Data Stewards as ambassadors between the NFDI and the community

Dirk von Suchodoletz[1], Timo Mühlhaus[2], Dominik Brilhaus[3], Hajira Jabeen[4], Björn Usadel[5], Jens Krüger[6], Holger Gauza[7], and Cristina Martins Rodrigues[8]

[1]0000-0002-4382-5104
[2]0000-0003-3925-6778
[3]0000-0001-9021-3197
[4]0000-0003-1476-2121
[5]0000-0003-0921-8041
[6]0000-0002-2636-3163
[7]0000-0003-0191-3680
[8]0000-0002-4849-1537

The NFDI consortium DataPLANT focusing on fundamental plant research, provides data stewards as a core element of its strategy for dissemination of common standards, concepts of research data management, and workflow services. Data stewards play a special hinge role between service providers, individual researchers, groups, and the wider community. They help to bridge the gap between the scientists working in the lab and the technical solutions and services. Project groups and individual researchers will profit from direct support in their daily tasks ranging from data organization to the selection and continuous development of the proper tools, workflows and standards. This leads to a community-wide dissemination and development of data management strategies especially suited to support plant research. In particular, the convergence of researcher and repository requirements is of great importance, and crucial for the success of RDM in general. Additionally, the data steward service concept of DataPLANT is designed for effective capacity building and training to ensure sustainability in the research landscape.

1 Motivation – What is a data steward?

The slow adoption and dissemination of common standards, the concepts of research data management, and workflow services is still a hindrance to collaboration, data sharing-and-reuse, as well as open science in many scientific communities [1, 2]. The responsible and informed

handling of research data is part of good scientific practice [3, 4]. The central goals of DataPLANT [5, 6] are, to provide appropriate infrastructure and workflows, and to train researchers of varying experience towards data stewardship and research data management (RDM). In the long run, such qualification measures should be included in the relevant curricula. The task for the support and community domain of the project is to prepare tailored content for the various data management mechanisms over the entire lifecycle.

Hence, data stewards are experienced individuals with strong communication skills, expertise in plant biology, bioinformatics tool development and familiar with heterogeneous infrastructure. Data stewards operate at the core of DataPLANT and fulfill a special hinge role between the various stakeholders and the wider community to bridge the gap between researchers and technical infrastructure (see Figure 1).

DataPLANT introduces a community-integrative approach of data stewardship that is supported by internally governed and associated data stewards with aligned functions. Internally governed data stewards are funded and orchestrated by the NFDI consortium itself. With a focus on DataPLANT's core mission, they support multiple consortia and individual research groups. This allows the DataPLANT consortium to provide on-site support for the individual project partners and participants either in person or remotely. Associated data stewards are funded by and seated at DataPLANT project partners such as collaborative research centres, typically familiar with local scientific workflows and RDM practices. The common goal of data stewards is to integrate institutional and community RDM concepts as well as aligning the standards in the domain and infrastructural support environments both on a practical and operational level [7]. This bidirectional communication fosters to interlink RDM activities within the community.

2 Contribution to the community

Data stewards target the community on different levels and provide specifically tailored data management strategies that enable the community to use existing standards and facilitate the use of technology and infrastructure for data management [8]. Through the community-integrative model, they interact directly with core facilities, research groups and individual researchers. As the major (*omics) data providers, core facilities play a special role in the development and dissemination of DataPLANT. They are experts in measurement technologies that are central to the community and know most about method-specific metadata and infrastructure requirements. Due to their community network and diverse client base, they take a multiplier role, allowing an indirect reach out to participants, plus possible links to other scientific communities and NFDI consortia. Data stewardship of core facilities thus has a manifold effect by finding an RDM solution that suits the facility and improving user-friendliness for clients who use the same DataPLANT mechanisms established in other facilities. Research groups profit from data stewards in multiple ways. Data stewards advise on data management and standards related questions of a grant application or during the setup phase of a research project. Project managers and principal investigators can request information on the ongoing activities in standards development. In addition, data stewards offer proven and well established procedures to handle research data aiming at the improvement of digital lab organisation according to the FAIR data principles [9].

