How are we FAIR-ing? Creating a FAIR self-assessment checklist for data repositories

Lauren Phegley[1] and Lynda Kellam[2]

Abstract

In 2023, a team from a local grant-funded medical data repository requested guidance from Penn Libraries on evaluating the extent to which their repository was FAIR-enabling. They wanted to know if their repository had the policies, infrastructure, and documentation to allow data that is deposited to be FAIR – Findable, Accessible, Interoperable, and Reusable. After a consultation with the repository team, our research data experts discovered that many of the current self-assessments of the FAIR guidelines were for data creators rather than data repository managers. In addition, we wanted a self-assessment tool similar to the process and guidance created by CoreTrustSeal but focusing explicitly on the FAIR Principles. In answer to their request, the Penn Libraries Research Data Engineer conducted a literature review and coalesced current guidance and assessment tools on the principles. After this review of the existing documentation, a small team developed a FAIR Principles self-assessment tool for repository teams. In addition to several iterations of the tool, we also met with the repository team for feedback on making the tool more understandable. Our conversation provided insights into the challenges of explaining the FAIR Principles to those without information science or data backgrounds. The discussion and creation of this self-assessment tool helped develop a more transparent and trustworthy repository. This paper will discuss our process for developing the assessment, the goals for utilizing the tool, and the lessons learned. Reporting our findings as they currently stand will prompt the research data management field to ruminate on the adoption of FAIR Principles for data repositories. We also intend to encourage conversation on the usability of the FAIR Principles for professionals without an information science or data background.

Keywords

FAIR Principles, data repository, data sharing, assessment

Introduction

The first publication on the FAIR Principles was in the 2016 article “The FAIR Guiding Principles for scientific data management and stewardship” by Wilkinson et al. The intent of the FAIR Principles is to create discipline agnostic guidance to enhance the reproducibility of data. The four foundational principles – Findability, Accessibility, Interoperability, and Reusability – are a minimal set of guiding practices that improve the experience for the data creator and downstream users. The FAIR Principles were created as guidance for any individual or system that encounters the data throughout its

lifecycle. Each of the words that make up the FAIR acronym are considered high level principles and have specific indicators (labeled F1, F2, etc.) to use in evaluating if a dataset meets the principle. Findability is focused on the ability for machines and humans to identify the data through high quality metadata and indexing. The Accessibility principle expects a free, open, and standardized way of accessing the data through their identifier. The Interoperable principle expects the entities data and metadata to follow a shared, defined vocabulary that assists in having broadly understood meanings. The Reusable principle focuses on how well defined the licenses, provenance, and metadata are for future users. To assist users of the assessment, we linked each of The Fair Principle indicators to a GO FAIR resource page that elaborates and explains its purpose and how it is used. GO FAIR is a stakeholder driven initiative dedicated to helping implement the FAIR Guiding Principles, and one of the most thorough resources for learning more about the FAIR Principles (n.d.).

In February of 2023, a grant-funded medical data repository team reached out to the Penn Libraries' Research Data & Digital Scholarship unit seeking a way to assess their implementation of the FAIR Principles. A small group consisting of the Research Data Engineer, the Director of Research Data & Digital Scholarship, the Director of the Holman Biotech Commons, and the Bioinformatics Librarian at the Penn Libraries met with the repository team and an individual from the Perelman School of Medicine's computing services to discuss their goals. Upon meeting, the data repository team explained they were invited to a symposium by their grant funder to discuss how their repository aligns with the FAIR Principles. The issue was translating the high-level, conceptual FAIR Principles into specific criteria that they could use to evaluate their repository. Our team decided that we would search for an existing FAIR Principles assessment that met their needs.

Our search for an assessment instrument for FAIR-enabling data repositories that fit the team's needs uncovered a multitude of resources that did not quite fit. The data repository team needed a tool that did not expect previous information science training or require a detailed understanding of the FAIR Principles. Since the publication of the FAIR Principles (Wilkinson et al., 2016), there has been a plethora of research into these ideals and a deep commitment to improving the data sharing landscape. Despite this, many resources are jargon-filled, created by the information science profession for the information science profession, or focused on assessing data objects rather than the repository. Based on our findings, we decided to build upon the hard work of many scholars of the FAIR Principles to create an assessment instrument for data repositories to evaluate how FAIR-enabling they are (Phegley et al., 2024). Our goal was to create an accessible self-evaluation tool that could be used by repository staff to assess their current level of FAIR adherence, allowing them to reflect on how to take realistic, actionable steps toward improvement.

