Developing a holistic research data management strategy for a university

Making preservation planning and long term access first grade citizens

Dirk von Suchodoletz
Jan Leendertse
eScience Department, Computer Center, University of Freiburg
Germany
@rz.uni-freiburg.de
0000-0002-4382-5104
0000-0001-5676-493X

Saher Semaan
eScience Department, University Library, University of Freiburg
Germany
@ub.uni-freiburg.de
0000-0001-7487-7348

Klaus Rechert
Rafael Gieschke
Professorship in Communication Systems, University of Freiburg
Germany
@rz.uni-freiburg.de
0000-0002-8807-018X
0000-0002-2778-4218

Björn Goldammer
Dimitri Tolkatsch
Freiburg Research Services, University of Freiburg
Germany
@frs.uni-freiburg.de
0000-0003-1768-7315
0000-0003-2266-2163

Abstract – A modern, integrated research data management (RDM) enables reproducible and verifiable research, the linking of interdisciplinary expertise, the sharing of research for comparison and integration of different analysis results and metadata studies to derive further knowledge. An increasing adoption of FAIR principles and requirements by funding agencies on open data has significantly benefited overall quality, reuse, and sharing of research results. For scientists, a proper data management is a crucial element to prove their findings and make them reproducible. As a consequence, RDM had to become an integral part of the science support infrastructure in todays research institutions in the past decade. Scientist of various disciplines should be supported over the complete data lifecycle starting from holistic planning of future projects to RDM related services provided: As RDM is a multidimensional endeavor requiring various skills, tasks ranging from community specific to community needs are optimally handed to the best qualified provider. The presented concepts and considerations are work in progress while establishing an organizational frameworks for a research university. Completely reproducible data publications including the relevant data-sets' context are the ultimate goal. These require appropriate service components like EaaSI. The university strives to profit from the overlaps in RDM and digital preservation and to define the handover of tasks from the first to the latter.

Key words – research data management, planning, FAIR, long-term access, re-use of data, data management plan, data publication

I. Introduction

Research data management has received a growing amount of attention as many scientific disciplines are increasingly data-driven [1]. Universities as public bodies are expected to adhere to the *FAIR* principles [2]. These state that data should be made available in a findable and accessible manner, i.e. in open and public repositories, and be interoperable and reusable, i.e. published in non-proprietary formats, sufficiently annotated and reproducible. Compared to the establishment of traditional services like university libraries and archives digital RDM is a fairly novel development still waiting to get fully embraced by all stakeholders. It fosters the linking of interdisciplinary expertise and the combination of different analytical results. Crossing domain boundaries achieved through comparative and integrative analyses provides additional insight in the examination of research questions that goes far beyond individual fields. Successful collaborative work and leveraging data of different modalities—from many sources and experiments, and pre-processed or pre-analyzed using a variety of algorithms—requires contextualization of the data according to the respective research objective.

The challenge of how to preserve computations and IT based workflows became increasingly well understood in the past decade [3]. Those issues are already addressed by preservation strategies for more than 20 years [4]. They are used by memory organizations dealing with software artifacts as cultural heritage. How-

17th International Conference on Digital Preservation
iPRES 2021, Beijing, China.

DOI: 10.1145/nnnnnnn.nnnnnnn

ever, the adoption of those preservation aware procedures did not reach all scientific domains. They are not broadly implemented while preparing concrete research projects or planning continuous access and reuse concepts. The crystallization of those considerations how to preserve the value of research data or scientific workflows are not sufficiently embedded into the institutional strategy and policy documents. Especially the links to already existing services or conceptualized data management frameworks beyond abstract statements are seldom seen in application forms for external funding or self-binding codes of conduct (CoC).

