EasyChair Preprint

№ 9351

How FAIR Is NUM? – Lessons Learnt from a FAIR Survey Within the German Network University Medicine (NUM)

Lea Michaelis, Rasim Atakan Poyraz,
Michael Rusongoza Muzoora, Kerstin Gierend,
Alexander Bartschke, Dagmar Waltemath and Sylvia Thun

November 22, 2022

How FAIR is NUM? – Lessons learnt from a FAIR survey within the German Network University Medicine (NUM)

Lea Michaelis[1][0000−0001−9691−2677], Rasim Atakan Poyraz[2][0000−0002−5705−7396], Michael Rusongoza Muzoora[2][0000−0002−1384−1509], Kerstin Gierend[3][0000−0003−0417−3454], Alexander Bartschke[2][0000−0002−5849−482X], Dagmar Waltemath[1,4][0000−0002−5886−5563], and Sylvia Thun[2][0000−0002−3346−6806]

[1] Core Unit Data Integration Center, University Medicine Greifswald, D-17475 Greifswald, Germany
[2] Core Unit eHealth and Interoperability, Berlin Institute of Health at Charité, Berlin, Germany
[3] Department of Biomedical Informatics at the Center for Preventive Medicine and Digital Health, Medical Faculty Mannheim,Heidelberg University
[4] Medical Informatics Laboratory, University Medicine Greifswald, D-17475 Greifswald, Germany

Abstract. The imminent need to harness large amounts of data, possibly within a short period of time, became extremely apparent during the Covid-19 pandemic outbreak. A particular solution for the collection of COVID-19 data across German University Hospitals was a dedicated Corona Data Exchange Platform (CODEX+), which had been developed within the German Network University Medicine. German Network University Medicine funded 21 subprojects in 2021/22, and the adherence with the FAIR principles had been discussed and planned. The FAIR principles for data stewardship have become a prominent building block of health research data management. They enable research networks to evaluate how good they comply with current standards in open and reproducible science. It is thus important to provide a general overview of the FAIRness of data across German Network University Medicine the projects. To be more transparent, but also to provide guidelines for scientists within the network on how to improve the data reusability, we disseminated an online survey within the German Network University Medicine and across individual projects and research datasets. The expectation was to identify positive examples of FAIR data in the German Network University Medicine thus to motivate other projects to take similar routes. Despite the results of the survey not encompassing the entire network, the analysis could support decisions about the future direction of research data management in German Network University Medicine and other biomedical research networks.

Keywords: FAIR · Survey · Medical Informatics.

1 Introduction

The ongoing digitalization creates large amounts of digital data in health care and research. However, the majority of data remain difficult to find, access, and reuse, due to license rights and GDPR compliance. Research networks hence spend time and effort on implementing research data strategies, including interoperability and data sharing. A large portion of FAIR compliance requirements are identified during FAIR assessments, which are usually based on the FAIR Guiding Principles for Data Stewardship [8] and then adapted to the specific domain or project of interest. The German Network University Medicine (NUM) was founded in April 2020, after the COVID-19 outbreak, with the determined goal to coordinate German COVID-19 strategies and research activities. Within NUM, the CODEX platform was developed to collect and share COVID-19 related data from all German hospitals to help the government make informed decisions [2]. As such, data interoperability, and standardisation were a prerequisite. CODEX aimed at establishing a nationwide, uniform, privacy-compliant infrastructure for storing and providing COVID-19 research datasets [5]. We ran a self-assessment of FAIRness across different projects funded within NUM [7] to comprehend the current status of "FAIRness". The survey is thematically divided into general questions about the projects and experiences with FAIR evaluations, followed by four sections each addressing one of the FAIR categories.

