Master Thesis Proposal

Yiqing Qu

Department of CIT, Purdue University

November, 2023

Purdue University

West Lafayette, Indiana

Master Thesis Proposal

Contents

Introduction

Background

Scientific data management has evolved significantly over the years, with a consistent emphasis on ensuring data not just produced but also efficiently utilized. To create a clear guideline, the FAIR principles have emerged over the years. Since the proposal of FAIR principles in 2016, It has greatly influence the development of infrastructures, and become an important part for transparent, collaborative, and efficient research methodologies. While the FAIR principles are fundamental and abstract guidelines for scientific data management, there are some following study on the technological implementations. FAIR principles experienced a evolution over the years. It consists four core principle concepts: Findability, Accessibility, Interoperability, and Reusability. These principles make sure the data are easily found and discovered by researchers, easily to access, easily to suppport smooth collaboration and are repeatable and are used again for different studies. FAIR is no longer just a theoretical principle but a practical necessity. The importance of having datasets adhering to FAIR data principles becomes increasingly evident as research are becoming more collaborative and interdisciplinary, especially in big-data platforms.

Significance of FAIR Principles in Contemporary Research

Research Question and Project Introduction

In the realm of FAIR data principles, while the theory is frequently referenced, its technological application varies and are still challenging. My research aims to address the specific question of how to design and implement a data resource portal in Django that adheres to the FAIR data principles. This portal will be capable of extracting dataset metadata and publishing it, ensuring FAIR compliance. Moreover, both the datasets and the Django application will be connected with a FAIR evaluation report generator. This tool will assess FAIRness using specific test cases. The research project is expected to

ensure FAIRness within the Django application and to provide insights into the practical development problems related to Django, message queues, RESTful APIs and Kubernetes. A notable innovation will be the ability to make FAIR-compliant datasets indexable and searchable on Google Dataset Search.

Objective and Scope of the Review

The primary objective of this literature review is to gain a comprehensive understanding of the FAIR data principles, especially within the context of software implementation and FAIRness evaluation. The literature will cover FAIR principles' conceptual framework and their significance in contemporary research, an analysis of the existing technological applications between theory and actual implementations, methodologies for designing FAIR-compliant portals and tools for metadata extraction and FAIRness evaluation. Additionally, some case studies of FAIR implementation will also be reviewed.

Limitations

Literature Review

FAIR Principles

Background and Conceptual Framework. In recent years, there has been a growing emphasis on enhancing the reusability and accessibility of scholarly data. The FAIR Data Principles, which stands for Findable, Accessible, Interoperable, and Reusable, have emerged as a set of concise and measurable guidelines to achieve this aim. The principles are proposed and then formally published by Wilkinson et al. (2016). Unlike other initiatives that prioritize the human researchers, the FAIR principles emphasize the machine automation to find, use, and reuse data. This ensures that data remains both comprehensible to human researchers and actionable by computational agents, addressing the challenges posed by the expansive scale and intricacy of modern scientific data.

The FAIR data principles are not the first initiative as data management principles experienced great evolution in the history. With the increasing volume of data generated across multiple domains, there arose an urgent need to enhance the infrastructure supporting data reuse. Recognizing this, a diverse group of stakeholders from academia, industry, funding agencies, and publishers designed the FAIR Data Principles. Since the old understanding of data principles cannot operate efficiently at the scale demanded by modern e-Science, making data understandable and actionable for machines becomes a necessity.

The FAIR principles have proven of great importance in data management across various domains. Their universality ensures that data, even when poorly produced, does not make it harder to reuse in scientific progress. Moreover, with the increasing importance of data transparency and integrity by publishers and institutions, ensuring data adheres to the FAIR principles has become crucial.

Community has contributed greatly in understanding and implementing these principles. The RDA FAIR Data Maturity Model by Bahim et al. (2020) has the main contribution of building standards and approaches for FAIR data, which can identify and assess the FAIRness of digital objects. Such work encourages research communities to create tools for assessing their data's FAIRness. Notably, initiatives like the European Open Science Cloud underscore the importance of the FAIR principles in promoting Open Science policies, emphasizing wide and early knowledge sharing.

In conclusion, the FAIR principles and related work like the FAIR Data Maturity Model provide a robust framework to ensure the quality, understanding, and consistency of the datasets.

