Bausteine Forschungsdatenmanagement
Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanagerinnen und -managern

Bringing Neuroscientific Data to Sustainability

Embedded Data Stewardship in CRC 1280

Marlene Pacharra[i] Nina Olivia Caroline Winter[ii]
Tobias Otto[iii]

2024

Zitiervorschlag

Pacharra, Marlene, Winter, Nina Olivia Caroline, Otto, Tobias. 2024. Bringing Neuroscientific Data to Sustainability. Embedded Data Stewardship in CRC 1280. *Bausteine Forschungsdatenmanagement. Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanagerinnen und -managern* Nr. 2/2024: S. 01-12. DOI: 0.17192/bfdm.2024.2.8706.

[i]Ruhr University Bochum, Germany. ORCID: 0000-0001-6602-6746
[ii]Ruhr University Bochum, Germany. ORCID: 0000-0003-2966-4057
[iii]Ruhr University Bochum, Germany. ORCID: 0000-0002-9994-0910

Abstract

Information Management (INF) projects are essential to the establishment of infrastructures for the management of research data and metadata in Collaborative Research Centers (CRCs) funded by the German Research Foundation (DFG). However, the CRC's disciplines, the INF project's design, and the institutional setting all have a significant impact on the organization and daily responsibilities of INF data stewards. We present the measures and potential benefits of embedded data stewardship on research data management (RDM) in the case of CRC 1280 "Extinction Learning". Central aim of the RDM in CRC 1280 was the implementation of data sharing across all subprojects at an early stage of the research life cycle. Despite obstacles such as legal issues regarding sensitive data and a highly competitive culture in life sciences, a steady increase in the sharing of CRC-standardised metadata and data over the second funding period was achieved. This was accompanied by participatory strategy development, the implementation of supportive training formats and personal, on-site consultations by an embedded data steward with appropriate scientific background. In addition, the CRC's endeavours were led to sustainability by bridging the gap between researchers and central infrastructure: INF paved the way for the CRC to participate in the development of the university-central repository infrastructure, enabling the emergence of a sustainable RDM infrastructure tailored to the needs of the researchers.

Background: Challenges of research data management (RDM) in CRC 1280

Since 2017, psychologists, physicians, clinicians, computational neuroscientists, and biologists have investigated the mechanisms of extinction learning within the Collaborative Research Center (CRC) 1280 (speaker: Prof. Dr. Dr. h.c. Onur Güntürkün). Due to its strong interdisciplinary structure, CRC scientists produce heterogeneous research data, using a variety of measurement techniques such as microscopy, single cell recording, magnetic resonance imaging, or patient protocols. Research involves data from different species (mice, rats, pigeons, corvids, and humans) or simulated data from computational neuroscience. Moreover, it is conducted at four different institutions (Ruhr University Bochum, University of Duisburg-Essen, Leibniz Research Centre for Working Environment and Human Factors at TU Dortmund University, and University of Marburg) with local RDM policies, resulting in site-specific RDM workflows and storage strategies.

Legal and ethical concerns increase the hurdle for data collection, early data sharing and cross-group analysis in the CRC. Handling of data from human subjects must be planned and performed in line with requirements of the General Data Protection Regulation, which stipulates special care when dealing with sensitive data (e.g., patient protocols). Moreover, ethical concerns about exposing human subjects to stressful situations (social stressors, electrostimulation, medication) as well as conducting animal

research which includes e.g., hand-rearing corvids, understate the need for a sustainable use of the generated data. In addition, certain research cultures may discourage early data sharing (see Gallacher & Webster, 2024; Martone & Nakamura, 2022) especially in the life sciences, such as biology and medicine.

