprojekts generieren und veröffentlichen. Damit erfasst eDissPlus die ganze Bandbreite an Disziplinen und Forschungsprojekten, die an einer promotionsberechtigten Hochschule existieren, sodass eine Adaption der Projektergebnisse auf andere

1 So verzeichnet das Verzeichnis re3data.org (Stand Februar 2017) hauptsächlich disziplinspezifische Ein richtungen (1560). Die anderen recherchierbaren Angaben sind institutionell (465) und „other" (189). An gaben zu Community- oder fachübergreifenden Repositorien liegen nicht vor. Verfügbar unter: http://www.re3data.org/search.

Community-übergreifende Szenarien möglich ist. Unter anderem sollen in diesem Projekt ein Ingest-Prozess für die Publikation von elektronischen Dissertationen mit Forschungsdaten entwickelt und der Pflichtablieferungsworkflow der DNB in Hinblick auf die neuen Sammelobjekte erweitert werden. Der vorliegende Beitrag präsentiert Erkenntnisse, die in den ersten Monaten des Projekts gewonnen wurden und für die Forschungsdaten-Community von Interesse sind. Es handelt sich erstens um Überlegungen zu den Verantwortungsstrukturen, die aus Sicht der DNB bei der Langzeitarchivierung publikationsbezogener Forschungsdaten zum Einsatz kommen sollten und zweitens um die Ergebnisse von Interviews, die für eine Anforderungsanalyse auf Seiten der HU vor allem mit Promovierenden geführt wurden.

Verantwortungsstrukturen

Die DNB und ihr Sammelauftrag

Der Sammelauftrag der Deutschen Nationalbibliothek ist in der Verordnung über die Pflichtablieferung von Medienwerken an die Deutsche Nationalbibliothek (Pflichtablieferungsverordnung - PflAV) definiert. Diese enthält auch klare Aussagen über Forschungsdaten. So definiert § 9 (Weitere Einschränkungen der Ablieferungspflicht für Netzpublikationen), dass selbständig veröffentlichte Primär-, Forschungs- und Rohdaten nicht abzuliefern sind. Das impliziert eine Ablieferungspflicht für unselbständig publizierte Forschungsdaten, die laut § 7 abzuliefern sind, da „die Ablieferungspflicht auch alle Elemente, Software und Werkzeuge [umfasst], die in physischer oder in elektronischer Form erkennbar zu den ablieferungspflichtigen Netzpublikationen gehören, auch wenn sie für sich allein nicht der Ablieferungspflicht unterliegen“ (BMJV 2017).

Damit wird deutlich, dass die Zuständigkeit der DNB sich nicht auf alle Publikationstypen von Forschungsdaten erstreckt. Vielmehr ist ihr Sammelauftrag auf Daten beschränkt, die „erkennbar“ zu einer Netzpublikation gehören und damit unselbständig publiziert wurden. Aus diesen Vorgaben wurden in der ersten Projektphase Kriterien für eine noch zu schaffende Forschungsdaten-Policy der DNB abgeleitet. Es galt zu konkretisieren, wann Forschungsdaten erkennbar zu einer Dissertation gehören und in welchen Fällen die DNB oder andere Institutionen für deren Langzeitarchivierung verantwortlich sind.

Die DNB konnte sich nicht bereits bestehende Verantwortungsstrukturen in andern Ländern zum Vorbild nehmen, da vergleichbare Projekte offenbar noch nicht durchgeführt wurden.[2] Eine Ausnahme bilden Projekte, die sich mit der Verknüpfung von wissenschaftlicher Publikation, Daten und Autoreninformationen mittels persistenter Identifizierung (PID) beschäftigen. Ein Beispiel hierfür ist THOR (Fenner et al. 2015). In Baden-Württemberg gibt es zudem das Projekt bwDataDiss, das allerdings andere Schwerpunkte verfolgt (ALWR 2017).

2 Einen Überblick über die Aktivitäten europäischer Länder gibt die Publikation der Europäischen Kommission: „Access to and Preservation of Scientific Information in Europe“ (Tarazona Rua et al. 2015). Sie verdeutlicht, dass je nach Nation verschiedene Akteure für das Sammeln und den Zugang zu wissenschaft lichen Publikationen zuständig sind, was den Vergleich mit den deutschen Gegebenheiten sehr erschwert.

Welche Forschungsdaten gehören zu einer Dissertation?

Die „erkennbare Zugehörigkeit“ kann durch den Integrationsgrad der Daten in die Dissertation bestimmt werden. Ein Blick auf die Data Publication Pyramid, die im Rahmen des EU-Projektes Opportunities für Data Exchange (ODE) unter Teilnahme der DNB erarbeitet wurde (DNB 2013), visualisiert die möglichen Integrationsgrade (Reilly et al. 2011, 5). Die Integration von Daten und Publikation nimmt demnach von oben nach unten ab, bis im unteren Teil der Pyramide reine Datenpublikationen als selbständig publizierte Daten (Typ 4) sowie unpublizierte Daten (Typ 5) stehen, bei denen kein Konnex zu einer separaten Publikation existiert und die daher nicht unter den Sammelauftrag der DNB fallen. Interessant für den Sammelauftrag der DNB sind hingegen die Typen 1 bis 3.

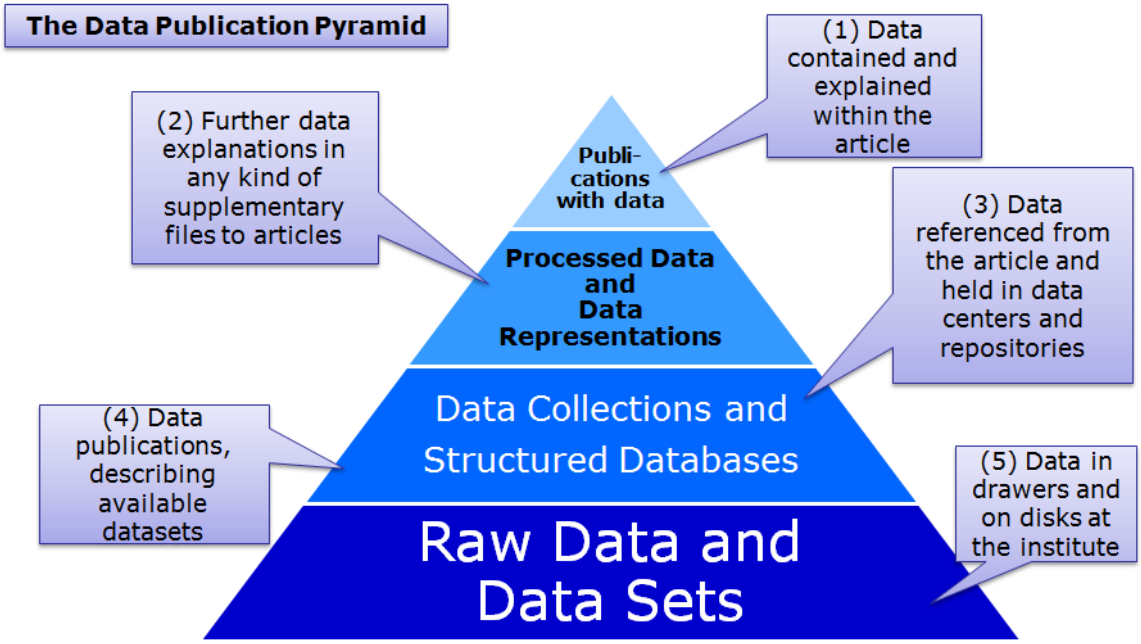


Abbildung 1. The Data Publication Pyramid (Reilly et al. 2011, 6).

Die Spitze der Pyramide bilden vollständig in eine Dissertationsschrift integrierte Daten (Typ 1). Das sind zum Beispiel Grafiken oder Tabellen, die in ein PDF eingebettet oder einer Printversion abgedruckt wurden. Forschungsdaten, die in dieser Form vorliegen, sammelte die DNB schon immer – sozusagen automatisch als Bestandteil der abgabepflichtigen Hochschulschrift. Auf diesen Typ musste eDissPlus deshalb auch nicht näher eingehen. Typ 2 beschreibt Daten, die als Supplemente einer Dissertation beigefügt sind. Varianten dieses Typs sind gedruckte Dissertationen, denen CD-ROMs oder andere Datenträger mit Forschungsdaten beiliegen. Überträgt man dieses Beispiel auf Netzpublikationen, dann manifestiert sich die erkennbare Zugehörigkeit der Forschungsdaten zur Dissertation darin, dass sie als Supplemente von Promovierenden zusammen mit der Dissertation auf den Publikationsserver der Hochschule geladen wurden. Der Integrationsgrad der Daten in die Dissertation ist immer noch hoch. Es ist anzunehmen, dass die beigegebenen Daten in enger Verbindung zur Dissertation und den darin präsentierten Forschungsergebnissen stehen. Supplemente sollten daher auch von der DNB gesammelt werden. Typ 3 illustriert die Möglichkeit, dass Forschungsdaten auf einem externen Repositorium liegen, in der Dissertati-

on aber referenziert werden (etwa durch Angabe einer PID). Man kann nun fragen, ob es sich hierbei noch um Daten handelt, die erkennbar zu der betreffenden Dissertation gehören und somit unter den Sammelauftrag der DNB fallen. Denn diese Daten gelten, sofern Sie über eine PID verfügen, zugleich auch als selbständig publizierte Daten. Deshalb werden die Typen 3 und 4 in der Data Publication Pyramid auch innerhalb einer Schicht verortet.

