FAIR Connect 1 (2023) 1–3
DOI 10.3233/FC-221508
IOS Press

Data Stewardship Plan templates designed to support the FAIR principles

Erik Schultes

FAIR Implementation Lead, GO FAIR Foundation, Poortgebouw Noord, Rijnsburgerweg 10, 2333 AA Leiden, The Netherlands
E-mail: eriks@gofair.foundation; Tel.: +310642448027; ORCID: https://orcid.org/0000-0001-8888-635X

Editor: Barbara Magagna (https://orcid.org/0000-0003-2195-3997)
Solicited reviews: Mijke Jetten (https://orcid.org/0000-0001-9114-2896); Valentina Pasquale (https://orcid.org/0000-0002-4499-9536)

Received 15 November 2022
Accepted 29 November 2022

Abstract. Data Stewardship Plan (DSP) templates prompt users to consider various issues but typically have no requirements for actual implementation choices. But as FAIR methodologies mature the DSP will become a more directive "how to" manual for making data FAIR.

Keywords: Data management plan, data stewardship activity, fair principle, knowledge of theories underlying fair implementation, fair evaluation of repositories for data deposition

For many years, Data Management Plans (DMPs) and the more broadly construed Data Stewardship Plans (DSPs) have been a key element in helping the researcher and their organization to maximize the value of the data they create (for simplicity, throughout this commentary we will use only the broader term DSP). There are two trends in FAIR approaches to data stewardship that are beginning to make an impact on the creation and use of DSPs. As FAIR methodologies mature, so too can the DSP as a practical "how to" manual, assisting the day-to-day work of the data steward and at the same time raising FAIR awareness among researchers.

First, concrete and specific FAIR methods and implementation choices made at a community level, can be embedded in DSP templates, making the DSP instances more directive. An example is the use of domain-specific identifiers, as explained below, clarifying immediately what is expected from the researcher-data steward team, regarding FAIR vocabularies. Up to now, DSP templates have typically been soft questionnaires, prompting the data steward/researcher to "consider" various (often disconnected) topics related to data stewardship [1]. Here, "soft" and "consider" imply that there has been little expectation (and often no requirement) for actual decisions. As such, the completed DSP instance is sometimes seen from the researcher's point of view as a purely bureaucratic exercise, and once completed, is rarely if ever again consulted during the course of the project. This perfunctory status given to data stewardship planning is a reflection of the largely unstructured and nonsystematic practices of data stewardship. Without broadly agreed upon solutions (i.e., standards), there is no possibility to expect definitive answers or to objectively evaluate them.

When there is a focus on FAIR, for which increasingly mature solutions and expectations are emerging, the DSP can become more definitive, explicit and directive. We no longer ask data stewards and researchers to consider issues but to follow a set of standards that could have been formulated elsewhere, by a trusted authority (such as requirements defined by the funding agency or recommendations offered by domain communities or national organizations). For example, the vague admonition “consider using controlled vocabularies” often seen in DSP templates can, when focusing on FAIR Principle I2 become a more decisive activity such as “use IUPAC-endorsed identifiers when describing chemicals, use UniProt identifiers when describing human proteins, use identifiers from the Global Biodiversity Information Facility when referring to biological species”. Of course, a directive like this is meant to support and expedite the work of the data steward, relieving the data steward from having to make a potentially very large number of (mostly uncontentious) decisions in isolation. The trusted authorities offering these directives may also provide some justification or guidance in the form of assessments on the intrinsic FAIRness of these resources, their open accessibility, ease of use, etc. This more directive approach of the DSP is not intended to second-guess the knowledge or skill of the data steward, and if the data steward should feel the need to deviate from the directives set in the DSP, they are then encouraged to do so with an explanation of why the alternative is, in this case, a better-fitted solution.

Second, with the emergence of research data “repositories” in the last decade, good data stewardship is often seen as the answer to the question: “in which repository will you deposit your data?” Indeed, the language of centralized repositories is the default in contemporary DSP questionnaires. Certainly, this is an improvement on the common practice of storing data on local devices with little or no metadata or public endpoints. But, as adherence to the FAIR Principles becomes a more advanced and routine expectation (especially from funders), there is often the assertion in DSPs that data will be made FAIR by simply depositing them in repositories. After all, there you will find a place for metadata, the data and upon a successful deposition, receive a persistent link such as a Digital Object Identifier (DOI). However, as we know from studies like that conducted by EOSC Nordic, data repositories currently have in general limited FAIR capabilities [5]. For example, centralized data repositories often use no controlled domain-specific vocabularies in metadata records, and they often have limited access control settings which are of vital importance for sensitive data such as surveys or patient medical records.