We were motivated to create this new resource instead of using a pre-existing assessment for multiple reasons. First, we uncovered more questions than answers during our investigation into finding an appropriate assessment. The FAIR Principles research sphere is bursting with scholarly products. Yet it seemed that many did not delineate between the data creator's and the repository team's responsibility for enacting the FAIR Principles. We wanted a resource that focused on evaluating the repositories' FAIR-enabling practices rather than the choices made by the data depositor. Second, we believe in the importance of the FAIR Principles and working towards increased adoption amongst data repository teams. We want to support data repository managers' capacity to conduct their own evaluations and improve their understanding of the FAIR Principles. Creating and sharing a new low

barrier to entry tool that acts as a stepping stone towards a better data sharing environment allowed us to enact our values.

Development of the Assessment

Our intention was to build a FAIR assessment instrument that, as David et al. (2020, p. 3) states, "must be realistic and pragmatic – what should be measured, and how to explicitly find the information needed." There is a rich offering of resources and assessments from scholars of the FAIR Principles, and we conducted a literature review to find materials that support creating this instrument. The Research Data Alliance (RDA) Fair Data Maturity Working Group created two resources that we relied on heavily for the specific questions and interpretation of the FAIR Principles into the data repository context: FAIR Data Maturity Model Specification and Guidelines (2020) and Results of an Analysis of Existing FAIR Assessment Tools (2019). The RDA resources, in addition to other materials, gave us a firm foundation of criteria that encapsulate the FAIR Principles (Behnke et al., 2020; Hahnel and Valen, 2020; L'Hours, 2022; Murphy et al., 2021).

Our assessment instrument is structured as a manual self-assessment checklist divided into four core FAIR sections (Findable, Accessible, Interoperable, and Reusable). Each category includes detailed FAIR Principle indicators as developed by Wilkinson et al. (2016). We linked the indicators to the associated GO FAIR description page to provide additional context (n.d.). The indicators follow the Wilkinson et al. labeling that is commonly used for referencing the FAIR Principles sub-sections. For example, indicator F1 is the first indicator of Findable. Each of the indicator sections has associated criteria describing a FAIR practice ("Dataset has an assigned identifier") (Figure 1, Label A). Most of the indicators have multiple criteria that are labeled according to the FAIR Principles indicator they respond to. For example, criteria F1-01 states, "Dataset has an assigned identifier," and criteria F1-02 is "Related works (connected literature, data, authors, project, and code) are able to be connected via persistent identifiers". Researchers then tick the checkbox(s) that best describes their repository's current adoption level of that FAIR practice (Figure 1, Label B).

Figure 1. Criteria F1-01 from the Implementation Checklist

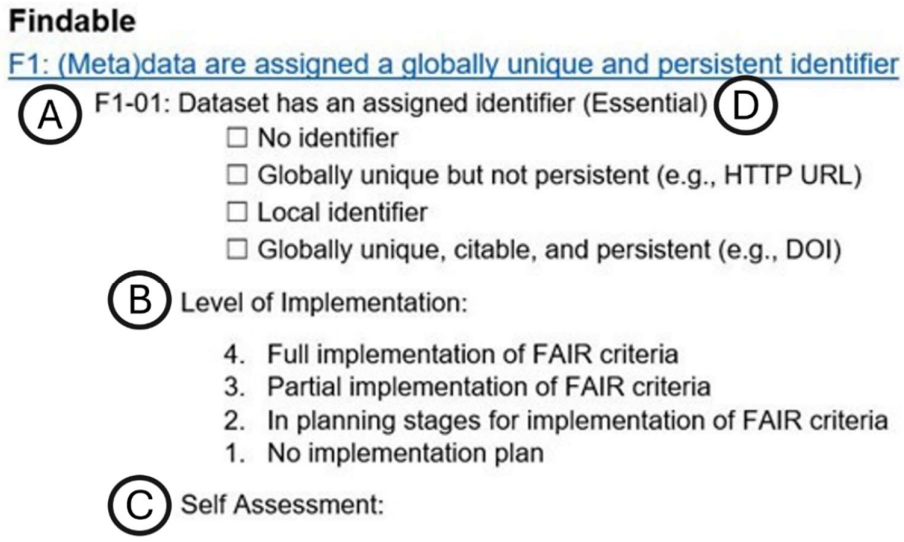

Figure 1: The first indicator for a Findable (F1) repository has criteria (F1-01) that allow evaluators to identify the repositories current adoption level. The evaluator is then able to explain their level of implementation and provide a self assessment.