In this field of tension, the neuroscience community is actively discussing strategies to overcome the Reproducibility Crisis (Open Science Collaboration, 2015), which addresses concerns that published research results are all too often not reproducible, potentially undermining scientific progress. According to Klingner et al. (2023) professional data management has the potential of facilitating reproducible research e.g. by improving data documentation and standardisation: However, the same authors highlight a lack of RDM literacy as well as standards for data and metadata among neuroscientists. These challenges underscore the internal imperative and external expectations for robust data stewardship within CRC 1280.

First funding period

In the CRC's first funding period (2017–2021) two scientific projects were established to take advantage of data resulting from streamlined experimental designs. At the technical level, these so-called Focus Groups on *Learning Dynamics* and *Neuroimaging and Genetics* were designed to integrate internally shared data from research groups in the CRC to run large-scale analyses. At the same time, preliminary work to establish an Information Management (INF) project within the CRC was carried out. For this, a close collaboration between the CRC and the central Research Data Services (RDS) team of the CRC's central institution, the Ruhr University Bochum (RUB), was commenced.

A requirement analysis for RDM was initiated even prior to the CRC's first funding period in 2017, reflecting early commitment to the creation of findable, accessible, interoperable, and reusable (FAIR, Wilkinson et al., 2016) research data. As a first step, a needs assessment was carried out via an online survey consisting of 64 questions and covering all aspects of the data life cycle in the proposed CRC groups. By explicitly asking about staff responsible for RDM in the groups, the survey determined contact persons who then met regularly with the central RDS team. Using the survey as a basis for discussions, two main RDM actions for the CRC were identified and worked on: the implementation of a central storage strategy and the development of a CRC 1280 data model to increase the reusability of internally shared data in the CRC.

As a central storage platform for the CRC the file sharing service of the RUB was chosen featuring a fine-granular rights and role concept that allows individual control over users and data. Moreover, a folder structure representing a nested data model was developed (Pacharra et al., 2023). The designed data model shares important features such as hierarchical folder structure and inheritance strategy with the then emerging Brain Imaging Data Structure (BIDS), a standard for neuroimaging experiments (Gorgolewski et al., 2016), but to allow easy implementation prescribes fewer e.g., file na-

ming rules than BIDS (Diers et al., 2024). For metadata, the CRC's experts agreed on 16 fields as a necessary basis for cross-group communication. With the help of the central RDS team, standardised vocabularies for metadata fields were introduced and a mapping to Dublin Core terms and DataCite properties implemented (Zomorodpoosh et al., 2023).

Second funding period

The actual implementation of data sharing at an early stage of the research data life cycle needed additional awareness, support, and motivation. Therefore, in the second funding period (2021–2025), an INF project was established.

At the same time, RUB committed to establish a central, open-source repository infrastructure *ReSeeD* for the use of the whole university (Frenzel et al., 2023). CRC 1280 is one of the collaborative research projects on the RUB campus with the most diverse participation of research areas according to the DFG classification (DFG, 2021), comprising social and behavioural sciences, biology, and medicine. By representing these diverse research areas, the CRC had the potential to actively shape the requirements for an infrastructure ensuring that it meets the needs of researchers. Thus, CRC 1280 was chosen as the first operational use case for *ReSeeD*. In this context, RUB granted resources to implement the CRC's data model into *ReSeeD*. In return, the CRC decided to participate in the development process via active feedback and the provision of test data and test users. To embed this commitment in the CRC's strategy, the expertise of a Principal Investigator (PI) from Cognitive Psychology was supplemented by a PI from the central IT of the RUB for the management of the INF project.

The INF project was equipped with resources for one data steward, implementing an embedded data stewardship model (Neuroth et al., 2019). The formal appointment of a data steward created a central point of contact for the approx. 90 CRC scientists regarding RDM. Her main task was to implement the measures described below to promote a faster internal exchange of reusable data for even more efficient work in the CRC in the second funding period. For the position a person with a neuroscientific background could be employed by the CRC speaker's Biopsychology work unit at the main institution's Psychology Faculty. This has the advantage that the barrier for communication was low and the understanding of workflows and challenges for data management within the research groups was ideal.