II. Problem statement

A typical research university hosts a wide area of diverse disciplines with varying scientific cultures. The types and amount of data handled differs significantly. Fast pacing scientific fields which embraced research data management methodologies and best practices for a while coexist with domains which still need base level nudging to FAIR data handling. Thus, there is a huge discrepancy between the actual state of abstract concepts on continuous access and reproduction of RDM versus the actual uptake in scientific institutions. This is not only due to different levels of digitization of research workflows but also to the actual implementation of institutional policies regarding good scientific practice in the digital age. Especially junior scientists do not necessarily receive the support and qualification by their supervisors and the institution's training programs. Nevertheless, there is an increasing pressure towards open and reproducible research by both grant providers and international scientific practice an university needs to adapt to. It is to be expected that the success of individual researchers and institutions is not only derived from the amount of traditional publications produced in high ranking journals but that it stems from research data made open for re-use and reproduction as well.

Many institutional strategies put the emphasis on the "now", but neglect long term access and re-use of data and workflows. Existing long term preservation and RDM solutions are often unknown, not considered or not affordable to some research groups. The cost models for RDM and required services are inconsistent across disciplines and institutions, and can form barriers for consistent lifecycle management of digital research. [5]

Reproducibility or scientific findings still has a way to go, and thus requires not only access to the data itself, but also to its context. The latter may need including technical metadata to replicate the actual data collection, measurements or computation with the software and instrumentation used as well as the the broader context in the sense of scientific field customs. [6]

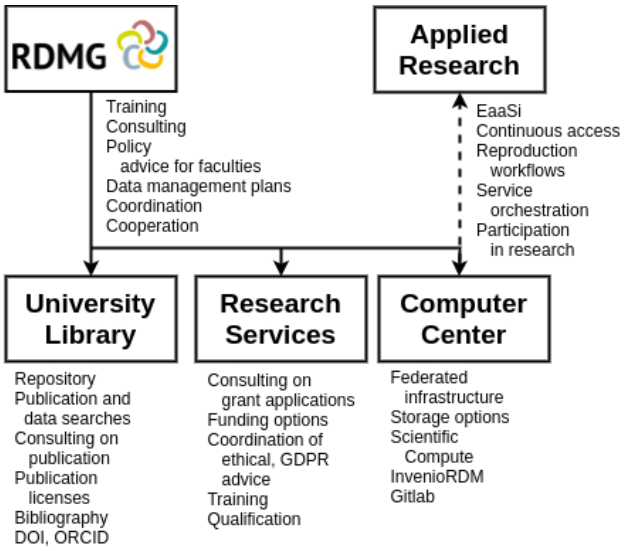

Figure 1: The organizational structure and task distribution within the Research Data Management Group at the University of Freiburg

III. Developing a strategy for the University of Freiburg

The FAIR Data and Linked Open Data principles provide crucial guidelines for any infrastructure receiving, processing and publishing research data. While various initiatives like the Research Data Alliance (RDA[1]) have made suggestions on best practices and processes towards fulfilling these principles, it is nevertheless always up to the individual institution's initiative to adhere to them. Thus, the University of Freiburg represents a typical example of the two velocities in RDM, maintaining a leading role in research regarding continuous access to digital assets and its contexts but pretty much neglecting the ongoing relevant developments in that domain in practice.

The actual workload of the Research Data Management Group (RDMG)[2] reveals that the initiation of a university-wide policy-based system of CoC tailored to the needs of scientific fields is at the beginning, the concrete demand for consulting is mainly driven by scientists applying for funding tracks of organizations asking for those codes. Similar requests arise at the end of funded research projects. The requirements of those projects include increasingly the obligation to store research data, which finds scientists regularly surprised.

A. Status-quo

The university as a research institution started comparably late into a structured approach towards RDM. Practitioners in the university library, the Freiburg Research Services and at the computer center inaugurated the RDMG bottom-up by forming a virtual organization out of existing science support personnel just three years ago (Fig. 1). The aggregated amount of efforts in

[1] See https://rd-alliance.org/rda-disciplines
[2] See https://rdmg.uni-freiburg.de

this field equals to just sligthly more than one FTE, comparably small for a self proclaimed research university of 330 professorships and over 4000 employees. Nevertheless, the formal structure is really beneficial regarding better synchronization of activities and improved coordination. One of the first tasks completed was the preparation of an university-wide policy statement obligating researchers to follow the guidelines for safeguarding good research practice of the German Research Council [7]. The group furthered strategy documents and translated them into a structured process. A year ago a new position got established in the RDMG to foster outreach and permeation into the faculties.