We intentionally built this tool as a manual self-assessment that allows for nuanced evaluation and thoughtful self-reflection. Inspired by the CoreTrustSeal self-evaluation process, we wanted repository staff to use the levels of implementation to reflect on current repository practices (CoreTrustSeal Standards and Certification Board, 2022). As such, there is a space at the bottom of each FAIR criteria for a narrative self-reflection to explain why they have implemented certain processes and how they want to improve (Figure 1, Label C). To know how to improve, the data repository team needs to know where they stand. Reflection encourages the team to engage with the FAIR process in a deeper way than would be possible with an automated tool. Undoubtedly, this translates into the checklist taking more time than an automated tool. The benefit of this process is that the repository team can learn the FAIR Principles overall rather than relying on an automated tool to tell them what might be wrong without context. While automated tools tools are useful, they do not assist in building a team's understanding of the FAIR Principles in relation to their own repositories.

Additionally, we created an evaluation that is not prescriptive in the FAIR-enabling criteria. In other words, the checklist can be applied to repositories without having to worry about an all or none approach to FAIR implementation. Not all repositories can achieve the same implementation levels due to policy requirements, budget constraints, repository infrastructure limitations, and disciplinary practices. All the criteria contribute towards creating a FAIR-enabling repository, but certain criteria are more crucial than others. This is why each criterion has an associated priority level of essential, important, or useful (Figure 1, Label D).

More than one adoption level checkbox can be checked at a time to create a flexible yet reliable evaluation, as more than one scenario might be true in a repository. The order of the adoption level checkboxes tends to increase in complexity and comprehensiveness. This allows repository teams to

see the next logical step for increased FAIR adoption for that criterion. They can provide a detailed explanation in the self-assessment section about current implementation, reasoning for why increased implementation may not be viable, and discuss their plans, if they want to improve. The level of implementation is assessed by repository personnel based on their own understanding of their repository. Due to the differences in repositories, there is currently no standard for the differences between implementation levels. As described in the assessment section, feedback from the data repository team showed us that it is difficult for individuals who are not experts in the FAIR Principles to evaluate their own level of implementation. Developing a standard for the levels of implementation of the FAIR Principles is a future opportunity for development.

This approach allows data repositories with various constraints to participate in a FAIR evaluation. For example, a specialist repository, such as a gut microbiome data repository, may only encourage standard compliance for criteria R1.3-02, "Repository requires or encourages data to comply with an applicable community standard." Gut microbiome data can take many forms (genomic data, proteomic data, etc.), where each data type may or may not have its own community standard. Because of this, it may be more realistic to encourage compliance with an applicable standard, rather than requiring data to comply with a standard. This is an example of a situation where an automated FAIR evaluation would miss the important context that a manual self-evaluation adds.

Assessment Feedback

Upon presenting the first iteration of the assessment to the data repository team, we met with them to get feedback on their experience using the tool. The data repository team amounted to about four people, all of them with extensive biomedical experience but less information science experience. Overall, they found our tool to be useful in figuring out where they were and where they wanted to go, but they had trouble interpreting the technical elements. They found that the assessment required a level of technicality they could not understand and requested additional descriptive information to help them. They suggested that it might be a more meaningful activity if an outside expert could act as a second reviewer. This is the existing model of the CoreTrustSeal process and while we agree this is a beneficial process, we want the assessment to be accomplished without expecting a secondary reviewer. Not all repository teams have a data librarian or FAIR Principles expert who can assist with this process.