Measures taken

Participatory processes and policy development

Data Stewardship activities at the start of the second funding period focused on RDM coordination processes and communication: To anchor RDM as a central strategic

measure in the CRC, a RDM Board was established comprising researchers from the various CRC disciplines, all status groups (early career researchers, PIs, support staff), the Focus Groups, and all CRC institutions. As a first step, the board developed a draft RDM policy between April and September 2022, which was unanimously approved by the CRC's PIs. The RDM policy (Pacharra et al., 2022) refers to a dynamic internal knowledge base that can be adapted to changing RDM needs. The policy defines the roles of the INF, Focus, and working groups in the CRC, including their tasks and responsibilities, which increases transparency regarding INF governance.

Internal knowledge base

Adherence to the CRC's policy means adapting site-specific workflows and changing the local RDM culture – a process that can be time-consuming and often lacks clear, short-term incentives. To avoid frustration, it has paid off to develop and implement low-threshold formats to broadcast and ingrain RDM knowledge, such as the CRC's knowledge base.

The internal CRC's knowledge base was set up in collaboration between INF, the CRC's central coordination project (Z) and Focus Groups to enable quick orientation: in detail, its aim is to provide all information for planning a new CRC study, for searching for CRC services, as well as documented and downloadable tools. The knowledge base is not limited to RDM (e.g. self-learning materials, best-practices, pipelines, scripts, glossary of available services and tools) but presents RDM as part of the CRC research process. For its technical implementation, the RUB Moodle web-platform is used: this tool allows a quick implementation, is familiar to many researchers, and allows access from different institutions. The CRC 1280 group on RUB's GitLab (https://gitlab.ruhr-uni-bochum.de/sfb1280) and the CRC's Zenodo community (https://zenodo.org/communities/sfb1280/) are used for the external dissemination of materials and resources.

Consultancy

The INF project's main focus is to offer consultancy services centred around group-specific data workflows. The primary objective is to foster researchers' comprehension of the collaborative data management workflow in the CRC and enable effective integration of their own data workflow with that of the CRC.

Researcher's trust in RDM processes is key to a cultural change towards early data sharing. Therefore, consultations were conducted on-site in person, where possible. Bochum, Dortmund, and Essen, which together comprise 98% of the CRC researchers, are each only a 30-minute drive away from each other. This proximity facilitates on-site consultation and the implementation of hands-on RDM support and workflow analyses, which are the main focal points of INF consultations (see Figure 1). Researchers

from Marburg were integrated through regular visits to the Ruhr area for INF consultations, as well as virtual consultations.