Interpretation and Implementation of FAIR Principles. While the FAIR principles provide a clear guideline, their interpretation can vary based on domain, discipline, and individual perspectives. The paper by Jacobsen, de Miranda Azevedo, et al. (2020) provides the interpretation and implementation considerations for each of the FAIR

principles. The authors discuss the importance of a common understanding of the scope, aim, and representative implementation choices for each FAIR principle to improve their stepwise application by different stakeholders.

For the Findable principle, the authors interpret it as requiring the use of persistent identifiers, metadata, and standardized vocabularies to enable data discovery. They also provide implementation considerations such as the use of metadata standards and the creation of metadata records.

For the Accessible principle, the authors interpret it as requiring data to be retrievable by humans and machines, and provide implementation considerations such as the use of open access licenses and the provision of access points.

For the Interoperable principle, the authors interpret it as requiring data to be structured in a way that allows for integration with other data sources, and provide implementation considerations such as the use of common data formats and the adoption of community standards.

For the Reusable principle, the authors interpret it as requiring data to be well-described and properly formatted to enable reuse, and provide implementation considerations such as the use of data repositories and the creation of data management plans.

Overall, this paper provides a valuable resource for FAIR principles implementation and assists accelerated global participation and convergence towards accessible, robust, widespread and consistent FAIR implementations.

FAIR Assessment: Methodologies and Tools. Assessing the FAIRness of datasets is as vital as the adoption of FAIR principles themselves. In this subsection, two papers are reviewed to provide insights on FAIR assessment methodologies.

The first paper by Devaraju et al. (2021) presents practical solutions, methodologies, and tools developed by the FAIRsFAIR project to pilot the FAIR assessment of research data objects in trustworthy data repositories.

To support the FAIR assessment, the paper proposes a minimum set of core metrics for the FAIR assessment of research data, building on existing work, including RDA outputs and evaluated and refined through several iterations. The metrics are mainly built on the indicators developed by the RDA FAIR Data Maturity Model Working Group. The paper also presents two applications of the metrics: an awareness-raising self-assessment tool and an automated FAIR data assessment tool. The FAIR-Aware online self-assessment tool aims at raising researchers' awareness about the value of making data FAIR before depositing into a repository. The paper also suggests future improvements, including the next steps to enable FAIR data assessment in the broader research data ecosystem. Overall, the methodologies and tools presented in this paper provide practical solutions for assessing the FAIRness of research data objects.

While the first paper focus on both the methodology and a example tool, the second paper by Sun, Emonet, and Dumontier (2022) provides a comprehensive comparison of three automated FAIRness evaluation tools, F-UJI, the FAIR Evaluator, and FAIR Checker. The FAIR Principles have gained broad endorsement by funding agencies and political entities in the scientific community, but they are largely aspirational and do not specify technical requirements that can be unambiguously evaluated. As a result, there are different approaches to evaluate the FAIRness of digital objects, including manual, semi-automatic, and automatic methods. The study examines three aspects of the automated tools: tool characteristics, evaluation metrics, and metrics tests for three public datasets. The results show significant differences in the evaluation results for tested resources, along with differences in the design, implementation, and documentation of the evaluation metrics and platforms. While automated tools do test a wide breadth of technical expectations of the FAIR principles, the study puts forward specific recommendations for their improved utility, transparency, and interpretability. The study concludes that future work could focus on standardized benchmarks to critically evaluate the functioning of these and future FAIRness evaluation tools.

Making Data FAIR: Practical Workflows and Tools. Making data FAIR often involves a structured workflow. In this subsection, two papers are reviewed to show the step-by-step process, from understanding the dataset's current state to its eventual FAIR transformation.

The first paper by Jacobsen, Kaliyaperumal, et al. (2020) presents a practical workflow for making data FAIR, which involves a step-by-step process that can be applied to any type of data. The workflow is divided into three phases: Pre-FAIRification, FAIRification, and Post-FAIRification, each with specific steps to ensure that data is FAIR. The authors also discuss various tools that can aid in the FAIRification process, including FAIR metrics, infrastructure, and data stewardship. They emphasize the importance of multidisciplinary teams and hands-on FAIRification workshops and projects, such as rare disease patient registries, in advancing the FAIR data principles.