They also had questions about how to evaluate their implementation level, as they themselves were not experts in the FAIR Principles. We were inspired by the CoreTrustSeal practice of the repository team evaluating their own level of implementation (CoreTrustSeal Standards and Certification Board, 2022). This suggestion led us to create two versions of our checklist, one where the reviewer can evaluate the level of implementation for each criterion and one where there is no level of implementation section. We also had the opportunity for an expert in the FAIR Principles to review the tool and provide feedback before we released the updated version on the University of Pennsylvania's institutional repository, ScholarlyCommons. The feedback from the team and the FAIR expert helped us understand what we could improve in the assessment and the information that researchers need to understand the FAIR Principles overall.[3]

Addition to the FAIR Landscape

Our assessment builds on the existing FAIR Principles research by shifting the focus of what is being evaluated. While the originating FAIR Principles article by Wilkinson et al. (2016) states, "The Principles define characteristics that contemporary data resources, tools, vocabularies, and infrastructures should exhibit to assist discovery and reuse by third-parties", many tools currently are oriented towards data objects. Only a few resources are oriented toward supporting FAIR-enabling data repository infrastructures. We anticipate that as the FAIR Principles continue to gain attention and have increased uptake, especially outside of the core library and data communities, the principles will continue to be adopted as a standard for other adjacent assets.

A second aspect of this assessment is that it differentiates between the data creator and the repository team regarding who is enacting the FAIR Principles. Our tool is made to evaluate the FAIR-enabling adoption levels of a repository, not to evaluate any individual dataset. For criteria that would traditionally be the responsibility of the data creator, such as data format or documentation, our assessment allows repository teams to indicate that they encourage or require a certain action. Assessment tools should not assume that the creators of the datasets will update the datasets to comply with FAIR Principles or that the individuals running the repository have the time or tools to do it themselves. We cannot fully assess a repository and its containing datasets as one unit until we realize sharing FAIR data has to be a conversation between the data creator and the repository manager, as neither can be FAIR-enabling alone.

From the beginning of this endeavor, we have approached the FAIR Principles as a goal to achieve rather than a marker of success or failure. We do not want repository managers to assume they fail if they are not fully FAIR-enabling. We want the tool to enable honest evaluation of the FAIR status of the repository and provide a plan with concrete steps towards further adoption of the FAIR Principles. Our goal was to take a step towards making the FAIR Principles in this assessment usable for those without an information science or data background. Approaching assessment from a place of shame or admonishment for not fully implementing the FAIR Principles negates the tool's usefulness by removing the possibility for learning and growth.

Lessons Learned

The first lesson we learned was that the current FAIR Principles information ecosystem expects a certain amount of background knowledge prior to introducing the principles. Few materials attempt to explain topics like "protocols" in an accessible way. Most materials seem to be written by information professionals for information professionals, which defeats the purpose of trying to increase adoption among all data stakeholders. Implementing tools to create or support FAIR-compliant data will only be useful if the individuals implementing them understand the entire FAIRification process (David et al., 2020). We need to create information and tools that support researchers beginning their FAIR journey without expecting knowledge on topics such as machine actionability or metadata. If we believe in the importance of FAIR data, we need to create scaffolded learning material and tools.

This lesson was informed by our experience reviewing the FAIR Principles literature and the repository teams' feedback on our assessment. Our experience demonstrated that there is a conceptual knowledge gap between the concept of data and the application of the FAIR Principles. The repository

team had a deep expertise with data, specifically bioinformatics data, but that knowledge did not directly translate into knowledge of the FAIR Principles. Even the Research Data Engineer on the Penn Libraries team who was familiar with the FAIR Principles had to spend many hours untangling jargon-filled sentences, defining terms, and attempting to weave the information into shareable knowledge.

The repository team let us know that while using the assessment, there were certain questions that they could not answer due to confusion over the criteria's meaning. We attempted to provide more clarification without bogging down the tool by linking to each of the principles developed by the GO FAIR Foundation (n.d.). The team did not find that the GO FAIR explanations clarified the criteria and had to skip sections of our assessment due to this. They also wanted to see examples of repositories that had fully implemented the FAIR Principles to emulate. We agree this would be an incredible learning tool, but we are unaware of a repository that fulfills these criteria in a way that can be demonstrated to individuals without requiring administrative-level access and an in-depth walkthrough by a repository staff member. This feedback demonstrates that there is a knowledge gap between the intentionally high concept FAIR Principles and the specific actions required to implement those principles.