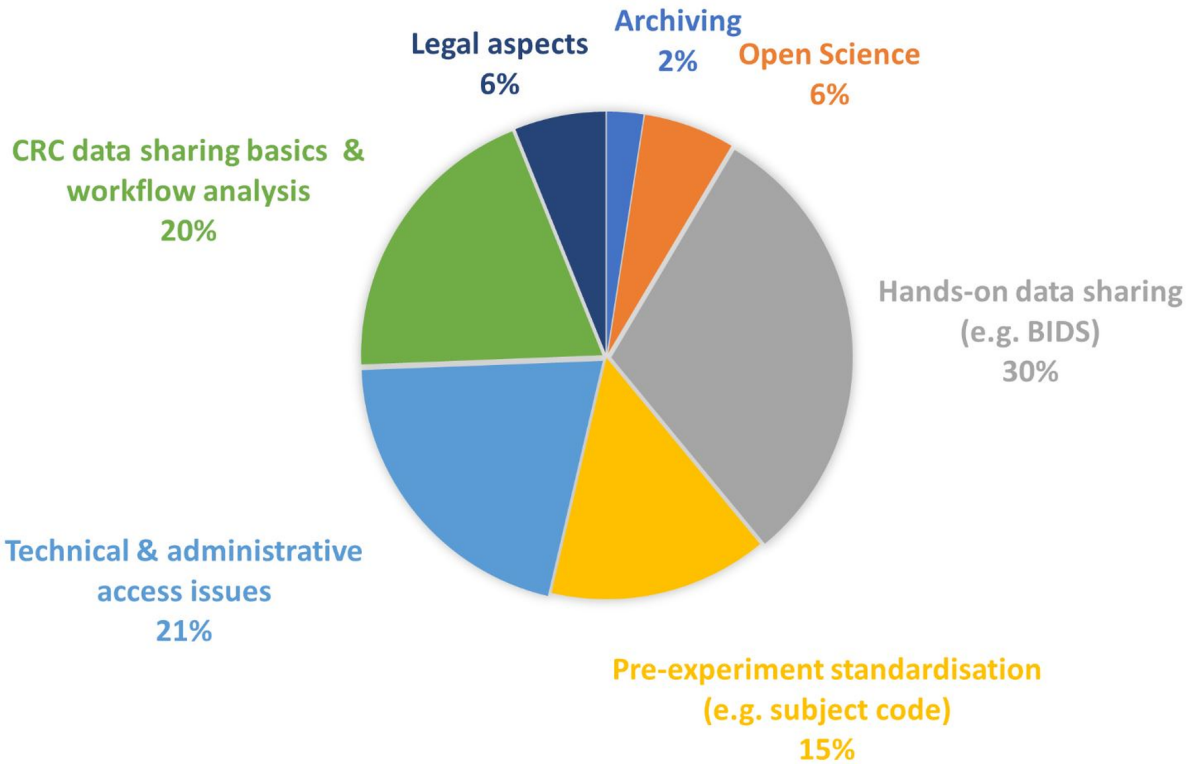

Figure 1: Thematic focus of the consultations (n = 82) with INF data steward (10/2021 - 03/2024).

Consultations also play a crucial role in establishing best practices and identifying weaknesses in current RDM, often leading to tool development by the INF project:

- Java-based applications for cross-platform use to automate tasks such as generating CRC folder structures or extracting CRC metadata from existing BIDS metadata (see https://gitlab.ruhr-uni-bochum.de/sfb1280)
- An open-source subject code generator (Diers et al., 2023) to facilitate the time- and purpose-limited pseudonymization of human subjects as recommended by the Ethics Committee of the German Psychological Society
- Functionalities for an established open-source analysis app to export raw data of electrodermal activity directly to BIDS (Otto et al., 2023)

These tools save time, make policy-compliant RDM more efficient and thus aid their widespread adoption. Nevertheless, dedicated training is necessary to raise awareness and avoid frustration when using any RDM tool.

Training and awareness

From October 2021 to March 2024, the INF project hosted four half-day virtual workshops, categorised thematically into two main groups. Firstly, general courses aimed to raise awareness of good data management and provide universally applicable methods and tools to improve data quality in line with the FAIR principles. These courses, including “Introduction to Research Data Management” and “Introduction to git and GitLab” (a versioning tool and web-based repository manager based on it) drew upon existing training modules and the expertise of the central RDS team. Secondly, discipline-specific training courses were organised to demonstrate how data management could be integrated into their daily research processes. Collaborating with external experts, INF organised workshops on BIDS and DataLad (Halchenko et al., 2021). The workshops were evaluated by the 11 to 27 participants using standardised online surveys. These surveys assessed the quality of the workshops, identified further desired training, and highlighted additional RDM needs. Based on this evaluation as well as ongoing consultancy with scientists, new training sessions and formats were developed (e.g. Christmas data management workshop series).

Recognizing potential motivational barriers associated with traditional workshop formats, INF also introduced a novel approach – a 4-part Christmas data management workshop series in 2021 and 2023. This series, spread over four weeks with one-hour sessions each week, served as a platform for brief recaps of previous workshops, presentations of INF-developed RDM tools, and discussions on CRC standards.

