Conference on Research Data Infrastructure

eventual DOI

Designing Data Infrastructure Maturity Profiles

[Cross-disciplinary Track: RDM Infrastructures]

Shuai Wang[1]*

[1]University Library, Maastricht University,
Grote Looiersstraat 17, 6211KG Maastricht, the Netherlands

*Correspondence: Shuai Wang, shuai.wang@maastrichtuniveristy.nl

Abstract

Data infrastructures (DI) [1], in particular data repositories and portals, are critical in the ecosystem for research data to be Findable, Accessible, Interoperable, and Reusable (FAIR) [2]. Tools specifically designed for modeling DIs are lacking, complicating the comparison of implementation specifications like persistent identifiers, communication protocols, and authentication services across different infrastructures. Consequently, assessing an infrastructure's alignment with the FAIR principles and its adherence to other recommended guidelines is difficult without reviewing comprehensive documentation, which is often not accessible to the public. In addition, diverse features were found in DIs with different implementation priorities depending on the types of metadata/data and domain-specific requirements [3]. For example, the PURE3D[1] Research Infrastructure has a 3D interface of the object. However, the integration of metadata has not yet been completed. Some metadata of DANS SSH Data Station was collected by the ODISSEI Portal [4] and Google Dataset Search[2] but it remains unclear when, how often, and with what selection criteria[3]. Moreover, ODISSEI Portal's documentation misses details on how the collected metadata is refined, enriched, and republished. Less studied is software dependency, which encompasses the inherited characteristics arising from utilizing existing software components (e.g., PURE3D uses Voyager[4]) and platforms (e.g., Dataverse [7] and CKAN[5]) [3].

An attempt towards a structured documentation of DIs is the re3data registry[6] [8]. By 26 April 2025, it includes 3,359 data repositories. For example, it provides essential features about the DANS SSH Data Station[7] [9], including its repository size, repository type, policy, persistent identifiers, metadata standards, etc. However, it is not sufficiently detailed how it aligns with the TRUST principles for data repositories [10]. For

[1] https://pure3d.eu/

[2] https://datasetsearch.research.google.com/

[3] Our examination on 26 April 2025 showed that, for example, a dataset named "DANS Nationale Enquête Arbeidsomstandigheden COVID-19 inclusief meting 4" (in English: DANS National Survey of Working Conditions COVID-19 including measurement 4) [5] was included by Google Dataset Search but not the ODISSEI Portal. The metadata of the Visio Divina Dataset [6] was collected by neither.

[4] https://smithsonian.github.io/dpo-voyager/

[5] https://ckan.org/

[6] https://www.re3data.org/

[7] https://www.re3data.org/repository/r3d100014195

example, the need for transparency calls for more detailed documentation with an examination of dependencies of software components, tools, licenses, etc. Moreover, the link between implementation choices and the FAIRness of its data/metdata remains insufficiently articulated. A recent attempt using the FAIR Implementation profile (FIP) for the ODISSEI Portal [11] demonstrated the potential of a more structured approach to documenting implementation specifications and how it can significantly facilitate analysis and research on DIs and their data/metadata. Documenting a DI's alignment with FAIR (software) principles[8] is advantageous, as a DI is fundamentally software.

This project addresses these challenges by providing a practical guide for assessing and advancing the maturity of data infrastructures. Inspired by existing FAIR assessment methodologies such as the FIP [14] and the FAIR Data Maturity Model [15], we introduce the *Data Infrastructure Maturity Profile* (DIMP), a structured assessment framework that maps high-level principles to concrete implementation decisions for the evaluation of DIs[9]. We systematically compare the above-mentioned principles/indicators/metrics and formulate implementation-driven questions. Furthermore, where necessary, we add questions regarding features captured by re3data, metrics proposed by previous DI assessments [3], checklists [16], and some software management plans (SMP) [17]. Finally, we adapt aligned questions to the context of DI design. By offering an actionable bridge between conceptual principles and operational practices, DIMP aims to be an instrument for documenting and comparing implementation choices, studying compliance with established principles, and benchmarking the progress of DIs and their implications on data management.

Resources

Supplementary material includes a table for the comparison of the above-mentioned principles, metrics, indicators, (FIP/SMP) questions, checklist, as well as a design of DIMP[10].

Author contributions

Shuai Wang: Conceptualization, Methodology, Formal Analysis, Writing – Original Draft, Writing – Review & Editing.

Competing interests

There is no competing interest as far as the author is aware.

Funding

This project was not funded.

[8] In this paper, we align metrics/indicators/questions with the FAIR4RS [12] and FAIRsoft [13] principles.

[9] As a proof-of-concept work, we limit our scope to data repositories and data portals. We exclude other registries and data archives.