B. Mapping on planning frameworks

The cornerstone of the strategy documents is to cover the complete data lifecycle with an emphasis on holistic planning during grant applications. Many problems regarding good scientific practice, reproduction of results and re-use of data, can be avoided through a structured and proactive approach regarding filetypes, agreements and licenses on data obtained or produced and the proper choice of tools and workflows. This implies a joint support of researchers especially in the planning phase. This supports the adherence to the expectations set by funding bodies regarding openness and re-usability of project results.

IV. Creating the basic institutional framework

The three founding departments of the RDMG already cover a significant part of the data life cycle requirements (Fig. 4), but need to complement these activities with concrete recommendation guidelines on the governance and the licensing of research data. While there is long-established support by the university library for researchers in the area of literature supply and search as well as electronic publishing (for ”small” research data sets) and metadata curation on the library’s own developed institutional repository, a shift in focus is emerging in the coming years from recommendations on appropriate literature and journals to relevant science portals and data repositories of the respective discipline.

In addition to distributed and collaborative research environments and repositories a researcher must be uniquely identified and her/his scientific output must be correctly linked. The University of Freiburg became a member of ORCID and obliged researchers to use this ID. Consistent links between authors and contributors and their research data enable proper aggregation and sharing of (meta)data between platforms via standardized protocols, e.g. OAI-PMH. In addition, access can be better regulated, visibility of scientific output increased and re-use simplified.

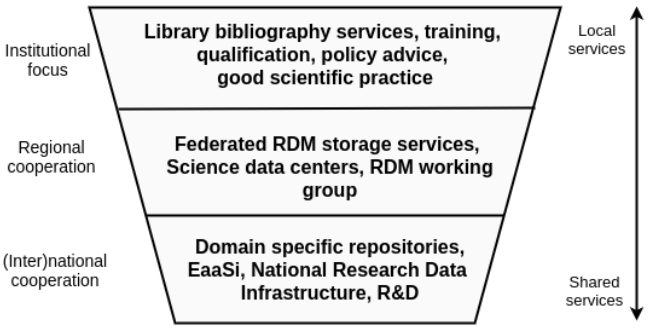

Figure 2: Between providing local support and regional and (inter-)national cooperation in RDM infrastructures and services.

V. Embedding the institutional strategy into a wider context

Covering all aspects of RDM is too big for a single institution and its entities. Thus the RDMG collaborates with other universities and research institutions with the statewide RDM working group (Fig. 3). It participates in the statewide Science Data Center initiative (through BioDATEN[3]) to dive deeper into the requirements for a scoped discipline and learn for the wider application. The university embraces the initiative for the National Research Data Infrastructure (NFDI) and leads consortia on fundamental plant research (DataPLANT[4]) and microsystems technology (Mat-Werk[5]). (Fig. 2)

A. Common working fields

As consortia leads, DataPLANT and Mat-Werk have to deal with cross-cutting topics defined by NFDI. [8] The consortia are obligated to scrutinize their portfolios in order to find common working fields with other consortia. The technical and organizational base for smooth collaboration will be an important argument for scientist to join certain universities. So the universities should embrace those developments and position themself as home institution for flexible research cooperations. The advantage for the universities might be that they can clean their in-house service portfolio in combination with intelligent service sharing.