Umgang mit referenzierten Daten

Eine Lösung für den Umgang mit referenzierten Daten eröffnet der inhaltliche Bezug von Daten und Dissertation. Dieser manifestiert sich in dem Beitrag der Forschungsdaten zum Verständnis der Dissertation, etwa indem sie Argumentationen oder Hypothesen der Promovierenden stützen oder falsifizieren und bestimmte Sachverhalte illustrieren. Eng damit verbunden ist der Anspruch, mithilfe der vorhandenen Daten die präsentierten Forschungsergebnisse reproduzieren und damit den Forschungsprozess selbst verstehen und nachvollziehen zu können. Ein denkbares Anwendungsfeld ist die Aufdeckung wissenschaftlichen Fehlverhaltens, so wie es die Empfehlungen der DFG zur Sicherung guter wissenschaftlicher Praxis beschreiben (DFG 2013). Folgt man dieser Argumentation, dann ist auch die Sichtweise berechtigt, dass referenzierte Daten ebenso wie Supplemente in einem engen inhaltlichen Bezug zur Dissertation stehen können. Allerdings entsteht dieser Bezug nicht durch eine einfache Datenzitation im Anmerkungsapparat des Dissertationstextes. Es könnte sich dabei auch um Daten handeln, die einem vergleichbaren Forschungsdesign entstammen, aber für die Dissertation nicht ausgewertet wurden. Um die erkennbare Zugehörigkeit der externen Daten zu einer Dissertation zu signalisieren, sollten die Promovierenden die Referenzen während des Abgabeprozesses der Dissertation explizit als solche ausweisen. Das kann durch den Eintrag von PIDs (z. B. DOI, URN, handle) in ein Webformular geschehen. In einem solchen Fall können die extern archivierten Daten als erkennbarer Bestandteil der Dissertation betrachtet und von der DNB gesammelt oder doch zumindest im Metadatensatz der Dissertation durch Angabe der PID referenziert werden.

Damit wird eines deutlich: Letztlich entscheiden die Promovierenden über die Bedeutung der Daten und ihre Unverzichtbarkeit im Hinblick auf Verständnis und Nachvollziehbarkeit der in der Dissertation publizierten Forschungsergebnisse.[3] Daher ist es Aufgabe der abliefernden Hochschulbibliotheken, durch propädeutische Maßnahmen (z. B. Schulungen, Workshops, Guidelines) bei den Studierenden ein Bewusstsein für die Bedeutung von Forschungsdaten zu erzeugen. Die Entwicklung entsprechender Maßnahmen ist folglich auch Teil des Projektes eDissPlus.

Zudem muss die Frage geklärt werden, ob die DNB extern archivierte Daten auch in ihr Langzeitarchiv übernimmt oder ob ein Metadatennachweis im Katalog ausreicht, zumal Forschungsdaten bereits vor Ablieferung der Dissertationsschrift auf einem Repositorium vorliegen können.[4] Die im vergangenen Jahr erschienenen Empfehlungen des Rates für Informationsinfrastrukturen (RfII) weisen hier eindeutig die Richtung, indem sie einen verteilten Archivierungsansatz vorschlagen, wonach „Aufgaben der Langzeitarchivierung als informationsinfrastrukturelle Daueraufgaben durch ggf. miteinander vernetzte Einrichtungen im Rahmen der Nationalen For-

3 Ohnehin tragen die Datenproduzenten, in diesem Fall die Promovierenden, Verantwortung für die Daten qualität, wozu im Falle eines Dissertationsprojektes auch die Auswahl der beizufügenden Daten gehören sollte. Vgl. hierzu auch die Empfehlungen des Rates für Informationsinfrastrukturen (RfII 2016).

4 Vergleiche hierzu auch die Punkte 6 und 10 der unten präsentierten Interview-Zwischenergebnisse.

schungsdateninfrastruktur (NFDI) zu leisten und zu koordinieren“ sind (RfII 2016, 47). Externe Daten sollten somit grundsätzlich nicht von der DNB übernommen, sondern in den bereits genutzten Repositorien verbleiben und lediglich im Metadatensatz der betreffenden Dissertation referenziert werden. Allerdings ist zu beachten, dass Langzeitarchivierung oft missverständlich mit den Empfehlungen der Deutschen Forschungsgemeinschaft (DFG) gleichgesetzt wird, wonach Primärdaten zehn Jahre aufbewahrt werden sollen (DFG 2013). Langzeitarchivierung meint jedoch wesentlich längere Zeiträume und kennt auch kein Ablaufdatum (Neuroth et al. 2009). Es wird somit seitens der DNB zu prüfen sein, ob ein Repositorium im Sinne der Langzeitarchivierung vertrauenswürdig ist und über eine entsprechende Zertifizierung verfügt (z. B. Data Seal of Approval; nestor-Siegel). Fehlt eine solche Zertifizierung, müsste die DNB entscheiden, ob und wie sie die Daten übernimmt und archiviert. Da dies jedoch aufwendige Prüfroutinen und Schnittstellenentwicklungen voraussetzt, ist eine Lösung dieser Aufgabe im Rahmen von eDissPlus nicht wahrscheinlich. Die Übernahme und Langzeitarchivierung von Supplementen sollte aufgrund des größeren Integrationsgrades von Dissertation und Daten Vorrang haben.

Anforderungsanalyse

Erhebungsmethoden und Analyseverfahren

Während die DNB vor allem die Perspektive einer Archivbibliothek einnimmt, verfolgt die Universitätsbibliothek der HU eine stärker forschungszentrierte Sichtweise auf den Umgang mit Forschungsdaten. Um dementsprechend die Anforderungen der Promovierenden aus verschiedenen Fachdisziplinen zu ermitteln, werden derzeit zusätzlich zur Auswertung bestehender quantitativer Studien (Simukovic et al. 2013; KIT 2015; Tenopir et al. 2017) eine Reihe von qualitativen Leitfadeninterviews mit Promovierenden, Post-Docs sowie Gutachtenden von Dissertationen an der Humboldt-Universität zu Berlin durchgeführt. Interessant sind dabei – neben dem allgemeinen Umgang mit dissertationsbegleitenden Daten während des Forschungsprozesses – vor allem die Möglichkeiten ihrer Veröffentlichung. In den Interviews werden die für das jeweilige Dissertationsprojekt relevanten Forschungsdaten unter Berücksichtigung der möglichen Datentypen und Bearbeitungsphasen erhoben. Weiterhin wird nach spezifischen Rahmenbedingungen, der persönlichen Motivation, den Besonderheiten der fachkulturellen Praxis, der Evaluation bestehender Infrastrukturangebote sowie nach konkreten Desideraten gefragt.

Für die Befragungen wurden potentielle Interviewpartner über diverse Kommunikationskanäle wie die Humboldt Graduate School, Fachreferenten und Bibliotheksbeauftragte kontaktiert. Trotz großer Bemühungen war das Echo erstaunlich gering, weshalb die Interviewphase noch nicht abgeschlossen werden konnte. Zum Stand dieses Beitrags lagen 16 transkribierte Interviews vor aus den Fachbereichen Bibliotheks- und Informationswissenschaft, Erziehungswissenschaft, Europäische Ethnologie, Geografie, Geschichtswissenschaft, Kultur- und Sozialanthropologie, Kulturwissenschaft, Literaturwissenschaft, Medizin, Physik, Rechtswissenschaft sowie Soziologie. Zudem ist die Repräsentativität insofern eingeschränkt, als dass sich bislang hauptsächlich Personen mit einem besonderen persönlichen Interesse am Thema zum Gespräch bereit erklärten.

Für die Datenauswertung wurden aus den Interviewprotokollen jeweils zwischen 50 bis 80 relevante Einzelaussagen extrahiert, so dass bislang etwa 1100 Statements als Datenbasis zur Ver-

fügung stehen. Anhand einer qualitativen Inhaltsanalyse mit der Vergabe von Keywords und unter Einsatz des Erschließungs- und Recherchewerkzeugs *Statement Finder*[5] werden wiederkehrende Einstellungsmuster, Prozesse, Anforderungen und Desiderate geclustert sowie fachspezifische Besonderheiten identifiziert. Auf dieser Grundlage lassen sich typisierte Nutzungsszenarien und Anforderungsprofile modellieren, welche die unterschiedlichen Zielgruppen für die Archivierungs- bzw. Publikationsangebote der HU abbilden sollen.

Zwischenergebnisse

Zu den zentralen Erkenntnissen der Interviewphase gehört, dass gemessen am Ideal einer offenen Wissenschaft der transparente und nachhaltige Umgang mit Forschungsdaten erhebliche Herausforderungen zu bewältigen hat. Diese umfassen individuelle Aspekte (z.B. Motivation, Zeitressourcen, IT-Kompetenz), wissenschaftsstrukturelle (z.B. Konventionen, Promotionsordnungen, Anreizsysteme), infrastrukturelle (z.B. IT-Dienste, Publikationsworkflows) sowie rechtliche Herausforderungen (z.B. Datenschutz, Lizenzierung, Urheberrecht). Als erstes Zwischenergebnis lassen sich folgende Punkte festhalten:

1. *Thema Forschungsdaten randständig in Wissenschaftspraxis*
   Eher unerwartet und daher sehr lehrreich ist die Einsicht, dass das Thema der dissertationsbegleitenden Forschungsdatenpublikation aus Sicht der Forschenden derzeit sehr randständig erscheint. Bereichsübergreifend gibt es für solche Veröffentlichungsformen keine gängige Praxis, sondern allenfalls persönlich motivierte Einzelfälle. Impulse durch die Begutachtenden gibt es ebenso selten wie Best-Practice-Beispiele. Rezeption oder Nachnutzung der Forschungsdaten sowie eine entsprechende Kreditierung durch die Fachgemeinschaften werden kaum erwartet. In Fachdiskursen wird das Thema Forschungsdatenpublikation vor allem im Zusammenhang mit Vorgaben von Förderinstitutionen (z.B. DFG, BMBF, ERC) wahrgenommen.
2. *Kaum erkennbare Rahmenbedingungen für Forschungsdatenmanagement*
   Abgesehen von den externen Vorgaben der Förderinstitutionen sind selten fachspezifische Rahmenbedingungen oder gar konkrete Handlungsempfehlungen für den Umgang mit Forschungsdaten feststellbar. Insbesondere findet das Thema kaum Berücksichtigung in geltenden Promotionsordnungen. Unabhängig von der Frage einer Forschungsdatenpublikation wird zwar deutlich, dass auf die Einhaltung der 10-jährigen Aufbewahrungfrist von Forschungsdaten im Sinne der guten wissenschaftlichen Praxis Wert gelegt wird. Allerdings werden hierbei meist ad-hoc-Lösungen wie das Vorhalten der Daten auf privaten Speichermedien verfolgt. Promovierenden sind die spezifischen Ansprüche an das Forschungsdatenmanagement wie beispielsweise die Erstellung eines Datenmanagementplanes kaum bekannt.
3. *Speicherbedarfe und Datenformate überschaubar*
   In Übereinstimmung mit einer vorherigen Studie (Simukovic et al. 2013) erreicht das Datenvolumen eines Promotionsprojektes selten mehr als einen zweistelligen Giga-

5 Der Statement Finder wurde im Rahmen des DFG-Projektes „Future Publications in den Humanities“ (FuPusH) entwickelt und als Open-Source-Anwendung zur Nachnutzung bereitgestellt. Verfügbar unter: https://www2.hu-berlin.de/fupush/statement-finder/#/statements.

bytebereich. Bei den Datenformaten besteht zwar durchaus eine gewisse Heterogenität, aber in der Regel werden gängige fachspezifische Standardformate verwendet.