[10] For this submission, the supplementary material can be found on Google Drive by following the link: https://drive.google.com/drive/folders/12C_rlm85tffQIAaEohThpJ2tTFtFg3wV?usp=sharing. The design will be published for final submission after further revision by domain experts.

Declaration on Generative AI

The author employed ChatGPT to transform text from bullet points into a coherent structure. TexGPT (via Writefull on Overleaf) assisted in rephrasing certain sentences. Neither contributed to the generation of ideas nor offered anything further than organizing/paraphrasing the text.

References

[1] L. Dodds and P. Wells, "Data infrastructure," *The State of Open Data*, p. 260, 2019.

[2] M. D. Wilkinson, M. Dumontier, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, Mar. 2016, ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18. [Online]. Available: http://dx.doi.org/10.1038/sdata.2016.18.

[3] M. Ali, C. Alexopoulos, and Y. Charalabidis, "A comprehensive review of open data platforms, prevalent technologies, and functionalities," in *Proceedings of the 15th International Conference on Theory and Practice of Electronic Governance*, 2022, pp. 203–214.

[4] T. Emery, R. Braukmann, M. Wittenberg, J. van Ossenbruggen, R. Siebes, and L. van de Meer, "The odissei portal: Linking survey and administrative data," *Royal Netherlands Academy of Arts and Sciences (KNAW)*, 2020. DOI: 10.5281/zenodo.4302095.

[5] C. B. voor de Statistiek (CBS), *DANS Nationale Enquête Arbeidsomstandigheden COVID-19 inclusief meting 4*, version V2, 2022. DOI: 10.17026/dans-2a9-s4sw. [Online]. Available: https://doi.org/10.17026/dans-2a9-s4sw.

[6] S. Wang, H. Makimei, and W. van Peursen, *The Visio Divina Dataset: A.I. generated images using biblical text*, version V1, 2025. DOI: 10.17026/SS/QA271C. [Online]. Available: https://doi.org/10.17026/SS/QA271C.

[7] M. Crosas, "The dataverse network: An open-source application for sharing, discovering and preserving data," *D-lib Magazine*, vol. 17, no. 1/2, 2011.

[8] H. Pampel, P. Vierkant, F. Scholze, *et al.*, "Making research data repositories visible: The re3data. org registry," *PloS one*, vol. 8, no. 11, e78080, 2013.

[9] Re3data.Org, *Dans data station social sciences and humanities*, en, date accessed: 26 April 2025, 2023. DOI: 10.17616/R31NJNG3. [Online]. Available: https://www.re3data.org/repository/r3d100014195.

[10] D. Lin, J. Crabtree, I. Dillo, *et al.*, "The trust principles for digital repositories," *Scientific Data*, vol. 7, no. 1, May 2020, ISSN: 2052-4463. DOI: 10.1038/s41597-020-0486-7. [Online]. Available: http://dx.doi.org/10.1038/s41597-020-0486-7.

[11] A. Valdestilhas, M. Windhouwer, R. Siebes, and S. Wang, "Comparing fair assessment tools and their alignment with fair implementation profiles using digital humanities datasets," in *International Workshop of Semantic Digital Humanities*, 2025.

[12] N. P. Chue Hong, D. S. Katz, M. Barker, *et al.*, "Fair principles for research software (fair4rs principles)," *Zenodo*, 2022.

[13] E. M. del Pico, J. L. Gelpi, and S. Capella-Gutiérrez, "Fairsoft - a practical implementation of fair principles for research software," *bioRxiv*, 2022. DOI: 10.1101/2022.05.04.490563. eprint: https://www.biorxiv.org/content/early/2022/10/10/2022.05.04.490563.full.pdf. [Online]. Available: https://www.biorxiv.org/content/early/2022/10/10/2022.05.04.490563.

[14] E. Schultes *et al.*, "Reusable FAIR Implementation Profiles as accelerators of FAIR Convergence," in *Advances in Conceptual Modeling*, Springer, 2020, ISBN: 978-3-030-65847-2.

[15] M. D. Wilkinson, M. Dumontier, S.-A. Sansone, *et al.*, "Evaluating fair maturity through a scalable, automated, community-governed framework," *Scientific data*, vol. 6, no. 1, p. 174, 2019.

[16] T. S. S. Institute, *Checklist for a software management plan*, version 1.0, Dec. 2018. DOI: 10.5281/zenodo.2159713. [Online]. Available: https://doi.org/10.5281/zenodo.2159713.

[17] C. Martinez-Ortiz, P. Martinez Lavanchy, L. Sesink, *et al.*, *Practical guide to software management plans*, version 1.0, Oct. 2022. DOI: 10.5281/zenodo.7248877. [Online]. Available: https://doi.org/10.5281/zenodo.7248877.