B. Infrastructural solutions

To allow for cross-disciplinary access and reuse of research data, a certain level of standardization and coordination is required for several metadata and data properties. Several scientific communities share the view of a Research Data Commons as an overarching virtual expandable infrastructure to leverage user involvement and collaborative data-driven research. This includes for example joint cloud services like the de.NBI and bwCloud, access to high performance computing and collaborative workspaces, and a common authenti-

[3] See https://biodaten.info
[4] See https://nfdi4plants.de
[5] See https://nfdi-matwerk.de

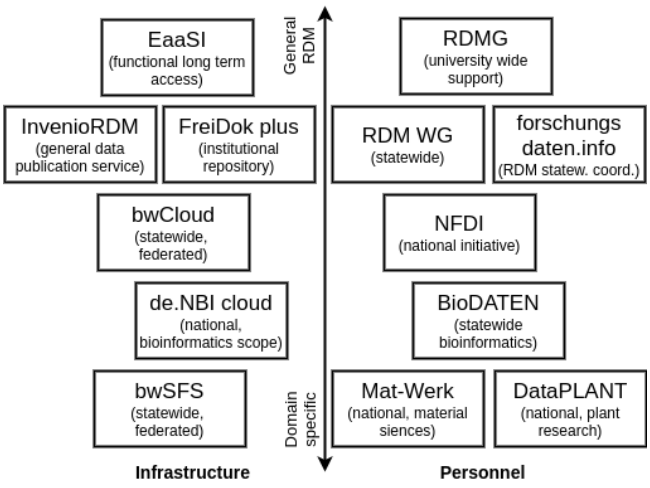

Figure 3: Personnel and infrastructural support the University of Freiburg is actively involved in on different layers.

cation and authorization infrastructure. The Research Data Council calls for a common strategy for interacting with the existing large-scale compute and data infrastructures in Germany and the need for harmonization among these centers. The University of Freiburg computer center and the university library bring in significant contributions regarding technical infrastructure and operation concepts and expertise. The infrastructure is partially shared with other universities and scientific communities in computing and run in a federated operational model. The scalability is an inherent feature of the operation model expressed through a collaborative infrastructure. (Fig. 3)

The storage system bwSFS (storage for science) [9] is a federated storage offering a technical base. It is added by higher service layers for archiving, publishing, and versioning. Complementary to the already established institutional repository FreiDok+[6] InvenioRDM will be used for archiving and publishing.

VI. Preservation of the research context

In pretty much every domain, preserving just the data objects risks losing access to the research context, and thus, eventually the ability for data interpretation and data reuse. Hence, data, data-processing software and sometimes even base-level technology stacks need to be considered in a joint context [6]. General considerations for long-term (ten years and more) reuse, validation and reproduction of research outputs is still in its infancy. The University of Freiburg brings in a strong team working on sustainable long-term access for over 15 years [6]. Concepts and practice of software citation have been developed with national and international consortia, as well as guidelines and infrastructure to manage and preserve software dependencies which should be made available to a wide range of research and memory institutions. Still, with technical progress and especially the advance of virtualization, container, cloud and related technologies, research environments became interconnected and interactive, and research data and software intertwined, such that ensuring meaningful access to data and reuse requires constant attention and development [10]. In order to ensure FAIR data principles, especially long-term re-usability of a wide variety of research outputs, novel methods are required for all scientific domains, and to be integrated in research data management strategies.

VII. Holistic preservation planning and continuous access

Preservation of research data, including the ability of re-using, reproducing or replicating of software-based methodology, is not yet fully embedded into typical research data management strategies. Both domains and especially their communities of practice overlap significantly, but do not speak a coherent language or pursue a single strategy. For the scientific community, in particular for active research projects, simplified access, integration into contemporary infrastructure, and usability are the main focus. Long-term access and specifically reuse, are usually not a major concern yet. By integrating infrastructure addressing long-term access and reproducible research within current workflows, the RDMG is working towards the concept of *continuous access*, with the goal that any RDM step is connected with a preservation strategy in an automated way, e.g., a successful commit of code or a successful run of a workflow, will trigger a pipeline of events, such that code dependencies are collected and maintained [10], [11], containers are preserved [12] or the workflow can be re-run, albeit without the high performance of dedicated compute facilities, from dedicated long-term infrastructure such as EaaSI.