The second lesson we learned was that FAIR researchers and educators need to separate what is attributed to a FAIR-enabling data repository and the dataset aligning to the FAIR Principles. The ability to share a FAIR dataset is a collaboration between a repository's infrastructure and its datasets. The checklist was developed for repository managers to assess the repository's infrastructure and policies and evaluate how much they allow full FAIR adherence. For example, when building checklist criteria for FAIR Principle R1: "(Meta)data are richly described with a plurality of accurate and relevant attributes," the criteria asks if the repository requires or encourages documentation to accompany datasets (Phegley et al., 2024). Documentation is incredibly important for FAIR data, but we did not want the repository to be deemed less FAIR-enabling because the depositor chose not to create rich documentation. Instead, the repository shows it is FAIR-enabling by requiring or encouraging rich documentation with all datasets. Meaningful FAIR data sharing only happens when the repository is built to support FAIR data sharing, and the data is also created to meet the FAIR Principles.

Future Directions

Throughout the process of creating the FAIR data repository checklist, we realized the importance of scaffolded FAIR Principles instruction. As research data management educators, we need to create accessible information on the FAIR Principles that professionals without an information science or data background can use. Unfortunately, directing someone to the FAIR Principles does not address what the principles mean in context or how to implement them. We lack jargon free guidance with example case studies and visuals that scaffold towards a comprehensive understanding of the FAIR Principles.

We hope this article and the associated checklist will prompt conversation in the research data management field on evaluating the FAIR Principles in a data repository. The FAIR assessment checklist is an early iteration of what we hope to improve based on feedback from data repository teams, FAIR experts, and the research data management community. We are especially eager to hear from individuals who have used the checklist to evaluate their data repository. As our community works

towards supporting FAIR-enabling data repositories, we look forward to this assessment being used as a stepping stone towards a more robust understanding of the FAIR Principles.

References

Bahim, C., Dekkers, M. and Wyns, B. (2019). "Results of an analysis of existing FAIR assessment tools". https://doi.org/10.15497/rda00035.

Behnke, C., Bonino, L., Coen, G., Le Franc, Y., Parland-von Essen, J., Riungu-Kalliosaari, L., and Staiger, C. (2020). "D2.3 Set of FAIR data repositories features". FAIRsFAIR. https://doi.org/10.5281/ZENODO.3631527.

CoreTrustSeal Standards And Certification Board. (2022). "CoreTrustSeal Requirements 2023-2025". https://doi.org/10.5281/ZENODO.7051011.

David, R. et al. (2020). "FAIRness literacy: The Achilles' heel of applying FAIR principles", *Data Science Journal*, 19, p. 32. https://doi.org/10.5334/dsj-2020-032.

GO FAIR Foundation. Interpreting FAIR. https://www.gofair.foundation/interpretation (Accessed: 2024-04-10).

Hahnel, M. and Valen, D. (2020). "How to (easily) extend the FAIRness of existing repositories", *Data Intelligence*, 2(1–2), pp. 192–198. https://doi.org/10.1162/dint_a_00041.

L'Hours, H. (2022). "Report on a maturity model towards FAIR data in FAIR repositories (D4.6)". https://doi.org/10.5281/zenodo.6699520.

Murphy, F., Bar-Sinai, M. and Martone, M.E. (2021). "A tool for assessing alignment of biomedical data repositories with open, FAIR, citation and trustworthy principles". *PLOS ONE*. Edited by F. Naudet, 16(7), p. e0253538. https://doi.org/10.1371/journal.pone.0253538.

Phegley, L., Kellam., L., de la Cruz Gutierrez, M., and Rajpal, N. (2024). "FAIR assessment checklist for data repositories". https://doi.org/10.48659/DTQC-2A45.

Research Data Alliance FAIR Data Maturity Model Working Group. (2020). "FAIR data maturity model: Specification and guidelines". https://doi.org/10.15497/RDA00050.

Wilkinson, M.D. et al. (2016). "The FAIR guiding principles for scientific data management and stewardship". *Scientific Data*, 3(1), p. 160018. https://doi.org/10.1038/sdata.2016.18.

Endnotes

[1] Lauren Phegley is the Research Data Engineer at the University of Pennsylvania. Her email is lphegley@upenn.edu and ORCID is 0000-0001-7897-1841.

[2] Lynda Kellam is the Director of Research Data and Digital Scholarship at the University of Pennsylvania. Her ORCID is 0000-0002-3263-859X.

[3] The authors are incredibly thankful to Laurence Horton, Research Data Specialist at the Digital Curation Centre, for his enthusiastic feedback on the assessment.