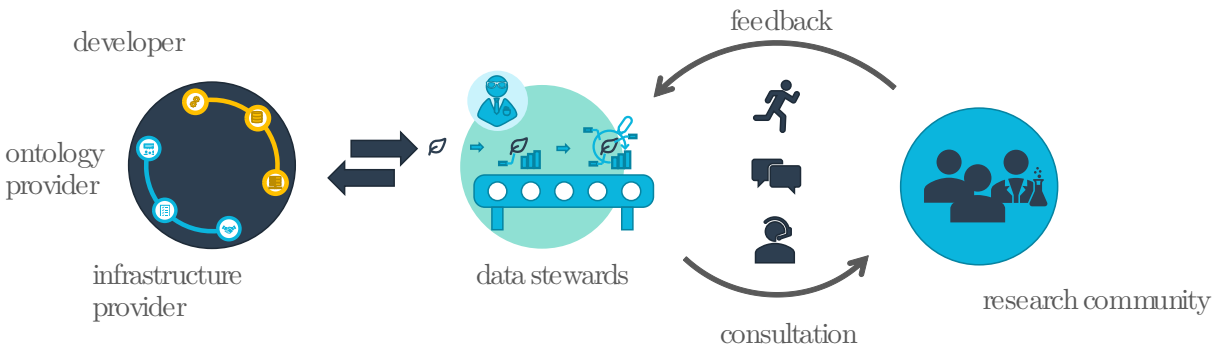

Figure 1: The hinge role of DataStewards between the community and infrastructure

2.1 Dissemination and development of data management strategies

A holistic planning phase including a data management plan (DMP) is a prerequisite of a successful grant application and project start. Together with the participants, data stewards develop a plan fitting their project requirements. 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. DataPLANT employs a data-centric approach towards FAIRness of plant biological data. At the heart of this approach lies the ARC (annotated research context) [10] as the data packaging format for research objects, which expands the widely established metadata grammar of ISA [11] to enrich the ARC with content and provides further context e.g. on the workflows and tools used. Its flexible and open nature guarantees long-term accessibility and sustainability. The central DataPLANT mechanisms of data stewardship and data management planning evolve around the ARC environment, accessible directly or through the DataPLANT Hub [5]. Data stewards help developing the ARC environment to offer a common suite of suitable data formats, standards, and repositories for an increasing range of data types and integrate associated tools and workflows for data processing and publication. These developments are elaborated in the DMP and enable the community to use the DataPLANT technologies and infrastructures and facilitate data publication in community-specific repositories.

2.2 Converging researcher and repository requirements

As the sustainability of DataPLANT depends on the convergence between its data-centric approach and the current state of the individual plant science communities, data stewards participate in implementing suitable operating procedures into the participant groups. Proper metadata description is the basis for data findability and accessibility. Data stewards support a structured collection of metadata for common experimental and computational workflows by drafting metadata templates and guiding participants on creating templates or adapting existing ones to their needs. They foster compliance with the submission requirements of end-point repositories and associated metadata standards and minimal reporting guidelines. This ensures that metadata is (readily) usable independent of DataPLANT services. To facilitate the collection of metadata at its point of emergence, data stewards support the FAIRification of the whole scientific process – from experiment planning to data acquisition and processing. The light-weight standardization convention of the ARC environment can easily be adapted to or implemented into daily laboratory routines. Data stewards help the participants to develop suit-

able solutions for data storage and sharing, for the lab organisation or to adapt local software packages. Through the development of digital workflows such as Galaxy [12] and Nextflow [13], they enable access to necessary infrastructures and harness remote resources. Data FAIRness and preparation of high quality ARCs for sharing and publication is assured by active participation of data stewards during the iterative cycles of metadata annotation and data handling. The development of ARCs is a bidirectional, iterative effort. Data stewards continuously monitor and evaluate participant feedback on tools and services. This process of incorporating case-by-case specific requirements into a widely adoptable consensus, shapes ARC's flexibility and the further route of development of tools and services. Retracing participant input and adaptations will propel the development of the ARC environment and facilitates to address frequently missed information in metadata templates, fragmentary ontologies, and existing standards. Furthermore, the direct and timely interaction with the active research community enables the flexible integration of future developments, including new techniques and data types.