In contrast, emerging FAIR technologies opens the door to solutions that make use of domain-relevant community standards (including controlled vocabularies, metadata schemas and data models), customized ontology-based access controls, and inherently decentralized architectures (where centralized repositories may co-exist among many other kinds of resources that store and serve data). Applications like FAIR Data Points can provide thin proxy-layers on top of existing centralized resources, boosting their FAIRness without expensive re-tooling [2,4]. FAIR data networks like this support global search on data assets (i.e., extending between established repositories) [7]. FAIR Data Stations can support distributed learning environments where analytical algorithms are sent to the data, rather than the other way around (making it easier to include sensitive data in data analyses) [3,6,8]. As these types of solutions mature and become more reusable, they can be made part of the aspirations reflected in the DSPs and included as resources to be used or otherwise deployed by data stewards. It has always been anticipated that as FAIR technologies become more readily available, achieving high-levels of FAIR will become easier for researchers and data stewards. In the future, FAIRification will become routine, with the technical aspects moving under the hood of commonly used software applications and methods. For the time being, DSP templates should be crafted with FAIR solutions explicitly in mind, with directives that are as concrete as possible, and removing where possible the uncertainty and guesswork from the practice of data stewardship.

In summary, more and more we look to trusted authorities like funders, knowledge sharing institutions (e.g., the Dutch Health-RI), and disciplinary RIs (e.g., for life science in Europe, ELIXIR) for recommendations on emerging FAIR infrastructure technologies. Such organizations, having the resources to vet and (when necessary) build the appropriate FAIR enabling resources, should take a leading role in disseminating these technology choices in well-documented DSP templates. These organizations will help data producers to move away from the zoo of vague and non-committal DSP templates to more definitive “blue-prints” that can guide actual data production.

Acknowledgements

Many thanks to Rob Stierum for constructive suggestions and comments.

Funding

The author reports no funding.

Disclosures and conflict of interest

Erik Schultes was not involved in the peer-review process.

References

[1] Horizon 2020 Data Management Plan Template [Internet]. [cited 2022 Nov 15]. Available at: https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm.

[2] O.M. Benhamed, K. Burger, R. Kaliyaperumal, L.O.B. da Silva Santos, M. Suchánek, J. Slifka et al., The FAIR Data Point: Interfaces and Tooling, *Data Intell.* (2022), 1–18.

[3] Citizen Centred and Controlled COVID-19 Data for Reuse (C4 Yourself) [Internet]. [cited 2022 Nov 15]. Available at: https://www.health-holland.com/project/2021/2021/citizen-centred-and-controlled-covid-19-data-reuse.

[4] L.O.B. da Silva Santos, K. Burger, R. Kaliyaperumal and M.D. Wilkinson, FAIR Data Point: A FAIR-Oriented Approach for Metadata Publication, *Data Intell.* (2022), 1–21.

[5] B. Meerman, S. Pittonet Gaiarin and S. Muradore Gallas, EOSC-NORDIC FAIRification study testing F-UJI [Internet], 2021 [cited 2022 Nov 15]. doi:10.5281/zenodo.5533896.

[6] S. Österle, See the Swiss Personal Health Network (SPHN) Data Ecosystem for FAIR Data, [cited 2022 Nov 15]; Available at: https://www.youtube.com/watch?v=pqV0qp4oisM.

[7] M. van Reisen, F. Oladipo, M. Stokmans, M. Mpezamihgo, S. Folorunso, E. Schultes et al., Design of a FAIR digital data health infrastructure in Africa for COVID-19 reporting and research, *Advanced Genetics* **2**(2) (2021).

[8] S. Warnat-Herresthal, H. Schultze, K.L. Shastry, S. Manamohan, S. Mukherjee, V. Garg et al., Swarm learning for decentralized and confidential clinical machine learning, *Nature* **594**(7862) (2021), 265–270. doi:10.1038/s41586-021-03583-3.