4. *Teilweise hohe Anzahl an einzelnen Forschungsdaten (z.B. Quellenmaterialien)*
   Wie granular Forschungsdaten als Publikationsobjekte anfallen steht in Abhängigkeit zum jeweiligen Forschungsprojekt. Es kann sich um eine einzelne Tabelle handeln oder um 8000 Abbildungen. Hier zeichnet sich Beratungsbedarf ab beispielsweise hinsichtlich des Bündelns zu bestimmten logischen Publikationseinheiten. Werden hohe Auffindbarkeit und separate Nachnutzung angestrebt, besteht Interesse an einer möglichst granularen Aufschlüsselung der Daten, einschließlich einer jeweiligen PID.
5. *Forschungsdatenmanagement höher priorisiert als Forschungsdatenpublikation*
   Während die Veröffentlichung von Forschungsdaten aus unterschiedlichen Gründen weitgehend nachgeordnet betrachtet wird, misst man dem Forschungsdatenmanagement durchgehend hohe Bedeutung zu. Dazu zählen aus Sicht der Promovierenden vor allem die Speicherung und Versionierung der Daten, die Verwaltung der Zugriffsrechte sowie die Dokumentation. Die Bedeutung von Datenmanagementplänen wird anerkannt, auch wenn die meisten Befragten keine entsprechenden Erfahrungen vorweisen. Großes Interesse besteht hinsichtlich der zum Teil unbekannten universitären Infrastruktur- und Beratungsangebote zum Forschungsdatenmanagement. Für Promovierende erweist sich die Perspektive einer Langzeitarchivierung bzw. -verfügbarkeit ihrer Forschungsdaten durch die DNB teilweise als Motivation, eine Forschungsdatenpublikation in Erwägung zu ziehen.
6. *Forschungsdatenmanagement in datengetriebenen Disziplinen meist intern*
   Datenintensive Fachbereiche wie Naturwissenschaften, Medizin, Geografie oder Digital Humanities bieten oft institutsinterne IT-Infrastrukturen und Lösungen zum Forschungsdatenmanagement, die als funktionierend bewertet und im Vergleich zu Angeboten der Universitätsbibliothek oder dem Rechenzentrum bevorzugt werden.
7. *Hoher Beratungsbedarf in Bereichen qualitativer empirischer Sozialforschung*
   Auch wenn die Befragungen selbst nur einen sehr selektiven Einblick bieten, zeichnet sich ab, dass in Fachbereichen mit einem hohen Anteil an qualitativer Sozialforschung wie in der Kultur- und Sozialanthropologie, der Europäischen Ethnologie oder der Sozialgeschichte ein besonderer Bedarf an Beratungs- und Dienstleistungsangeboten besteht. Dies wurde unter anderem auch in einem gemeinsamen Workshop mit dem Fachinformationsdienst (FID) „Kultur- und Sozialanthropologie“ an der Universitätsbibliothek der HU Berlin herausgearbeitet. Gründe dafür werden vor allem gesehen in der Heterogenität der in diesen Fächern auftretenden Forschungsdaten (z.B. Feldtagebücher, Interviews, Filmaufnahmen, Digitalisate von Artefakten und Archivalien), der Sensibilität der oft personenbezogenen und nur schwer anonymisierbaren Daten sowie einer im Vergleich zu klassischen MINT-Fächern weniger ausgeprägten IT-Affinität. Zum einen besteht Unsicherheit in der Frage, inwieweit ethnologische bzw. historische Quellenmaterialien überhaupt als Forschungsdaten angesehen werden können; zum anderen erscheint die Relevanz solcher Quellen und daher auch die Motivation einer Veröffentlichung besonders hoch.
8. *Forschungsdatenpublikation stark abhängig von individueller Motivation*
   Das Interesse am Thema Forschungsdaten scheint vor allem persönlich motiviert zu sein. Generell lassen sich für eine Forschungsdatenpublikation drei Motivationstypen

unterscheiden: a) eine pragmatische, bei der Forschungsdaten aus situativer Notwendigkeit und eventuell zur Transparenzsicherung der Dissertationsschrift begleitend publiziert werden; b) eine reputationale, bei der die Forschungsdatenpublikation in der Erwartung zusätzlicher wissenschaftlicher Kreditierung veröffentlicht wird und schließlich c) eine ethische, die Zugänglichkeit und Nachnutzbarkeit von Forschungsdaten als wichtigen Baustein offener Wissenschaft ansieht. Empirisch scheint der dritte Motivationstyp bei den Promovierenden die derzeit dominante Form zu sein.

9. *Forschungsdatenpublikation oft unabhängig von Dissertationsschrift*
   Entgegen der Ausgangsannahme, dass die Dissertationsschrift und die zugehörigen Forschungsdaten gleichzeitig publiziert werden, zeigt sich, dass besonders dann, wenn die Datenpublikation nicht von vornherein im Publikationsplan vorgesehen ist, der Abschluss der Promotion zunächst gegenüber einer Forschungsdatenpublikation priorisiert wird. Daher wird oft eine möglichst frühzeitige Veröffentlichung der Dissertationsschrift angestrebt. Aufbereitung und Publikation der Forschungsdaten werden eher für den Zeitraum nach der Titelzuerkennung angestrebt.
10. *Migration der Forschungsdaten von Fremdrepositorien ggf. erwünscht*
    Es gibt weiterhin Fälle, bei denen Forschungsdaten bereits vor Abschluss der Arbeit auf externen Plattformen veröffentlicht wurden. Entweder sollten dadurch Nachnutzungsszenarien frühzeitig realisiert werden oder die Forschungsdaten sind zugleich in übergreifende Zusammenhänge wie kooperative Forschungsprojekte eingebunden. Hier besteht das Interesse, diese bereits extern veröffentlichten Daten möglichst in direkter Verbindung zur Dissertationsschrift nochmals zu veröffentlichen. Es wird auch betont, dass der natürliche Erscheinungsort einer Forschungsdatenpublikation zur Promotion die jeweilige Hochschule ist.
11. *Kontrolle über Weitergabe eigener Forschungsdaten statt Open Research Data*
    Viele Promovierende möchten selbst entscheiden, wem und unter welchen Bedingungen sie ihre Forschungsdaten zur Einsicht und Nachnutzung überlassen. Ein Nachweis von Forschungsdaten durch Metadaten in Suchmaschinen bzw. Discovery Systemen gilt durchaus als wichtig, um die eigene Sichtbarkeit zu erhöhen. Die Freigabe der Daten selbst soll dagegen nur nach Rücksprache erfolgen.
12. *Dokumentation kaum standardisiert und sehr fachspezifisch*
    Für die Dokumentation von Forschungsdaten existieren unterschiedliche disziplinäre Anforderungen, aber selten Standards bzw. Vorlagen. Hier existiert eindeutig ein Desiderat. Die Dokumentation von Forschungsdaten wird oft als erheblicher zusätzlicher Aufwand bewertet, der viele Promovierende von einer Forschungsdatenpublikation absehen lässt.

Zusammenfassung

Der erste Teil des Beitrages reflektierte mögliche Verantwortungsstrukturen für die Langzeitarchivierung dissertationsbezogener Forschungsdaten. Aus dem Gesagten ergeben sich Eckpunkte für eine Forschungsdaten-Policy der DNB, die auch auf andere Publikationsformen erweiterbar ist. Hierbei sind insbesondere die unterschiedlichen Integrationsgrade von Forschungsdaten in

eine Dissertation (Supplement oder Referenz) und die Bedeutung des inhaltlichen Bezugs der Daten zur Dissertation als Selektionskriterium zu beachten.

Im zweiten Teil wurden Zwischenergebnisse der Anforderungsanalyse für die Archivierung und Veröffentlichung von dissertationsbegleitenden Forschungsdaten vorgestellt. Als vorläufiges Fazit lässt sich festhalten, dass bisher eher punktuelle Bedarfe bestehen. Die relative Randständigkeit des Themas in der Wissenschaftspraxis lässt vermuten, dass bestehende Initiativen eher wissenschafts- und förderpolitisch und damit extrinsisch zu den Fach-Communities motiviert sind. Insgesamt sollte also das zu entwickelnde Dienstleistungsspektrum neben den unabdingbaren technischen Rahmenbedingungen vor allem beratend ausgerichtet sein und die Möglichkeit enthalten, bei Bedarf auch eine direkte Begleitung auf Einzelfallebene anzubieten.

Literaturangaben

Arbeitskreis der Leiter wissenschaftlicher Rechenzentren in Baden-Württemberg (ALWR). 2017. „bwDataDiss - bwData für Dissertationen“. Online verfügbar unter https://www.alwr-bw.de/kooperationen/bwdatadiss/. Zuletzt geprüft am 13.08.2017.

Bundesministerium für Justiz und für Verbraucherschutz (BMJV). 2017. „Verordnung über die Pflichtablieferung von Medienwerken an die Deutsche Nationalbibliothek“. Online verfügbar unter http://www.gesetze-im-internet.de/pflav/index.html. Zuletzt geprüft am 13.08.2017.

Deutsche Forschungsgemeinschaft (DFG). 2013. *Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission "Selbstkontrolle in der Wissenschaft"; Denkschrift.* Ergänzte Auflage. Weinheim: Wiley-VHC. doi:10.1002/9783527679188.oth1.

Deutsche Nationalbibliothek (DNB). 2013. “ODE - Opportunities for Data Exchange”. Last modified February 08, 2013. Online verfügbar unter http://www.dnb.de/DE/Wir/Projekte/Archiv/ode.html. Zuletzt geprüft am 13.08.2017.