3 Capacity building

Significant dissemination to the community is achieved through a comprehensive training program that introduces DataPLANT services and tools as well as general data literacy and analysis capabilities to the researcher [14]. Individual consultation of participants will be complemented with on-site workshops for research groups adapted to the needs of the community and the stage of association with DataPLANT. During the onboarding phase, the activities cover general data management practices and familiarization with DataPLANT tools and services. In-depth expertise on specific topics is elaborated with respective stakeholders in the participating groups. For a continuous exchange between the data stewards and the research groups, DataPLANT encourages the appointment of data management representatives (DMRs), who – similar to core facilities – act as relevant multipliers. They take a bidirectional role by (i) spreading knowledge on data management, standards and services in their groups and (ii) reporting back common hurdles and requirements. Both DMRs and core facility heads will specifically be addressed and qualified by DataPLANT data stewards. In addition to workshops, a continuously updated knowledge base provides teaching materials, tutorials for tools, services and best practices that reflect the development of DataPLANT. The ultimate goal of DataPLANT is to enable the researcher to produce ARCs without or only minimal support by the data steward. Training is not exclusive to participants, but likewise enables the continuous qualification of data stewards ("train the trainer"). Data stewards attend training and workshops to keep track of all relevant developments in the field as well as international activities and achievements. In regular meetings and through a central data stewardship knowledge base, data stewards exchange on best-practices, qualify on new standards, learn on legal issues, updates on extended modified ontologies and metadata schemas as well as on potential new workflow and software options. FAIRification use cases at the participants' sites are shaped into general best practices and common data stewardship tasks. This rich support resource will particularly be useful to freshly onboarding data stewards, but may also be transferred into the plant science or NFDI community to set new standards for data stewardship in general. Besides disseminating DataPLANT mechanisms, the data stewards consulting and qualification capacities need to be extended over time. This challenge to personnel development is shared with other consortia in the NFDI as well and addressed through cross-cutting activities [15].

4 Data steward dispatch model

Substantial data stewardship time is allocated to consulting services and capacity building, in addition to self-qualification and dissemination. Data steward support can be requested in any stage of the research process. The group of data stewards maintains connections with the community as they accompany scientists and research groups in the various stages of the research data life cycle. Until the data stewardship is institutionalized, we follow a distribution model to optimize leveraging effects in the community. Therefore, efficient scheduling of resources suggests focusing the support on data generating hubs within the community. However, in order to follow the consortium's objectives of transparent communication and broad user involvement, a balanced mechanism that ensures fair allocation of resources is envisioned with the following dispatch model:

1. First time request is (automatically) granted but goes with conditions (e.g commitment to the NFDI objectives, provisioning of the data to the NFDI).

2. FairShare: Available data stewards hours are divided by the number of requests. Additionally, 30% are reserved for future requests.

3. Later, the allocation could take input parameters like the size of a research group, the provision of additional resources (e.g grant money, material costs of their accepted grant) and bonus points.

4. The bonus points are allocated to groups or individuals after quality assessment of the provided data, and these points can be translated into additional hours or resource allocation using an evaluation system.

5. In the future extra points may be awarded for exemplary data sets published and referenced.

6. During phases of higher loads, the multiple incoming requests can be ordered by waiting time. Groups which interacted more recently with a data steward will wait comparably longer than researchers who used their services a longer time ago. A weighted queue can be maintained for high load, less resource time-period.

The preliminary strategy combines factors of fair distribution of resources with incentive schemes to improve the metadata quality and FAIRness of data sets. Given that it is challenging to know the demand in advance, it is anticipated that this set of rules will be further polished and adjusted according to the existing resources and data management demands from the community. Special requests, conflicts which are not solvable on that layer will be passed on to the Senior Management Board to decide. Additionally, this body takes steering responsibility to adapt the distribution if necessary, after a ramp-up period followed by an evaluation of the process. We assume a rising demand from the wider community.