A. *Joint creation of data management plans*

A holistic planning phase documented a data management plan (DMP) is a prerequisite of a successful grant application and project kick-off (Fig. 4). Together with the individual researchers, the RDMG develops plans fitting their project requirements along the guidelines of the funding organization. The DMP of the proposed project estimates the required funds and compute resources as well as the amount for data to be stored and published in the long run. Reproducibility in this domain includes the tracking of tools and their version and the piping sequence if more tools were used. Research cannot be planned ahead in every details, but the environment used for trying and erring can be prepared to keep the steps traceable. Such an environment is at the same time a good base for later publication or archiving. It focuses on open standards and tools to guarantee continuous accessibility and sustainability. The re-usability is not only a technical issue. Inconvenient licenses may prevent a hassle-free use in new research projects. But not every computation or workflow is possible with open-source tools, so that mitigation

[6] See https://freidok.uni-freiburg.de

measures for those cases should be applied.

B. Legal aspects

Legal aspects such as licensing of data and software, intellectual property rights, data protection, and privacy are of significant importance for scientists. Research data often stems from international collaborations or is shared with international colleagues. A specific concern with regard to the development of an university and faculty policies is to clarify parameters for the commercial use of data. There is a general need by the research community in legal support. Here, the university still has to close a gap in appropriate staffing. Such an expert should both coordinate with other research institution on common questions. Sensitive and especially person-related data to be used for research adds particular challenges, both on the ethical and the legal side. While not all scientific communities might encounter GDPR-related concerns, ownership and responsibility play a role.

VIII. Sustainable financing

Concepts for fair distribution of (financial) responsibilities between faculties and central administration is needed and a long-term perspective of financial sustainability of RDM services and infrastructures on and off site should be developed. Several options exist to provide base line funding and cover dynamic requirements:

- Restructure the existing service landscape of repositories, storage and compute resources at the university library and the computer center to accommodate required RDM components as well.
- Let the benefiting entities like the faculties contribute a fair share by e.g. paying the fees for external repositories researchers store data at and providing personnel for the necessary coordination tasks.
- Acquire additional funds for the core research portion of RDM from science supporting agencies.

The latter can be achieved in part through the design of future grants for research. As funding agencies ask for compliance in RDM, they should provide the necessary means through dedicated funding. This could be embedded into the application process and the co-development of data management plans of both researchers and RDM support. Consultation with the own scientific community e.g. through the NFDI could ensure that new projects work according to current workflows, with modern tools and jointly established standards regarding long term access to data sets. To be able to respond flexibly in this regard, a sufficient amount of staff should be allocated to support these communities through e.g. data stewards [13].

A conceivable approach for the future would be that newly submitted research proposals should apply for funds for support services in personnel or infrastructural form, which, if approved, flow directly to the RDM infrastructure and services. This allows in particular to deal well with smaller projects and fractions of positions for data management. It also ensures the sustainability and continuity of the experts employed by the RDM team. Recruitment and training of suitable personnel and follow-up employment after the end of the project are often unsolved challenges in the science domain. Participation of the researchers and involvement of the scientific communities in the design of such processes should be ensured via the governance structures within the university.

A. Risk-based approach to funding

Similar discussions within funding organizations are known and partly reflected in guidelines of some of the funding lines. Nevertheless, the actual granting is lagging behind on both sides, on the side of funding agencies as well as on the side of applying researchers. Part of this overarching problem is the assessment phase. The reviewers show also different velocities in recognizing the shift of IT servicing in scientific workflows. Some of them see the IT needs covered by acquiring hardware rather than using cloud infrastructure. These uncertainties may lead to conservative applications relying on well introduced, but more and more outdated planning schemes endangering the long-term preservation and re-use of research data.

IX. Outlook

Modern research data management is characterized by collaboration within the field and spanning cross-domain boundaries (Fig. 2). While the data organization, digital workflows and metadata standards are set by the different scientific communities, RDM teams need to provide guidance and a solid set of research data management related tools and services. By this means the increasingly blooming landscape of research data management (RDM) services and infrastructures help to facilitate the acquisition, processing, exchange, archival and application of research data sets. These developments pave the way to a cultural change towards adding further contributions to the researchers' scholarly records. Data and workflow publication is still a quite novel concept, but of increasing relevance. This adds to the institution's hightened visibility in the (international) scientific landscape and researchers benefit from better services, faster access to new findings.