Fenner, Martin et al. 2015. D2.1: *Artefact, Contributor, and Organisation Relationship Data Schema.* doi:10.5281/zenodo.30799.

Karlsruhe Institute of Technology (KIT). 2015.”Öffentlicher Abschlussbericht von bwFDM-Communities“. Online verfügbar unter http://bwfdm.scc.kit.edu/bwFDM-Communities.php. Zuletzt geprüft am 13.08.2017.

Neuroth et al. 2009. *nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung.* Version 2.0. Göttingen: Verlag Werner Hülsbusch,

Reilly, Susan, et al. 2011. *Opportunities of Data Exchange: Report on Integration of Data and Publications*. Online verfügbar unter hdl:10013/epic.40198.d001. Zuletzt geprüft am 13.08.2017.

RfII – Rat für Informationsinfrastrukturen. 2016. *Leistung aus Vielfalt: Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland.* Göttingen. Online verfügbar unter urn:nbn:de:101:1-201606229098. Zuletzt geprüft am 13.08.2017.

Simukovic, Elena, Maxi Kindling, Peter Schirmbacher. 2013. „Umfrage zum Umgang mit digitalen Forschungsdaten an der Humboldt-Universität zu Berlin“. Online verfügbar unter http://edoc.hu-berlin.de/oa/reports/reFIYMgduNiVE/PDF/22YavRAszVauc.pdf. Zuletzt geprüft am 13.08.2017.

Tarazona Rua, Maria et al., eds., 2015. *Access to and Preservation of Scientific Information in Europe*. Luxemburg: Publications Office of the European Union. doi:10.2777/975917.

Tenopir et al. 2017. Research Data Services in European Academic Research Libraries. *LIBER Quarterly* 27(1), 23–44. http://doi.org/10.18352/lq.10180.

One-stop publishing and archiving: Forschungsdaten für Promotionsvorhaben über Repositorien publizieren und archivieren: Eine landesweite Initiative im Rahmen des Projekts bwDataDiss am Beispiel des Karlsruher Instituts für Technologie (KIT)

Tobias Kurze[1], Regine Tobias[2], Matthias Bonn[3]

1,2 Bibliothek, KIT
2 Steinbuch Centre for Computing, KIT

Abstract. Nowadays research relies more and more on data to achieve progress in various scientific domains. To understand and to be able to reproduce results, it is essential that the underlying research data is available to scientists - even after a relatively long time.
bwDataDiss is an effort to provide infrastructure and services for a very specific group of researchers – namely PhD students – to enable them to store and archive their research data and also to make it available to other researchers.

Schlagwörter. Dissertation, Forschungsdaten, Langzeitarchivierung, OpenAccess

Einführung – Hintergründe für das Projekt bwDataDiss

bwDataDiss ist als dreijähriges Projekt, finanziert durch das MWK Baden-Württemberg gestartet. Das Projekt verbindet die Erfahrungen der Bibliothekswelt im Umgang mit Nutzern mit der Expertise der Rechenzentren im Bereich des Aufbaus und Umgang mit großen Speicherinfrastrukturen. Daher sind im Projekt jeweils die Bibliotheken und die Rechenzentren der Universität Freiburg und des Karlsruher Institut für Technologie beteiligt. Die Hauptmotivation für das Projekt bestand in der Errichtung einer Infrastruktur für die Veröffentlichung und den Erhalt von Forschungsdaten sowie die Bereitstellung dieser Infrastruktur in Baden-Württemberg. Durch bwDataDiss soll es auch kleineren Einrichtungen ohne große lokale Rechenzentren vor Ort möglich sein, auf einfachem und schnellem Wege die Repositorien der Bibliotheken um neue Forschungsdatenservices zu erweitern.

Und die Erweiterung der Services für dieses Anwendungsfeld ist auch dringend nötig, denn die Diskussion um die Veröffentlichung und den Erhalt von Forschungsdaten ist derzeit in der Wissenschaftswelt ein großes Thema[1]. In der Denkschrift der DFG zur „Sicherung guter Wissenschaftlicher Praxis“ wird darauf hingewiesen, dass „Primärdaten als Grundlagen für Veröffentli-

1 Siehe zum Beispiel: Rat für Informationsinfrastrukturen: Leistung aus Vielfalt. Empfehlungen zu Struktu ren, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland, Göttingen, 2016.

chungen (…) auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, zehn Jahre lang aufbewahrt werden (sollen).“[2]

Inzwischen ist in der Wissenschaftsgemeinschaft die Bestätigung, dass zugrundeliegende Forschungsdaten für die Überprüfbarkeit von Forschungsergebnissen häufig unverzichtbar sind, fast ein Allgemeinplatz. An vielen Orten formieren sich daher entsprechende Policies an den Hochschulen, die erste Schritte und Herausforderungen im Umgang mit der Thematik regeln.[3] Gleichzeitig wächst der Bedarf an entsprechender Infrastruktur, diese Forschungsdaten langfristig zu archivieren und den Zugang zu ihnen zu ermöglichen. Trotz vielen Diskussionen und Initiativen in den letzten Jahren besteht aber derzeit noch eine große Lücke bei der tatsächlich bereitgestellten Infrastruktur der Bibliotheken. Forschungsdaten sind, wenn überhaupt, in vielen Fällen in disziplinären Fachrepositorien abgelegt, denen es vielfach an nachhaltigen Betriebskonzepten mangelt und die in punkto Verbindlichkeit und Standardisierung häufig erst noch Neuland betreten.[4]

Angesichts der großen Herausforderungen des neuen Serviceumfelds wollte das Wissenschaftsministerium in Baden-Württemberg für bwDataDiss einen konkreten Rahmen spannen, so dass die Umsetzungserfolge für Infrastrukturanbieter leichter zu erreichen sind. Daher knüpfte man an bestehende Workflows und Vorarbeiten an. Ganz konkret handelt es sich hier um Forschungsservices für Nachwuchswissenschaftler: Denn zum einen entstehen im Rahmen von Forschungsprojekten und im Speziellen bei Doktorarbeiten häufig Forschungsdaten und zum anderen spielen Bibliotheken traditionell eine Schlüsselrolle im Dissertationsprozess: Die Pflichtabgabe zur Erlangung des Doktortitels erfolgt an allen Hochschulstandorten anhand der Veröffentlichung über die zugehörige Bibliothek. Die zugrundeliegenden Workflows sind also bereits existent und wurden von der analogen, auf gedruckten Exemplaren basierenden Abgabe in den letzten Jahren annähernd vollständig auf digitale Veröffentlichungsprozesse transformiert.

Allerdings unterscheiden sich eben diese Prozesse von Bibliothek zu Bibliothek im Detail und können auch relativ komplex sein. Um diese Komplexität und Diversität ein Stück weit abzubilden, bringen sowohl die Universitätsbibliothek Freiburg, als auch die KIT-Bibliothek ihr Wissen um promotionsbezogene Publikationsworkflows in bwDataDiss ein.

Und noch ein weiterer Anknüpfungspunkt war für das Projekt relevant: Bibliotheken haben üblicherweise weder die Möglichkeit, große Datenmengen zu speichern, noch Erfahrungen darin, diese Daten auf Langzeitverfügbarkeit zu analysieren. Daher fließt in bwDataDiss die Expertise von zwei großen Rechenzentren mit ein: Das SCC am KIT stellt die IT Infrastruktur und die Systeme zur Verfügung, die es ermöglichen, Forschungsdaten zu speichern und zu archivieren. Das Rechenzentrum der Universität Freiburg liefert Werkzeuge, um Daten auf Archivierbarkeit und Langzeitverfügbarkeit zu untersuchen – in bwDataDiss auch als Charakterisierung bezeichnet.

2 DFG – Deutsche Forschungsgemeinschaft: Sicherung guter wissenschaftlicher Praxis. Denkschrift, Empfehlungen der Kommission zur Selbstkontrolle in der Wissenschaft, Bonn, 2013: http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf, S. 21 f. : zuletzt geprüft am 24.2.2017.

3 Zum Beispiel die Forschungsdaten-Policy am KIT vom 17.10.2016. http://www.rdm.kit.edu/downloads/FDM-Policy_final.pdf, zuletzt geprüft am 24.2.2017.

4 Eine gute Übersicht über Forschungsdatenrepositorien gibt re3data: http://www.re3data.org/.

Aktueller Stand im Projekt

Im letzten Jahr der Projektförderung präsentiert sich nun bwDataDiss als ein Dienst der KIT-Bibliothek, der Hochschulen des Landes Baden-Württemberg beim Aufbau einer lokalen Infrastruktur für die Langzeitarchivierung und Bereitstellung von Forschungsdaten von Promovierenden unterstützt und über ein landesweites Portal präsentiert. Als Voraussetzung für die Nutzung der Services von bwDataDiss muss zwischen der jeweiligen Bibliothek (bzw. Universität) und dem KIT ein Vertrag abgeschlossen werden. Promovierende, die bei einer teilnehmenden Bibliothek ihre Dissertation abgeben, können dann zusätzlich Forschungsdaten über bwDataDiss archivieren und publizieren. BwDataDiss unterstützt alle Disziplinen und Datentypen. An die Abgabe der Forschungsdaten wird lediglich die Bedingung geknüpft, dass diese einen finalen, für die Nachnutzung aufbereiteten Charakter innehaben. Der genaue Umfang des Services der Bibliothek und die endgültige Auswahl der Forschungsdaten liegt im Ermessensspielraum der zuständigen Bibliothek. Der Dienst ist eng mit den jeweiligen Repositorien vor Ort verbunden und kann auch auf weitere Publikationstypen ausgeweitet werden, die unabhängig von der Dissertationsabgabe fungieren. Im Zentrum steht, dass der Dienst vollständig in lokale Workflows der Bibliothek integriert werden kann.

Es haben sich drei Modelle herauskristallisiert, die die Nutzung des Dienstes sowohl für große Bibliotheken mit einer leistungsfähigen IT-Infrastruktur als auch für kleinere Bibliotheken attraktiv macht. Die Modelle unterscheiden sich in erster Linie anhand der Integrationstiefe der bwDataDiss-Infrastruktur in die vorhandenen Workflows und sind in Abbildung 1 dargestellt.