As an integral part of the Open Science Task Force within the CRC, the INF project actively promotes Open Science practices. This effort includes supporting workshops conducted by external trainers in the CRC on topics such as preregistrations (Nosek et al., 2018) and enhancing validity and transparency in animal studies (Diederich et al., 2022). In line with this, INF provided initial consultancy to early career researchers within the CRC in establishing their own *ReproducibilTea* Journal Club (Orben, 2019) for Open Science. Additionally, INF collaborates with RDS to conduct workshops during the RUB’s Open Week.

Data Cleaning Days

To keep the motivation hurdle for taking up INF support as low as possible, the INF project organises Data Cleaning Days, which include both training and consultancy, at least once a year. On these days, all CRC researchers are asked to stop their experiments and focus solely on RDM (e.g., back-ups, creating metadata). The special event character is highlighted by on-site visits by the data steward.

During these visits, researchers share the results of their RDM and get tailored RDM training in individual or group sessions (max. 5 researchers). In addition to the normal thematic foci of consultations (see Figure 1), the Data Cleaning Days are used particularly to share new data and metadata with the CRC (see Figure 2), to receive immediate

feedback about the quality of shared data and metadata, and to get practical help in the process. In feedback loops with the researchers, INF must curate the newly uploaded data and metadata to ensure compliance to CRC standards and the CRC's RDM policy (Pacharra et al., 2022) for example by checking folder structure and metadata using open-source CRC tools such as the Study Checker (Diers et al., 2024) and MetaDatabaseAnalyzer (Pacharra, 2024). This can also lead to (meta-)data cleansing (see Figure 2).

