

First Steps Towards a Structured FAIRification Workflow for the DZD CORE DATASET at the German Center for Diabetes Research

Esther Thea Inau^{1,*}, Angela Dedie², Ivona Anastasova², Andreas Birkenfeld², Brigitte Fröhlich², Martin Hrabě de Angelis², Michael Roden², Renate Schick², Yaroslav Zdravomyslov², Atinkut Alamirrew Zeleke¹, Dagmar Waltemath¹ and Martin Preusse²

¹Department of Medical Informatics, University Medicine Greifswald, Greifswald, Germany

²German Center for Diabetes Research (DZD), Germany

1. Introduction

The German Center for Diabetes Research (DZD) conducts large clinical multicenter studies in the field of diabetes and metabolic research [1]. It has established the DZD CORE DATASET (CDS) which lists the clinical parameters relevant for diabetes research in related clinical studies.¹ Various initiatives, concepts and practices have been initiated to implement the principles of findability, accessibility, interoperability and reusability (FAIR) in data stewardship [2]. In this vein, we hypothesise that the implementation of a formalised, provenance-enabled and semantically enriched representation of (meta)data will add value to this dataset and further save time spent on data exploration, data selection and data processing. The objectives of this work are to conduct a baseline evaluation of the FAIRness of the DZD CDS and to formulate first steps in its FAIRification.

2. Methods and Results


We incorporated the FAIR Cookbook [3], the GO FAIR website² and the SATIFYD tool³ [4] in a baseline evaluation of the FAIRness of the DZD CDS and then formulated practical recommendations on how to improve its FAIRness. We then went on to implement some measures

14th International SWAT4HCLS Conference, February 13–16, 2023, Basel, Switzerland

*Corresponding author.

✉ inaue@uni-greifswald.de (E. T. Inau)

ORCID: 0000-0002-8950-2239 (E. T. Inau); 0000-0002-8191-2583 (A. Dedie); 0000-0003-3671-725X (I. Anastasova); 0000-0003-1407-9023 (A. Birkenfeld); 0000-0002-7898-2353 (M. H. d. Angelis); 0000-0001-8200-6382 (M. Roden); 0000-0001-7838-9050 (A. A. Zeleke); 0000-0002-5886-5563 (D. Waltemath); 0000-0003-4789-0592 (M. Preusse)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

¹<https://medical-data-models.org/45430>

²<https://www.go-fair.org/fair-principles/>

³<https://satisfyd.dans.knaw.nl/>

that we suppose will lead to a better FAIRification score. A higher FAIRification score is the benchmark for success of this work. At the time of conducting this work the CDS was only available as an MS Excel file on the DZD website. A substantial part of it was not available in any structured form nor did it have an accompanying data dictionary. To improve its FAIRness score, we converted the CDS into the recommended format for spreadsheets, annotated the parameters therein with LOINC [5], licensed the dataset, indexed the metadata in a searchable resource and enriched the dataset with metadata. These steps resulted in an overall increase in the FAIRness score by 47%.

3. Discussion and Conclusion

The DZD strives to set standards for good scientific management of clinical studies. The development of a common DZD CDS increased the interoperability of clinical studies under the DZD umbrella. The next step towards reusability and comparability of studies is the adherence to the FAIR principles. The data owners therefore agreed to direct their efforts into structuring the data and mapping the local codes to standard terminologies in order to provide understandable, valuable and fit-for-purpose data. The mapping task is not trivial and requires domain knowledge as well as understanding of the used standard terminology terms to make sure that the semantic meaning is correctly translated from the local codes to the applied standards. Data validation remains a critical step to ensure that the data generated and codes used are valid. According to the SATIFYD, a dataset containing an open license is FAIRer than one containing a restricted, embargo or any other license. The FAIR Cookbook however indicates that the type of license chosen does not cause differences in the FAIRness. This is a common misconception about licensing in matters FAIR [6]. The implementation of a machine actionable format will come with the cost of time and immense effort because the DZD CDS is not encoded in any structured format. The return on this investment is increased certainty of the future data readability. We postulate that the results of this evaluation will help formulate the first steps towards a FAIRification workflow for related DZD datasets, facilitate the planning for the resources that the FAIRification journey will require and motivate the stakeholders to engage in this journey.

References

- [1] B. Niesing, Deutsches Zentrum für Diabetesforschung, *Diabetes aktuell* 19 (2021) 55–56.
- [2] E. T. Inau, et al., Initiatives, concepts, and implementation practices of FAIR (findable, accessible, interoperable, and reusable) data principles in health data stewardship practice: protocol for a scoping review, *JMIR research protocols* 10 (2021) e22505.
- [3] P. Rocca-Serra, et al., D2.1 FAIR Cookbook, 2022. doi:10.5281/zenodo.6783564.
- [4] D. Slamkov, V. Stojanov, B. Koteska, A. Mishev, A Comparison of Data FAIRness Evaluation Tools, *Proceedings* <http://ceur-ws.org> ISSN 1613 (2022) 0073.
- [5] C. J. McDonald, et al., LOINC, a universal standard for identifying laboratory observations: a 5-year update, *Clinical chemistry* 49 (2003) 624–633.
- [6] B. Mons, The VODAN IN: support of a FAIR-based infrastructure for COVID-19, *European Journal of Human Genetics* 28 (2020) 724–727.