Abbildung 1. Integrationsmodelle von bwDataDiss

Der Projektpartner Universitätsbibliothek Freiburg folgt Modell 1, das es ermöglicht, die Workflows vollständig in FreiDok *plus*[5], dem institutionellen Repository der Universitätsbibliothek Freiburg, abzubilden. In den letzten Jahren wurde es zu einem Forschungsdateninformationssystem weiterentwickelt, um die komplette Forschungslandschaft der Universität abdecken zu können. Vor diesem Hintergrund ermöglicht FreiDok *plus* auch die Veröffentlichung und Archivierung von Forschungsdaten in einem Guss und stellt entsprechende Workflows bereit. Der Integrationsaufwand war dementsprechend hoch.

Einen anderen Ansatz verfolgte die KIT-Bibliothek mit dem zentralen Repository KITopen[6], die die Mehrwerte von bwDataDiss weniger nachprogrammieren, als direkt mit dem Repository verbinden möchte. Der Fokus der IT-Anbindung liegt darauf, zwar eine möglichst einheitliche

5 https://freidok.uni-freiburg.de/
6 https://www.bibliothek.kit.edu/cms/kitopen.php

Nutzerkommunikation anzustreben, aber bei den konkreten Anwendungen auf die Features von bwDataDiss zu verweisen. Ganz ohne Nutzerbrüche kommt der Forschungsdatenworkflow so nicht aus, aber der Integrationsaufwand wird um beträchtliche Komponenten verringert.

Das Projekt bwDataDiss konnte im Verlauf des Projekts noch ein weiteres, drittes Modell entwickeln, das dem Wunsch des Ministeriums nach Nachnutzbarkeit voll entspricht: Durch die komplexe Portalentwicklung von bwDataDiss kann eine Bibliothek auch auf relativ schnellem Wege zur Einführung dieses neuen Forschungsdatendienstes kommen, indem das Repository und bwDataDiss IT-technisch völlig getrennt voneinander betrieben werden. Dieser Ansatz kommt in dieser Form auch kleineren Bibliotheken zu Gute (Modell 3).

Im Folgenden werden die technischen Grundlagen von bwDataDiss erläutert. Die Ausführungen werden schwerpunktmäßig anhand des Modells 2 erläutert, gehen aber in den Details auch auf die Spezifikationen für Modell 1 und Modell 3 ein bzw. sind auch für diese gültig.

Konzepte und Workflows

Prinzipien

Beim Entwurf von bwDataDiss wurden folgende Prinzipien berücksichtigt:

- Möglichst einfach nutzbar für die Hauptkunden, nämlich Promovenden bzw. Forschende im Allgemeinen
- Datenintegrität jederzeit sicherstellen
- Flexible Möglichkeiten der Integration in Bibliothekssysteme

Aufgabenteilung

Zwischen der Bibliothek und bwDataDiss gibt es eine strikte Aufgabentrennung:
Bibliotheken:

- Ansprechpartner für Forschende und Promovierende
- Erfassung und Kontrolle für Metadaten

bwDataDiss:

- Archivierung von Daten
- Bereitstellung von archivierten Daten zur Nachnutzung durch die Wissenschaftsgemeinschaft
- Charakterisierung der Forschungsdaten und Bereitstellung derer Ergebnisse

Da Bibliotheken über ihre Repositorien eine zentrale Rolle im institutionellen Publikationsprozess spielen, ist es konsequent, die Abgabe von Forschungsdaten – die als Teil der Dissertation oder als Teil von weiteren Publikationen entstanden sein können – in eben diesen Prozess zu integrieren. Der Workflow besteht im Groben aus den folgenden Schritten (Reihenfolge vernachlässigbar):

1. Der Forscher überträgt seine Publikation (Bibliothekswebseite)

2. Der Forscher stellt Metadaten für die Publikation bereit (Bibliothekswebseite)
3. Die Bibliothek prüft die Publikation und die bereitgestellten Metadaten und kontaktiert ggf. den Forschenden

Bei Forschungsdaten sind mindestens zwei zusätzliche Schritte nötig:

- Übertragung der Forschungsdaten (zu bwDataDiss oder zur Bibliothek, dies hängt vom konkreten Integrationsszenario ab)
- Zu den Forschungsdaten gehörende Metadaten bereitstellen (zu bwDataDiss oder zur Bibliothek)

Möglichkeiten der Integration einzelner Komponenten

Upload

bwDataDiss unterstützt mehrere Wege der Integration in die Bibliothekssysteme, um Forschungsdaten vom Promovenden zum Archiv zu transferieren. Entweder werden die Daten direkt vom Promovenden zu bwDataDiss (und dann weiter ins Archiv) übertragen, oder die Daten werden temporär auf Servern der Bibliothek zwischengespeichert und später von dort zu bwDataDiss übertragen (Modell 1). In jedem Fall aber prüft die Bibliothek die bereitgestellten Metadaten.

Details des direkten Datentransfers: Es gibt zwei Möglichkeiten, den direkten Datentransfer vom Promovenden zu bwDataDiss zu organisieren, wobei die Bibliothek entscheidet, welche Lösung umgesetzt wird.

Für eine einheitlichere Sicht auf die Bibliotheksseite (Modell2) kann die Uploadkomponente von bwDataDiss in die Bibliothekswebseite integriert werden. Obwohl die Komponente auf der Bibliotheksseite integriert ist, überträgt diese die Daten direkt an bwDataDiss.

Die andere Möglichkeit (Modell 3) besteht darin, den Nutzer zum bwDataDiss Portal weiterzuleiten und dort die Daten hochzuladen. Allerdings stellt dies für den Nutzer natürlich einen Bruch dar.

In Abschnitt Upload Komponente finden sich weitere Details bezüglich des Uploaders.

Metadata

BwDataDiss bietet verschiedene Möglichkeiten, um Metadaten entgegenzunehmen. Die einfachste Methode besteht im Ausfüllen eines Webformulars auf dem bwDataDiss Portal. Dies setzt aber offensichtlich voraus, dass der Nutzer bwDataDiss ansteuert und damit die Bibliothekswebseiten verlässt (Modell 3).

Um bwDataDiss unsichtbar im Hintergrund zu halten, können die Metadaten auch entweder über eine Schnittstelle von der Bibliothek zu bwDatDiss übermittelt oder aber von bwDataDiss per OAI-PMH abgerufen werden.

Weitere Details zu Medataten finden sich im entsprechenden Abschnitt.

Umsetzung

bwDataDiss wurde in PHP mit Hilfe von Symfony – einem Web Application Framework – implementiert. Außerdem wird JavaScript für einige Funktionen, wie z. B. den zuverlässigen Upload sehr großer Dateien, genutzt.

Es werden drei Benutzerbasisrollen unterschieden: Bibliotheksnutzer, bwDD-Administratorrolle und reguläre Nutzer. Ein Nutzer, der über die bwDD-Administratorrolle verfügt, kann anderen Nutzern weitere Rollen zuweisen.

Web-Frontend

Im Prinzip ist bwDataDiss als Selbstbedienungsportal entworfen, allerdings kann eine Bibliothek die Nutzung auf manche Funktionen beschränken. Es ist erreichbar unter: https://bwdatadiss.kit.edu. Das gesamte Portal ist in Englischer und Deutscher Sprache verfügbar.

API und Datei Management

bwDataDiss stellt eine moderne ReST Schnittstelle (API) bereit, die sowohl xml als auch json Antworten liefern kann und über eine Schlüssel-basierte Authentifizierung verfügt. Die API stellt Funktionen bereit, um Datensätze zu erstellen, Metadaten zu bearbeiten, Dateien in Stücken hochzuladen, Datensätze nach Zustand abzurufen, etc. Außerdem können darüber Archivierungsaufgaben angezeigt und deren erfolgreiche Durchführung gemeldet werden. Ein Großteil der über das Portal bereitgestellten Funktionalität lässt sich auch über die API nutzen.

Upload- Komponente

bwDataDiss stellt ein mächtiges Upload-Werkzeug bereit. Wie im Abschnitt Prinzipien geschildert, ist Datenintegrität eines der wichtigsten Entwurfsziele von bwDataDiss und muss auch durch den Uploader sichergestellt werden. Der Uploader transferiert die Daten vom Promovenden zu bwDataDiss und ist fähig, mit Dateien beliebiger Größe umzugehen. Der Upload kann außerdem unterbrochen und zu einem späteren Zeitpunkt fortgesetzt werden. Weiterhin werden vor und nach dem Upload Prüfsummen berechnet und verglichen, um Integrität sicherzustellen. Die Funktionsweise des Uploaders ist in Abbildung 2 dargestellt.

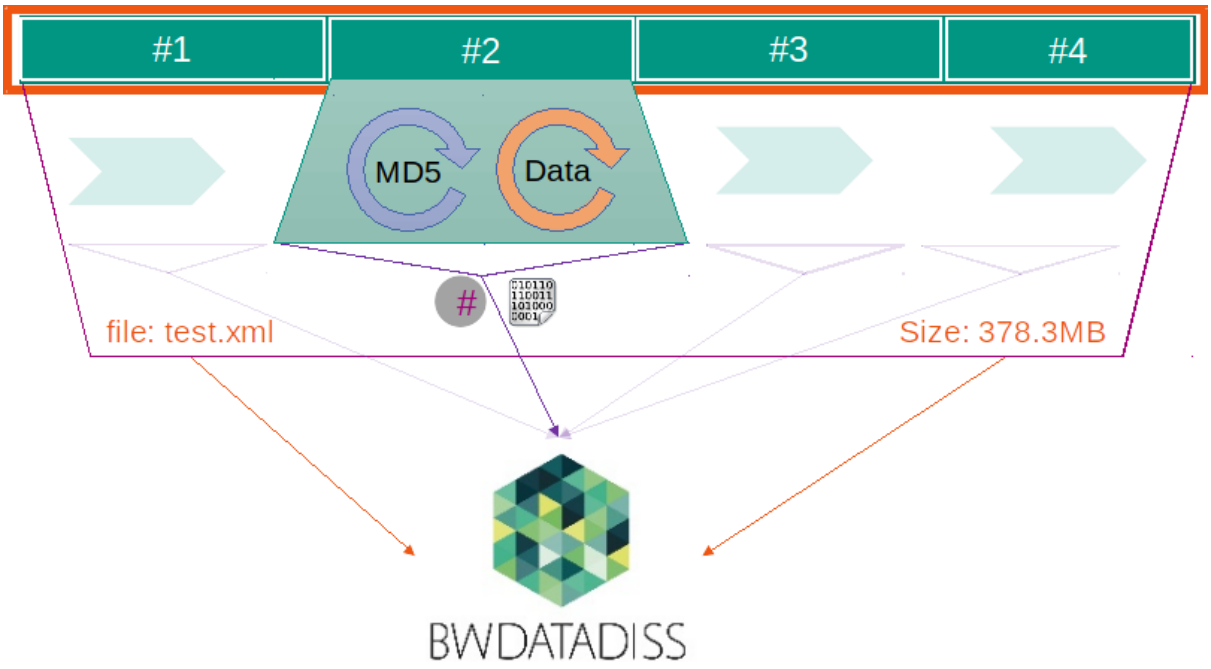


Abbildung 2. bwDataDiss Uploader: Funktionsweise

Beim Upload einer Datei wird diese in Stücke einer festen Größe aufgeteilt und dann separat verarbeitet: zunächst wird eine Prüfsumme berechnet, dann werden die Daten des Dateiausschnittes, zusammen mit der Prüfsumme zu bwDataDiss übertragen. Ein Beispiel ist in Abbildung 2 dargestellt. Der Uploader ist in JavaScript geschrieben und kann mehrere Threads nutzen, um Dateien parallel zu bearbeiten.