To solidify these developments they need to be supplemented with holistic planning and means for long-term reproducibility. Data stewards in a hinge role between domain specific expertise and experience in those new concepts and services can foster these developments through guidance and training of the research groups. It changes the role of university libraries focusing broader on data publication and the necessary services on long-term access. From this, library profession-

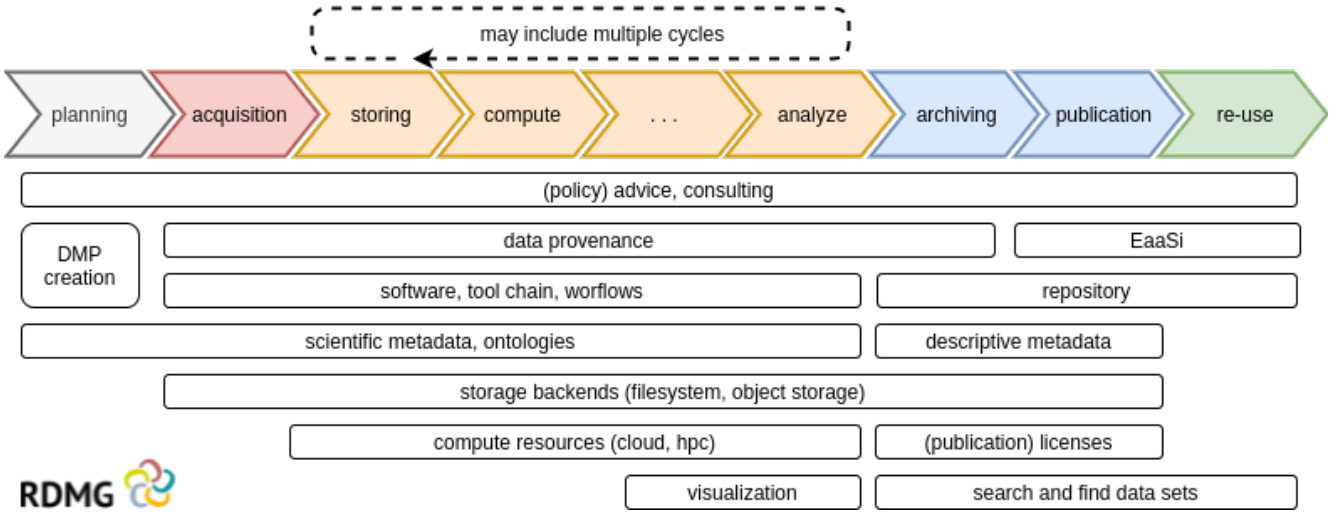

Figure 4: Project planning is the crucial starting point for a holistic RDM strategy. The RDMG provides support and guidance over the complete data lifecycle after successful project funding.

als should adjust their role become consultants of researchers how to manage reserach data. All this is an ongoing process where all stakeholders should be committed to policies as a common ground.

A shift to organizing IT as a landscape of services makes it easier to assign needs to resources. Termed project could be easier accomplished in time, if staffing from a drained job market were replaced by sharing models. Resource sharing in a service landscape would allow an efficient support of a larger number of groups. One pre-condition for such a shift is a governance structure with clear responsibilities in conjunction with a flexible cost management. Supporting providers, i.e. computer centers or libraries, should embrace the new flexibility of cloud services.

RDMG sees this as its obligation to raise awareness to look beyond the scope of single research endeavors. This includes trainings, consulting, and connecting to base services covering their needs. Within this process, questions regarding data protection and security in the sense of availability can be addressed.

A peculiarity at Freiburg university is the tight link between practical support and cutting edge research in long-term access. A major step in this regard is to maintain access to suitable technical ecosystems e.g. through emulation frameworks like EaaSI which form a baseline infrastructure that could be shared among a wide range of research institutions. EaaSI is one answer for those producing code as part of their research or using special software environments. However, the long-term establishment of EaaSI is not solved, yet. It is part of the provider's strategic operations to moderate the recalibration of services.