2 Methods

We invited all partners of the German Network University Medicine to participate in an online self-assessment [7]. The questionnaire is composed of 33 questions with over 100 possible answer options. The aim of the survey was to evaluate the degree “FAIRness” across the different NUM Sites and Projects. For us to obtain reliable data in the cross-site survey, we decided to use REDCap [4], a secure, GDPR compliant and widely used Research Electronic Data Capturing Tool. The questionnaire was analysed according to 1) the granularity of questions, 2) specific terminology used, and 3) to comprehend the level of implementation for each FAIR principle. The questionnaire was then shared within a small community of "FAIR experts" to provide feedback on the suggested questions. By including GDPR experts feedback, we ensured that our survey was GDPR compliant. We reused previously established questions from online FAIR frameworks and FAIR self-assessment tools, in particular ARDC [1] and FAIR enough [3], to record the adherence to the FAIR principles. We disseminated the survey link within the NUM and Twitter.

3 Results

During the productive survey time period (1.9 - 14.11.2022), over 100 entries were recorded, however, only five of these entries were fully submitted from a

total of three NUM projects. None of the participants had used a specific FAIR evaluation tool before, but one participant had previously participated in a FAIR evaluation. The participants were asked about data management activities, and three participants had Standard Operating Procedures (SOP) in place for data management. Three other participants had used a data management tool. One participant did not know if any data management activities were happening in the project.

Findability 80% of the participants assign local identifiers to their data, 60% use web addresses (URLs), and 40% use globally unique, citable, and persistent identifier (Fig. 1). 20% of the participants did not know if there is any assigned identifier to their data. A local identifier is considered the minimum requirement for data items. Two participants (40%) stated that the metadata file in use contains unique identifiers to the actual data. 40% of the entries used a domain-specific repository to encode metadata in, and 40% used a generalist public repository such as GitHub or Confluence. Interestingly though, two participants (40%) did not know if their metadata had any identifiers assigned. A question was asked to identify if sufficient metadata was provided to make data findable, understandable and reusable. Two participants (40%) reported about providing only minimal metadata for required fields, one participant also filled additional fields beyond the required fields. Two participants reported to provide rich metadata with as much information as possible. 60% of the participants have made additional documentation about their data by adding readme files, versioning and provenance information. However, at the same time, two participants selected, "I do not know".

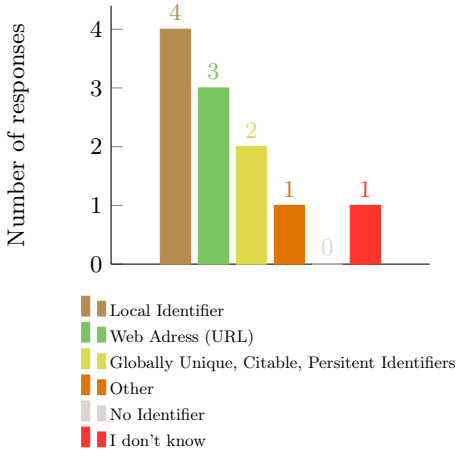

Fig. 1. The graph shows if the participants' data have any identifiers assigned

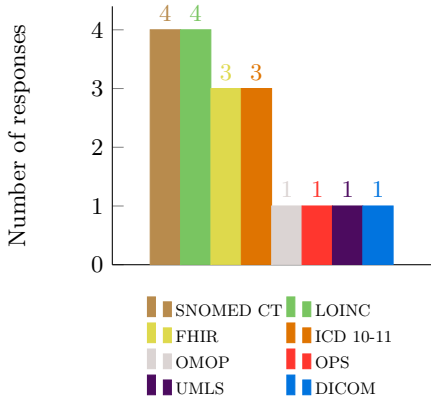

Fig. 2. What semantic interoperability standards are being used

Accessible Two participants (40%) reported to only have full access to the project data if the subjects explicitly stated conditions such as ethics approval for sensitive data. One participant (20%) had solely access to the metadata. Furthermore, 80% of the participants stated that their data of interest could be accessed online by individual arrangements. 60% of the participants used open access (CCO) licenses which had also been recommended by the NUM. One person (20%) did not know if the NUM recommended any licenses, and one person (20%) stated that no recommendation existed.

Interoperable All participants replied that their data was available in a structured, open standard and machine-readable format (Fig. 2). The semantic interoperability standards were: FHIR, SNOMED CT, LOINC, ICD 10-11, and OMOP CDM. Contextual information about the data was provided via persistent identifiers and/or reference to other datasets. The provenance of their data included the origin of data and versioning history. 60% of the participants used either ARTDecor, Simplifier or FHIR Provenance tools for data modeling.