Benutzerauthentifizierung und bwIDM

Um Benutzer zu identifizieren setzt bwDataDiss auf einen anderen Dienst, nämlich das „Föderierte Identitätsmanagement der baden-württembergischen Hochschulen (bwIDM)“. bwIDM ermöglicht Nutzern von Universitäten und Hochschulen in Baden-Württemberg. sich per Shibboleth (SAML-basiertes Single Sign On) zu authentifizieren. Für Benutzer, die keiner Universität (mehr) angehören – und sich somit nicht per bwIDM anmelden können – wurde eine Einladungsfunktion geschaffen. Solch eine Einladung kann von der jeweiligen Bibliothek ausgelöst werden, um externe Nutzer zu bwDataDiss einzuladen.

Authentifizierung und Integration durch Bibliotheken

bwDataDiss stellt eine Reihe von Diensten bereit, die durch Bibliotheken genutzt werden können. Das einfachste Integrationsszenario sieht dabei schlicht die Archivierung von Daten und das Bereitstellen von Metadaten vor.

Wie im Abschnitt Aufgabenteilung dargestellt, ist die Bibliothek der Hauptansprechpartner von Promovenden und der Ausgangspunkt für den Veröffentlichungsprozess der Dissertation.

Nun hat eine Bibliothek mehrere Möglichkeiten, die Übertragung der Forschungsdaten zu organisieren: Die Forschungsdaten werden vom Promovenden zur Bibliothek übertragen und später dann von dieser (unter Nutzung der API) weiter zu bwDataDiss. Dies erfordert aber natürlich, dass die Bibliothek entsprechende Systeme vorhält.

Eine andere Möglichkeit besteht darin, den Uploader von bwDataDiss in die Webseite der Bibliothek zu integrieren und die Daten direkt vom Promovenden zu bwDataDiss zu übertragen.

Um authentifizierte Benutzer-Kontexte von der Bibliothekswebseite zum API-Key basierten bwDataDiss Backend delegieren zu können, authentifiziert eine Bibliothek einen Nutzer zunächst per Shibboleth WebSSO. Dann kann sich die Bibliothek selbst mit dem HMAC-SHA256 signierten Benutzer-Shibboleth Token authentifizieren und von der bwDataDiss API einen API-Key für den Benutzer erhalten. Auf diese Art können auch asynchrone API-Anfragen von der Bibliothek an bwDataDiss gestellt werden, ohne den initial hergestellten Nutzerkontext zu verlieren.

Daten Charakterisierung

Datenformate variieren stark zwischen verschiedenen Forschungsdisziplinen und da bwDataDiss nicht auf Forschungsdaten bestimmter Disziplinen beschränkt ist, besteht trotzdem Bedarf an qualitätssichernden und die Langzeitarchivierbarkeit vorbereitenden Maßnahmen. Insbesondere die Langzeitarchivierbarkeit ist in der heutigen Zeit nicht einfach zu erreichen – da schon im geplanten Zeitraum von 10 Jahren Formate unlesbar werden könnten. Um dieses Risiko besser einschätzen zu können, müssen die Formate der gespeicherten Daten bewertet und hinterlegt werden. Für die hinterlegten Daten können dann ggf. entsprechende Erhaltungsstrategien entwickelt werden. Für diese Analyse der Formate und deren Bewertung wurde im Rahmen von bwDataDiss ein entsprechender Dienst entwickelt und integriert.

Erhaltungsrisiken

Das Ziel der Charakterisierung ist es, eine Übersicht über mögliche Erhaltungsrisiken bezüglich Nachnutzbarkeit der Daten bereitzustellen. Dafür muss die logische und strukturelle Repräsentation der Daten – das Dateiformat – bewertet werden. Dies gilt insbesondere wenn eine Dokumentation und Software zu den Dateiformaten existiert. Anhand dieser Informationen können dann Vorhersagen bzgl. langfristiger Nutzbarkeit erstellt werden.

Die Ergebnisse des Charakterisierungsdienstes können für zwei Zwecke genutzt werden: Zum einen als Werkzeug, um Daten vor der Archivierung zu bewerten und Rückmeldung bzgl. Datenformaten zu geben. Anhand dieser Rückmeldung können Wissenschaftler Empfehlungen für Dateiformate gegeben und deren Aufmerksamkeit bzgl. geeigneter Dateiformate gesteigert werden. Zum anderen können die Charakterisierungsergebnisse genutzt werden, um eine Softwaresammlung zu pflegen, die benötigt wird, um mit entsprechenden Daten zu arbeiten bzw. eine Emulationsumgebung bereitzustellen.

Daten Charakterisierung

bwDataDiss stellt einen RESTful Charakterisierungsdienst zur Analyse von Daten bereit. Es wurde FITS als Werkzeug zur Analyse der Dateien ausgewählt, da es verschiedene Charakterisierungswerkzeuge in einem einzelnen, anpassbaren Java-Framework bündelt.

Eine Charakterisierungsanfrage kann durch Stellen einer POST-Anfrage generiert werden, die sowohl eine Referenz zu einem bwDataDiss Datensatz, als auch – optional – auf eine Policy enthält. Aus Effizienzgründen werden die Daten in ein ISO9660 Container (CD-ROM / DVD format) gebündelt. Damit ist ein Vorab-Download der Daten überflüssig, da der Container aus der Ferne eingehängt werden kann und nur Daten, die für die Charakterisierung benötigt sind, übertragen werden. Für die HTTP Anfragen werden entsprechend Range-Requests genutzt. Da die Daten nur im Speicher gehalten werden, stellen parallele Anfragen und Festplattenplatz kein Problem dar.

Beispielanfrage: http://bwdatadiss.eaas.uni-freiburg.de:8080/bwdatadiss/FileFmtCheck/init

```
{
   "objectUrl": "http://bwdatadiss/myset.iso",
   "policyUrl": "http://bwdatadiss/base-policy.txt"
}
```

Der Dienst gibt sofort eine Session ID zurück, welche für Abfragen bzgl. des Charakterisierungsstatus genutzt werden kann. Abhängig vom Umfang der Daten kann es eine Weile dauern bevor die Charakterisierung abgeschlossen ist. Mithilfe der Session ID können die Ergebnisse abgerufen werden – liegen diese noch nicht vor, muss der Aufruf wiederholt werden.

Das bwDataDiss Charakterisierungsergebnis ist eine Dateiformatverteilung bzw. die Anzahl der Dateien pro Dateityp (PRONOM ID).

Beispiel: http://132.230.3.211:8080/bwdatadiss/FileFmtCheck/getResultSummary?sessId=5

```
{
   "summary": [
   {
     "type": "x-fmt/111",
     "value": "GREEN",
     "count": "242"
   },
   {
     "type": "fmt/16",
     "value": "GREEN",
     "count": "2"
   },
   {
     "type": "x-fmt/411",
     "value": "RED",
     "count": "1"
   }
}
```

Es kann auch ein detailliertes Ergebnis angefordert werden, welches eine Liste von Dateien (inkl. relativem Pfad) zu jeder PRONOM ID enthält. Wenn zusätzlich eine Policy angegeben wurde, wird zu jedem Format eine „Bewertung“ angehängt. Im obigen Beispiel enthält die Policy Ampelfarben, wobei den PRONOM IDs „x-fmt/111“ (Plain Text) und „fmt/16“ (PDF) die Farbe Grün und „x-fmt/411(Windows Executable COFF) die Farbe Rot zugewiesen wird.

Archivintegration

Das Archiv basiert auf dem High Performance Storage System (HPSS) und stellt ein hierarchisches Dateisystem zur Verfügung, auf welches per SFTP zugegriffen werden kann. HPSS verfügt über einen integrierten, sehr großen Festplatten Cache, der (logisch) oberhalb des eigentlichen Bandarchivs angeordnet ist. Die Integration des Archivs in bwDataDiss wird realisiert durch zwei verschiedene Techniken:

1. Einen FUSE basierten SSHFS Mount Punkt in das bwDataDiss Hostsystem, welches das SFTP Protokoll versteckt und ein „normales“ lokales Verzeichnis bereitstellt, auf das mit gewöhnlichen Dateiwerkzeugen zugegriffen werden kann.
2. Einen direkten Zugriff per SFTP-Softwarebibliothek ohne Einbindung in das Dateisystem des Hostbetriebssystems.

In beiden Fällen besteht allerdings eine wichtige Einschränkung: Der Zugriff kann deutlich langsamer sein, als ein lokales Dateisystem bzw. übliche NFS/CIFS Mounts. Die gilt insbesondere dann, wenn die Daten von Archivbändern abgerufen werden müssen und nicht vom Festplattencache kommen. Daher ist das Schreiben in das Archiv üblicherweise ausreichend schnell, das Lesen hingegen – von Dateien die nicht mehr im Cache sind – kann sehr langsam sein. Insbesondere kann es lange dauern, bis ein angeforderter Datenstrom überhaupt Daten liefert. Daher musste eine asynchrone, entkoppelte Lösung für den Zugriff auf die Daten im Archiv entwickelt werden.