Acknowledgements

Part of the activities and insights presented in this paper were made possible through the collaboration in the Science Data Center BioDaten and the NFDI consortium DataPLANT. DataPLANT, 442077441, is supported through the German National Research Data Initiative (NFDI 7/1).

References

[1] A. J. G. Hey, Ed., *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, Washington: Microsoft Research, 2009.

[2] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[3] T. Allançon, A. Pietri, and S. Zacchiroli, "The software heritage filesystem (swhfs): Integrating source code archival with development," in *2021 IEEE/ACM 43rd International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, IEEE, 2021, pp. 45–48. doi: 10.1109/ICSE-Companion52605.2021.00032. [Online]. Available: https://arxiv.org/abs/2102.06390.

[4] F. C. Y. Benureau and N. P. Rougier, "Re-run, repeat, reproduce, reuse, replicate: Transforming code into scientific contributions," *Frontiers in Neuroinformatics*, vol. 11, p. 69, 2018, issn: 1662-5196. doi: 10.3389/fninf.2017.00069. [Online]. Available: https://www.frontiersin.org/article/10.3389/fninf.2017.00069.

[5] H. Hockx-Yu, "Superb stewardship of digital assets–developing a strategy for digital archiving and preservation at the university of notre dame.," in *iPRES*, 2017.

[6] S. A. Iqbal, J. D. Wallach, M. J. Khoury, S. D. Schully, and J. P. Ioannidis, "Reproducible research practices and transparency across the biomedical liter-

ature," *PLoS biology*, vol. 14, no. 1, pp. 1–13, 2016. doi: 10.1371/journal.pbio.1002333.

[7] Deutsche Forschungsgemeinschaft, "Guidelines for safeguarding good research practice," *Code of Conduct. de (Sept. 2019). Publisher: Zenodo. doi*, vol. 10, pp. 12–13, 2019.

[8] F. O. Glöckner, M. Diepenbroek, J. Felden, J. Overmann, A. Bonn, B. Gemeinholzer, A. Güntsch, B. König-Ries, B. Seeger, A. Pollex-Krüger, J. Fluck, I. Pigeot, K. Toralf, T. Mühlhaus, C. Wolf, U. Heinrich, C. Steinbeck, O. Koepler, O. Stegle, J. Weimann, T. Schörner-Sadenius, C. Gutt, F. Stahl, K. Wagemann, T. Schrade, R. Schmitt, C. Eberl, F. Gauterin, M. Schultz, and L. Bernard, "Berlin Declaration on NFDI Cross-Cutting Topics," eng, Sep. 2019. doi: 10.5281/zenodo.3457213.

[9] F. Bartusch, K. Glogowski, U. Hahn, M. Janczyk, S. Kaminski, J. Krüger, V. Lutz, G. Schneider, M. Seifert, D. von Suchodoletz, T. Walter, and B. Wiebelt, "Defining the future scientific data flow for multi-disciplinary research data," in *E-Science-Tage 2019: Data to Knowledge*, Heidelberg: heiBOOKS, 2020, pp. 110–127, isbn: 978-3-948083-14-4. doi: 10.11588/heibooks.598. [Online]. Available: https://books.ub.uni-heidelberg.de/heibooks/reader/download/598/598-4-88224-1-10-20200325.pdf.

[10] K. Rechert and R. Gieschke, "Citar – preserving software-based research," *Int Journal of Digital Curation*, 1 2020. doi: 10.2218/ijdc.v15i1.716.

[11] K. Rechert, J. Oberhauser, and R. Gieschke, "How long can we build it? ensuring usability of a scientific code base," *Int Journal of Digital Curation*, vol. 16, no. 1, 2021.

[12] K. Rechert, T. Liebetraut, S. Kombrink, D. Wehrle, S. Mocken, and M. Rohland, "Preserving containers," *Forschungsdaten managen*, pp. 143–151, 2017.

[13] G. Peng, "The state of assessing data stewardship maturity–an overview," *Data science journal*, vol. 17, 2018. doi: 10.5334/dsj-2018-007.