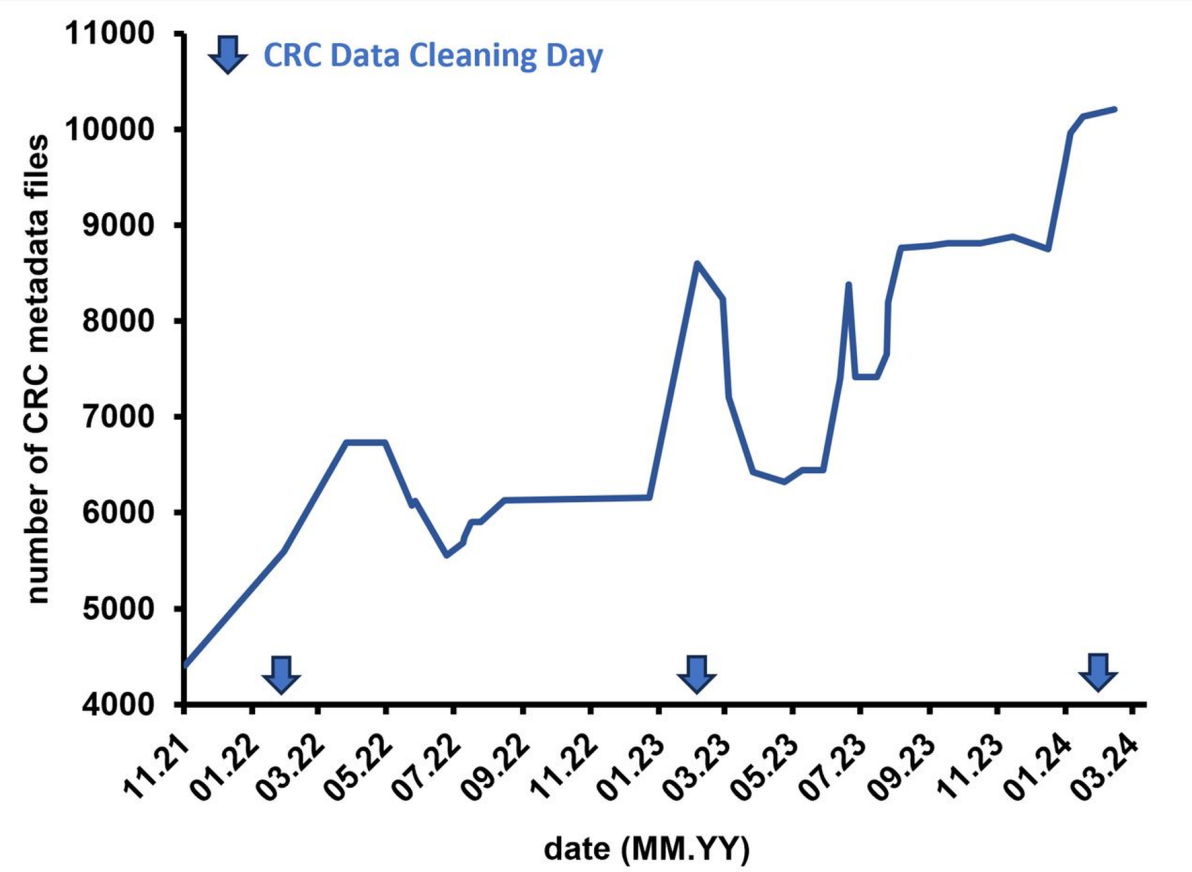

Figure 2: Number of CRC metadata files on the central file share (11/2021 - 03/2024) with initial increase and subsequent metadata cleaning after a CRC Data Cleaning Day.

Sustainable workflows in RDM infrastructure

INF facilitates the continuous collection and analysis of specific RDM requirements and was thereby able to refine the specifications for the new central repository infrastructure *ReSeeD*. Innovative features in *ReSeeD*'s developments arose from CRC feedback:

- Hierarchical data storage reflecting CRC folder structure

- Daily and bulk ingest workflows
- Review workflow involving central university staff and peer-review by researchers (Pacharra et al., 2024), which will further formalise the current curation workflow

Moreover, INF selects CRC test data for *ReSeeD* and supports beta testing. By coordinating beta tests with researchers and analysing user feedback from the CRC, INF is helping to bridge the gap between central infrastructure and scientific workflows (Pacharra et al. 2024) and drive the development of a sustainable RDM infrastructure tailored to the needs of researchers.

Conclusion

Embedded data stewardship within CRC 1280 demonstrates opportunities to foster sustainable RDM across diverse scientific disciplines. The use case emphasises measures aimed at bridging cultural barriers between research and data management to improve understanding and communication between researchers and infrastructure.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 316803389 SFB 1280 (project INF). We thank Sandra Linn for proofreading and helpful comments.

References

DFG (2021). *DFG classification of subject areas and review boards (2020-2024)*. Accessed April 23, 2024 from https://www.dfg.de/resource/blob/175336/46d0a8c5c14d0f6530da28b4d25bf589/fachsystematik-2020-2024-en-grafik-data.pdf

Diederich, K., Schmitt, K., Schwedhelm, P., Bert, B., & Heinl, C. (2022). A guide to open science practices for animal research. *PLoS biology*, *20*(9), e3001810. https://doi.org/10.1371/journal.pbio.3001810

Diers, E., Pacharra, M., Merz, C. J., Ernst, T. M., & Otto, T. (2023). *Subject Code Generator (v1.1)*. Zenodo. https://doi.org/10.5281/zenodo.7634563

Diers, E., Diers, O., Pacharra, M., & Otto, T. (2024). *Study Checker*. Zenodo. https://doi.org/10.5281/zenodo.13968793

Frenzel, J., Esser, A., Pacharra, M., Otto, T., Schramm, A., Barthauer, R., & Winter, N. O. C. (2023). *Making data management feel easy: Integration of a Hyrax data repository into the research process* [Poster]. 1st Conference on Research Data Infrastructure (CoRDI 2023), Karlsruhe, Germany. Zenodo. https://doi.org/10.5281/zenodo.8343752