Die Entkopplung wurde realisiert, indem die Archivintegration vom Apache bzw. PHP basierten Web-Dienst getrennt wurde. Um einen asynchrone Archivanbindung zu realisieren, wurde ein separater Hintergrundworker implementiert, der die bwDataDiss REST API nach Archivierungsaufgaben (schreibe Daten ins Archiv, lese Daten aus dem Archiv) abfragt und das interne bwDataDiss Datenmodell auf das SFTP-Dateisystem abbildet (Abb. 3). Dabei können die beiden o. g. Varianten (SSHFS/FUSE-Dateisystem bzw. SFTP-Direktzugriff) unabhängig voneinander für verschiedene Archivoperationen (lesen/schreiben) und verschiedene Zugangspunkte genutzt werden. Dies ermöglicht die Anbindung mehrerer Archive gleichzeitig, es ist lediglich notwendig, dass sie direkt per SFTP oder einem anderen, ins lokale Dateisystem integrierbaren Protokoll ansprechbar sind.

Zusätzlich wurde in das bwDataDiss Hostsystem ein schneller CIFS-Kurzzeitspeicher eingebunden, um hochgeladene bzw. zum Download bereitgestellte Datensätze für den Webserver performant zugreifbar zu halten.

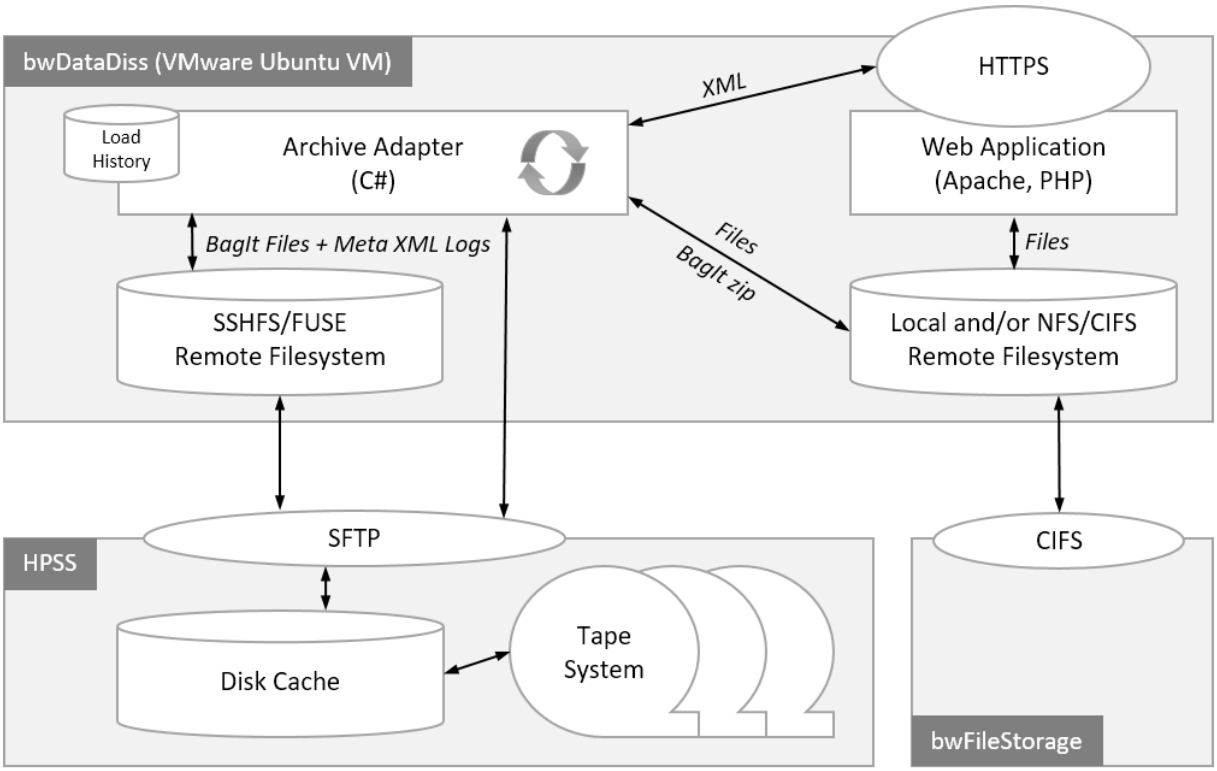


Abbildung 3. bwDataDiss und HPSS

Dort wird ein unkomprimiertes BagIt Verzeichnis angelegt, in welches die lokalen Dateien kopiert werden. Die Metadaten und das Transfer-Log werden ebenfalls in das Archiv geschrieben, was zu einer selbstbeschreibenden Archivdateisystemstruktur (gruppiert nach Bibliotheken) führt. Im Namensschema (Abbildung 4) sind sowohl der [Library-Name] und die [DataSet-ID] (in bwDataDiss) eindeutig, und die [User-EPPN] ist zumindest bei der jeweils zuständigen Bibliothek eindeutig.

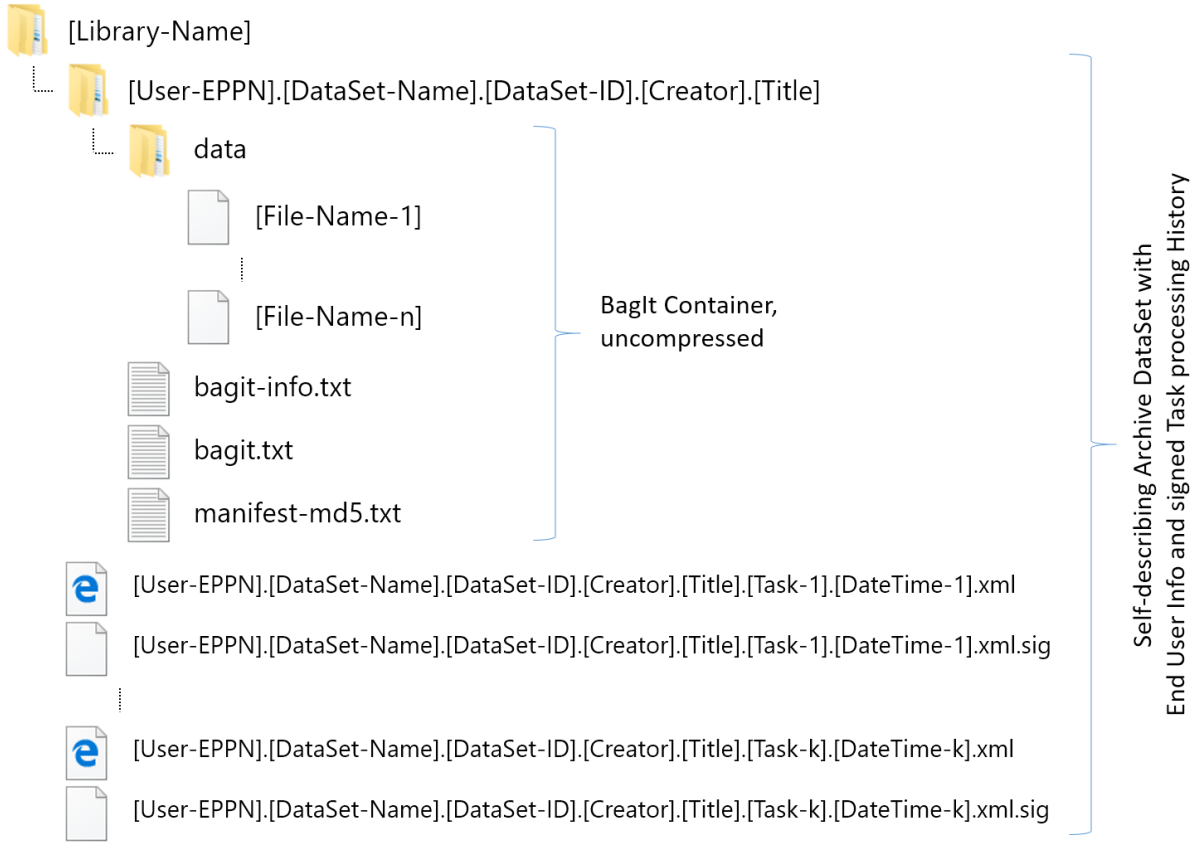


Abbildung 4. bwDataDiss Archivnamensschema

Um einen archivierten Datensatz zu lesen, wird das komplette BagIt Verzeichnis asynchron auf den lokalen Dateisystemcache kopiert, von wo aus der Webserver direkten Zugriff auf die Dateien hat (Abb. 3, CIFS-Mount von bwFileStorage). Optional kann auch eine Zip Datei vom Archivworker angefordert werden, die alle Dateien inklusive der BagIt Dateien enthält. Alle Archivierungsaufgaben werden mit Prüfsummen gegengeprüft und bei Erfolg der bwDataDiss API entsprechend mitgeteilt. Sollte ein Fehler auftreten, wird die Aufgabe wiederholt.

Metadata

bwDataDiss kann durch alle Hochschulbibliotheken im Land Baden-Württemberg genutzt werden und ist auf keine Fachdisziplin festgelegt. Das Metadatenschema von bwDataDiss trägt dem Rechnung und ist entsprechend generisch angelegt. Das Schema ist außerdem an den RADAR[7] Metadaten Kernel angelehnt.

7 Das Projekt RADAR oder Research Data Repositorium stellt Infrastruktur für das Forschungsdatenmana gement bereit und ist ein gemeinsames Projekt des FIZ Karlsruhe, des Steinbuch Centre for Computing (SCC), der Ludwig-Maximilians-Universität München (LMU) and der Technischen Informationsbibliothek (TIB).