5 Sustainability and outlook

To foster a broader adaptation of DataPLANT within the community and to grow with the demand for new participants, data stewardship should be complemented by co-funding or own personnel of new members. If a broad range of future individual project proposals or large-scale projects like collaborative research centres plan for personnel and infrastructure services directly

by contributing to the NFDI, a sustainable financing and reimbursement model can be created benefiting the broader community. Small projects can then receive qualified support from a range of experts according to their contribution. Data stewards in large projects get integrated into a broadly qualified team working on cutting-edge research and workflows. The consortium's and NFDI's governance structures ensure the orientation of the data stewards' support on the actual demands of the community.

Acknowledgements

CEPLAS has been supported by Deutsche Forschungsgemeinschaft within the Excellence Initiative (EXC 1028) and under Germany's Excellence Strategy – EXC 2048/1 – project 390686111. We acknowledge support for DataPLANT 442077441 through the German National Research Data Initiative (NFDI 7/1) and the Science Data Center BioDATEN which is supported by the Ministry of Science, Research and Art Baden-Württemberg.

References

[1] S. Rosenbaum, "Data governance and stewardship: designing data stewardship entities and advancing data access," *Health services research*, vol. 45, no. 5p2, pp. 1442–1455, 2010.

[2] G. Peng, "The state of assessing data stewardship maturity–an overview," *Data science journal*, vol. 17, 2018.

[3] Deutsche Forschungsgemeinschaft, "DFG guidelines on the handling of research data," https://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/guidelines_research_data.pdf, 2015, [Online; accessed 28-April-2021].

[4] "Guidelines for Safeguarding Good Research Practice. Code of Conduct," Sep. 2019, available in German and in English. [Online]. Available: https://doi.org/10.5281/zenodo.3923602

[5] "DataPLANT NFDI webpage," https://nfdi4plants.de/, [Online; accessed 16-April-2021].

[6] D. von Suchodoletz, T. Mühlhaus, J. Krüger, B. Usadel, and C. Rodriges, "Dataplant — ein nfdi-konsortium der pflanzen-grundlagenforschung," 2021.

[7] D. Iglezakis and S. Hermann, "4.4 disziplinspezifische und – konvergente fdm-projekte," in *Praxishandbuch Forschungsdatenmanagement*. De Gruyter Saur, 2021, pp. 381–398.

[8] D. Hausen, J. Rosenberg, U. Trautwein-Bruns, and A. Schwarz, "Data stewards an der rwth aachen university–aufbau eines flexiblen netzwerks," *Bausteine Forschungsdatenmanagement*, no. 2, pp. 20–28, 2020.

[9] 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.

[10] C. Garth, J. Lukasczyk, T. Mühlhaus, B. Venn, , K. Glogowski, C. M. Rodgrigues, and D. von Suchodoletz, "Immutable yet evolving: ARCs for permanent sharing in the research data-time continuum," 2021.

[11] S.-A. Sansone, P. Rocca-Serra, D. Field, E. Maguire, C. Taylor, O. Hofmann, H. Fang, S. Neumann, W. Tong, L. Amaral-Zettler *et al.*, "Toward interoperable bioscience data," *Nature genetics*, vol. 44, no. 2, pp. 121–126, 2012.

[12] J. Boekel, J. M. Chilton, I. R. Cooke, P. L. Horvatovich, P. D. Jagtap, L. Käll, J. Lehtiö, P. Lukasse, P. D. Moerland, and T. J. Griffin, "Multi-omic data analysis using galaxy," *Nature biotechnology*, vol. 33, no. 2, pp. 137–139, 2015.

[13] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame, "Nextflow enables reproducible computational workflows," *Nature biotechnology*, vol. 35, no. 4, pp. 316–319, 2017.

[14] S. Jones, R. Pergl, R. Hooft, T. Miksa, R. Samors, J. Ungvari, R. I. Davis, and T. Lee, "Data management planning: How requirements and solutions are beginning to converge," *Data Intelligence*, vol. 2, no. 1-2, pp. 208–219, 2020.

[15] F. O. Glöckner, A. Pollex-Krüger, K. Toralf, J. Fluck, B. König-Ries, C. Eberl, T. Schrade, A. Güntsch, B. Gemeinholzer, T. Schörner-Sadenius *et al.*, "Berlin declaration on nfdi cross-cutting topics," Jülich Supercomputing Center, Tech. Rep., 2019.