Reusable One participant reported that the project provided provenance information via adding version history of data. Another participant reported that the origin of data, processing history of data and version history of data are provided. One participant replied that they did not know if there was any provenance information available in their project data. Finally, one participant reported that their project did not include provenance information. 40% of the participants reported non-standard, text-based licenses being attached to the data, and 20% of the participants did not use any license attached their data. 40% of the participants were not able to answer the question.

4 Discussion

The overall participation in the survey was very low despite broad announcement. 107 people opened the survey, but only five of them completed it. We hypothesise that the survey questions demanded too much technical detail about the actual IT infrastructure, as had been reported by one participant. This might have discouraged the completion of the survey. More participants might be reached with an accompanied survey, allowing participants to ask questions and receive further explanations. However, the survey clearly showed that FAIRness is not yet a particularly high priority within NUM projects and that there is still a great potential for improvement. To motivate further FAIRification processes within NUM, a strategy for FAIRification is needed to ensure that data from within the German Network University Medicine can reliably be reused for research. Hence the benefits of FAIR principles should be communicated, possibly using positive examples of FAIR datasets within NUM. The German Corona Consensus Dataset (GECCO)[6] is a prime example of how a basis for compliance with the FAIR principles can be established in COVID 19 research. When considering the aspect of interoperability, the use of FHIR, as a syntactic

standard, makes machine-readable metadata available. Using international semantic standards e.g SNOMED CT, LOINC and ICD 10 ensures that a broadly applicable language for knowledge representation within the dataset is used.

5 Conclusion

We conclude that the NUM community lacks awareness for the FAIR data principles. Consequently, more information on the benefits of FAIRification should be provided and the leadership should develop a strategy for research data management which addresses the missing FAIR components such as provenance information, licensing, open data repositories. "FAIR experts", such as data stewards, should assist the NUM community members in the overall FAIRification process and provide a single point of contact for interested researchers. Support is needed in particular to 1) Cross check interoperability of datasets across NUM projects, e. g. LOINC, SNOMED CT; 2) Provide assurance on used codes and/or alternative mappings. Once this structure is in place, specific guidelines and rules should be defined within the NUM to change the mindset of the community and pave the way towards FAIR data across the NUM-Sites.

Acknowledgements This project has been funded by the German Ministry of Health (BMBF), FKZ 01KX2121. Thanks go to Martin Sedlmayr and Brita Sedlmayr for their feedback and help with the survey design.

References

1. Ardc fair data self assessment tool. https://ardc.edu.au/resource/fair-data-self-assessment-tool8, accessed: 2022-11-8
2. NUM homepage. https://www.netzwerk-universitaetsmedizin.de, accessed: 2022-11-7
3. Emonet, V., et al.: Towards an extensible fairness assessment of fair digital objects. Research Ideas and Outcomes **8**, e94988 (2022)
4. Harris, P.A., Taylor, R., Minor, B.L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., Kirby, J., et al.: The redcap consortium: Building an international community of software platform partners. Journal of biomedical informatics **95**, 103208 (2019)
5. Prokosch, H.U., et al.: The covid-19 data exchange platform of the german university medicine. In: Challenges of Trustable AI and Added-Value on Health, pp. 674–678. IOS Press (2022)
6. Sass, J., Bartschke, A., Lehne, M., Essenwanger, A., Rinaldi, E., Rudolph, S., Heitmann, K.U., Vehreschild, J.J., von Kalle, C., Thun, S.: The german corona consensus dataset (gecco): a standardized dataset for covid-19 research in university medicine and beyond. BMC Medical Informatics and Decision Making **20**(1), 1–7 (2020)
7. Waltemath, D., et al.: FAIR evaluation at the NUM sites - How FAIR is NUM? (Nov 2022). https://doi.org/10.5281/zenodo.7299373, https://doi.org/10.5281/zenodo.7299373
8. Wilkinson, M.D., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data **3**(1), 1–9 (2016)