Das Metadatenschema besteht aus nachfolgenden Punkten:

- Titel*
- Zusatztitel
- Ersteller*
- Beitragende
- Abstrakt*
- Schlagwörter
- Liesmich*
- Erstellungsjahr (Ende)*
- Creation Year (Beginn)
- Herausgeber*
- Jahr der Veröffentlichung*
- Klassifizierungen*
- Ressourcentyp
- Lizenz*
- Rechteinhaber
- Embargodatum
- Zusätzliche Metadaten

Punkte, die mit * markiert sind, sind Pflichtfelder und müssen bei bwDataDiss angegeben werde. Ein paar der Punkte bedürfen einer Erklärung:

Liesmich: Im Speziellen für Forschungsdaten ist es wichtig zu wissen, wie die bereitgestellten Forschungsdaten organisiert sind und ggf. genutzt werden können. Auch technische Informationen oder Hinweise zur genutzten bzw. der zu nutzenden Software können hier ihren Platz finden.

Klassifizierungen: bwDataDiss können beliebige Klassifizierungen übergeben werden – mindestens jedoch eine. Über das Portal werden Auswahlhilfen für die DDC und für eine Klassifizierung nach DGF Fachgruppen angeboten.

Resourcentyp: Vom RADAR Metadaten Kernel entliehen, kann es einen der nachfolgenden Werte, der den Typ der Forschungsdaten beschreibt, annehmen: audiovisual, collection, dataset, image, model, software, sound, text, workflow, other. Diese Beschreibung der Art der Forschungsdaten ist relevant für die Aufnahme in den Data Citation Index von Thomson Reuters /Clarivate Analytics.

Lizenz: bwDataDiss steht hinter Open Access und stellt daher alle möglichen CC Lizenzen zur Auswahl bereit (z.B.: CC-BY und CC-BY-SA). Allerdings können Situationen eintreten, wenn diese Lizenzen nicht ausreichend sind und erlauben es Bibliotheken daher, eigene Lizenzen zu hinterlegen.

Embargodatum: Unter Umständen ist es nötig, den Zugriff aus Forschungsdaten zu unterbinden. Dies kann durch die Einrichtung eines Embargos erreicht werden. Während des Embargos ist es möglich, einzelnen Nutzern trotzdem Zugriff auf die Daten zu gewähren.

Mit bwDataDiss in wenigen Schritten zum neuen Service der Forschungsdaten-veröffentlichung – am Beispiel KITopen

Wahl des Modells der Integration

bwDataDiss ist an der KIT-Bibliothek seit Anfang März im Produktivbetrieb. Damit ist es nun am Campus des KIT erstmals möglich, Forschungsdaten aller Disziplinen und Formate gemeinsam mit der Promotionsschrift kostenlos zu veröffentlichen. Neben der technischen Integration des Dienstes bwDataDiss mit dem Repository der KIT-Bibliothek waren dazu auch noch weitere organisatorische Schritte erforderlich. Das Repository KITopen ist Teil eines am KIT entstehenden Forschungsinformationssystems und befindet sich in einem großen technischen Umbruch. Schwerpunkte sind neben der Veröffentlichung von Volltexten und bibliographischen Daten von KIT-Wissenschaftlern die Bedienung der unterschiedlichen Berichtspflichten das KIT und insbesondere der Helmholtz-Gemeinschaft. Aus praktischen Gründen war daher die Abgabe der Promotionen über die KIT-Bibliothek nur in Form eines gesonderten elektronischen Formulars möglich. Die eigentliche Datenerfassung erfolgte über die Mitarbeiter in das Repository. Es galt daher für die Integration von bwDataDiss, diesen Sonderworkflow zunächst in das sonstige Repository-Umfeld einzupflegen. Darauf aufbauend erfolgten dann die weiteren Prozessschritte für die Veröffentlichung der Forschungsdaten.

Ziel war es, den Spagat zu schaffen und den Nutzern keine unnötigen Systembrüche zu verursachen und dennoch das Repository um Komponenten von bwDataDiss zu erweitern, um unnötige Doppelimplementierungsarbeiten zu vermeiden. Daher fiel die Wahl auf Modell 2 das beinhaltet, die Erfassung der Metadaten in die Workflows des Repository der Bibliothek vollständig zu integrieren und auch die Nutzerkommunikation über KITopen zu veranstalten. Für die technischen Spezialanforderungen wie den Upload großer Mengen an Forschungsdaten, die unterschiedliche Uploadzeiten und Speichernutzungskapazitäten mit sich bringen, sollte auf die technische Infrastruktur von bwDataDiss zurück gegriffen werden. In der Praxis des KIT nutzen die Promovierenden daher nun den Upload im Repository , welcher direkt die Schnittstelle von bwDataDiss anspricht. Siehe dazu auch: Upload im Abschnitt: Möglichkeiten der Integration einzelner Komponenten.

bwDataDiss stellt neben Kern- auch Hilfsfunktionen bereit. So sind am KIT ca. 40% aller Promovenden zum Zeitpunkt der Abgabe der Forschungsdaten nicht mehr am KIT beschäftigt bzw. verfügen über keinen Account mehr, mit welchem eine Anmeldung bei bwIDM möglich wäre. Für diese Fälle wird eine Einladungsfunktion von bwDataDiss bereitgestellt, die es ermöglicht, Accounts auf bwDataDiss zu erstellen. Des Weiteren ist es über die bwDataDiss API möglich, bwDataDiss Benutzerkonten und Passwörter zu prüfen und damit diese Accounts zur Autorisierung in anderen Systemen – wie KITopen zu nutzen. Nutzer können sich also auch mit bwDataDiss Accounts bei KITopen anmelden auch wenn eine Anmeldung per bwIDM nicht möglich ist.

Formulierung der Policy für die Nachnutzung von Forschungsdaten

Ein wichtiger Aspekt von bwDataDiss ist die Bereitstellung der Forschungsdaten zur Nachnutzung durch andere Forschende und Interessierte. Die Forschungsdaten sind dazu mit den entsprechenden Metadaten als auch mit persistenten Links und Identifier für die Zitation versehen. Laut Policy von KITopen sind alle Forschungsdaten grundsätzlich unter eine Open-Access-Lizenzen gestellt. Beim Upload können auf Wunsch des Forschenden Embargos eingerichtet werden, die die Nutzung der Forschungsdaten für einen gewissen Zeitraum verhindern. Auf eine Limitierung des Zeitraums wird vorerst verzichtet. Die Embargofrist kann von der Bibliothek gesteuert werden. Innerhalb dieser Frist können ausgewählte Nutzer Zugriff auf die Forschungsdaten erhalten. In der ersten Stufe der Einführung des Dienstes erfolgt das durch die Mitarbeiter des Teams KITopen.

Aufbau eines Services zur Qualitätssicherung und Beratung zur Langzeitarchivierung

Für die inhaltliche Vollständigkeit und Konsistenz der Forschungsdaten sind die Datengeber selbst verantwortlich. Hier muss sich die KIT-Bibliothek erst langsam an den neuen Service herantasten und schrittweise vorgehen. Zunächst liegt daher der Fokus auf der bewährten formalen Prüfung der Metadaten. Dabei wird darauf geachtet, dass die erläuternden Felder wie „Liesmich“ und „Abstract“ bzw. „Schlagwörter“ ihre beschreibenden Funktionen entsprechend erfüllen. Wichtig ist hier eine frühzeitige Rückmeldung an die Forschenden und der Einstieg in die Kommunikation. Darauf aufbauend ist der Aufbau weiterer Beratungsservices für die Langzeitarchivierung angesichts der heterogenen Landschaft der Forschungsdaten ein komplexes und drängendes Feld. BwDataDiss unterstützt diese neuen Qualitätsprüfungsprozesse durch Bibliotheken in Form von Dateitypcharakterisierungen und gibt entsprechende Rückmeldungen an die Datengeber bzw. Bibliotheken. Die Basisinstallation von bwDataDiss verweist dabei auf Empfehlung zu Dateiformaten von der Library of Congress und der Cornell University.[8] Im Fall von KITopen wird zunächst auf die automatisierte Rückmeldung der Charakterisierungsergebnisse an die Nutzer verzichtet, da man zunächst aufgrund der Rückmeldungen Erfahrungen im Umgang mit den Dateiformaten aufbauen möchte. Das dafür nötige Expertenwissen wird in den nächsten Jahren am KIT in verteilten Rollen kooperativ aufgebaut werden, so dass KITopen sich schrittweise erweitern wird. Die technische Implementierung von bwDataDiss erlaubt sogar, daran gemeinschaftlich und im Land verteilt zu arbeiten und Policies gemeinsam zu nutzen.

Rechtliches und Formales

Aufgrund der Projektbeteiligung erübrigt sich eine weitere vertragliche Regelung zwischen der KIT-Bibliothek und bwDataDiss. Für weitere teilnehmenden Bibliotheken ist aber in jedem Fall der Abschluss eines Kooperationsvertrags erforderlich.

Die durchzuführenden Maßnahmen um Datenschutzansprüchen zu genügen, unterscheiden sich je nach Integrationsmodell, können aber auch von Umsetzungsdetails abhängen. Nach unse-

8 https://ecommons.cornell.edu/page/support#format, http://www.digitalpreservation.gov/formats/fdd/browse_list.shtml, zuletzt geprüft am 24.2.2017.

rem Informationsstand ist im Falle von Modell 1 ein Vertrag zur Auftragsdatenvereinbarung zu vereinbaren, im Fall von Modell 2 kann dies auch der Fall sein und im dritten Modell ist ein solcher normalerweise nicht nötig. Allerdings möchten wir darauf hinweisen, dass Sie sich auf jeden Fall an ihren Datenschutzbeauftragten wenden sollten. Die erforderlichen Unterlagen liegen für bwDataDiss vor.

Schlusswort

bwDataDiss ist eine wichtige Initiative auf Landesebene die sich gut in die Dienste einer Bibliothek bzw. der jeweiligen Hochschule einpasst und flexibel integriert werden kann.

Mit den drei Modellen werden sowohl kleinere Einrichtungen, als auch größere, die ggf. einen höheren Integrationsaufwand betreiben können, adressiert.

Durch die Charakterisierung der Forschungsdaten wird es Bibliotheken erleichtert, die Qualität der Forschungsdaten zu sichern bzw. abzuschätzen. Außerdem wird es dadurch möglich, entsprechende Erhaltungsstrategien für Forschungsdaten im Archiv zu erstellen und eine langfristige Nachnutzbarkeit der Daten zu gewährleisten. Dies ist derzeit in der Landschaft von Serviceanbietern für Forschungsdatendienste weitgehend singulär.