Gallacher, J., & Webster, C. (2024). We must discuss research environments. *Royal Society Open Science, 11,* 231742, https://doi.org/10.1098/rsos.231742

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., Poline, J. B., … Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, *3*, 160044. https://doi.org/10.1038/sdata.2016.44

Halchenko, Y. O., Meyer, K., Poldrack, B., Solanky, D. S., Wagner, A. S., Gors, J., MacFarlane, D., Pustina, D., Sochat, V., Ghosh, S. S., Mönch, C., Markiewicz, C. J., Waite, L., Shlyakhter, I., de la Vega, A., Hayashi, S., Häusler, C. O., Poline, J.-B., Kadelka, T., … Hanke, M. (2021). DataLad: distributed system for joint management of code, data, and their relationship. *Journal of Open Source Software*, *6*(63), 3262. https://doi.org/10.21105/joss.03262

Klingner, C. M., Denker, M., Grün, S., Hanke, M., Oeltze-Jafra, S., Ohl, F. W., Radny, J., Rotter, S., Scherberger, H., Stein, A., Wachtler, T., Witte, O. W., & Ritter, P. (2023). Research data management and data sharing for reproducible research-results of a community survey of the German national research data infrastructure initiative neuroscience. *eNeuro*, *10*(2), ENEURO.0215-22.2023. https://doi.org/10.1523/ENEURO.0215-22.2023

Martone, M. E., & Nakamura, R. (2022). Changing the culture on data management and sharing: Overview and highlights from a workshop held by the National Academies of

Sciences, Engineering, and Medicine. *Harvard Data Science Review, 4(3)*. https://doi.org/10.1162/99608f92.44975b62

Neuroth, H., Rothfritz, L., Schmunk, S., Schrade, T., & Rapp, A. (2019). *Embedded Data Stewardship: A community-driven agile self-assessment framework for monitoring and improving the quality of research data management* [Poster]. 14th International Digital Curation Conference. Accessed April 23, 2024, from https://www.dcc.ac.uk/events/idcc19/posters

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science (New York, N.Y.)*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Orben A. (2019). A journal club to fix science. *Nature*, *573*(7775), 465. https://doi.org/10.1038/d41586-019-02842-8

Otto, T., Wolf, O. T., & Merz, C. J. (2023). *EDA-Analysis App* (5.13). Zenodo. https://doi.org/10.5281/zenodo.8407428

Pacharra, M. (2024). *MetaDatabaseAnalyzer.* Zenodo. https://doi.org/10.5281/zenodo.13912364

Pacharra, M., Frenzel, J., Schramm, A., & Otto, T. (2023). *Data stewardship in CRC 1280 "Extinction Learning": From policy to dedicated workflows in an institutional data management system* [Presentation]. Data Stewardship goes Germany 2023 Workshop (DSgG 2023), Dresden. Zenodo. https://doi.org/10.5281/zenodo.8388876

Pacharra, M., Frenzel, J., Winter, N. O. C., & Otto, T. (2024). *Towards tailored data curation workflows in a trusted repository: Strategies in a collaborative research centre in neuroscience* [Poster]. 18th International Digital Curation Conference (IDCC24), Edinburgh, Scotland. Zenodo. https://doi.org/10.5281/zenodo.10776740

Pacharra, M., Winter, N. O. C., Kumsta, R., Uengoer, M., Caviola, J. K., Ernst, T. M., Reichert, R., Yavari, F., Merz, C. J., Cheng, S., Linn, S., Wolf, O. T., Güntürkün, O., & Otto, T. (2022). *Research data management policy of the collaborative research centre SFB 1280 "Extinction Learning"* (v1.0). Zenodo. https://doi.org/10.5281/zenodo.8004432

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

Zomorodpoosh, S., Diers, E., Linn, S., Merz, C. J., Pacharra, M., & Otto, T. (2023). *MetaDataApp* (v2.0). Zenodo. https://doi.org/10.5281/zenodo.8040229