This paper provides a comprehensive guide for making data FAIR and highlights the growing importance of FAIR data in enabling efficient and error-free analysis of data from multiple sources by machines and humans alike.

While the first paper proposed one specific workflow and one specific tool to make data FAIR, the second paper by Thompson, Burger, Kaliyaperumal, Roos, and da Silva Santos (2020) provided a broader view for the workflow and several compatible tools for making data FAIR. A general workflow includes the following steps: data management planning, data production, data publication, data evaluation, and data finding and reuse. The authors also introduce several FAIR tools and resources that can be used to support each step of the workflow, such as the FAIR Data Point (FDP) for demonstrating compliance with FAIR principles, Data Stewardship Wizard for creating domain-relevant data management plans, and FAIRsharing for finding and reusing FAIR data.

Case Studies of FAIR Implementations. When conducting literature review, the case of FAIRshake attracts me by its mature design and capability. Therefore, a case study is included to provide a roadmap for upcoming design and implementation of my

research project.

FAIRshake is a toolkit designed to evaluate the FAIRness of research digital resources by Clarke et al. (2019). The platform was developed to meet the demands of the biomedical research community and integrates a number of community-accepted standards, including RDF, DATS, SmartAPI, and schema.org. FAIRshake is capable of facilitating FAIR assessments of a diverse set of digital objects, including datasets, tools, repositories, and APIs.

The platform enables the community to study the FAIRness of the resources they produce and use, and it is expected to become more automated as the FAIR metrics converge. FAIRshake is flexible enough to facilitate other related applications, such as scientific peer review.

However, there are some challenges and limitations associated with using FAIRshake. Before beginning to use the platform, users must have some training about concepts like FAIR metrics and rubrics. Additionally, many established community standards are not being employed within the biomedical research community, largely due to a lack of awareness.

According to the paper, FAIRshake has the potential to benefit researchers and institutions by providing a standardized way to evaluate the FAIRness of digital resources, which can improve their discoverability, accessibility, and reusability.

From the discusion of the paper, challenges and limitations associated with using the platform exist. The insight from this paper is that guidance of user interface and the design of the process are worth to be considered as important problems in my research project and to be resolved and optimized to ensure its effectiveness in the assessment process.

Comparisons with Other Principles. While FAIR data principles gained popularity in research data reusability research, another set of principles that has been influential in data management is the Linked Open Data (LOD) principles introduced by Bizer, Heath, and Berners-Lee (2008). In the review of the paper by Hasnain and

Rebholz-Schuhmann (2018), the LOD principles focus on the quality of data based on its accessibility, format, structures, and interoperability with other data sources across the web. Comparatively, the FAIR principles, while introduced for similar reasons, emphasize more on reusability. In assessing the FAIR principles against the LOD principles, it becomes evident that while FAIR incorporates aspects of LOD, it also extends and refines them in various ways.

In comparison of FAIR and LOD, there are advantages and limitations for each principle. LOD principles prioritize open data, guiding providers to improve data accessibility and reusability. FAIR, however, emphasizes clear licensing agreements, making it versatile for varied data scenarios. FAIR also consider metadata important for enhanced interoperability, which is the same with LOD's view on interoperable data.

Related Software Engineering Technologies

Django: Web Framework. Django is a high-level Python web framework that encourages rapid development and clean, pragmatic design. Known for its simplicity and robustness, Django provides a set of tools and features that enable developers to build secure, scalable, and maintainable web applications. In this proposal, Django serves as the backbone for developing web interfaces and handling backend processes. Its ORM (Object-Relational Mapping) capabilities, along with security features, make it an ideal choice for managing database interactions and ensuring data integrity.

Schema.org: Format of Structured Data. Schema.org is a collaborative community activity with a mission to create, maintain, and promote schemas for structured data on the Internet. By using Schema.org annotations, webmasters can embed structured data on their web pages, enhancing the discoverability and understanding of their content by search engines. In this proposal, the use of Schema.org formats will enable the standardization of data presentation across different parts of the system, facilitating better data integration and retrieval.

Globus: Service of Data Publication and Discovery. Globus is a cloud-based service designed to enable the secure, reliable, and easy publication and discovery of large-scale datasets. Its capabilities are particularly useful in research environments where data sharing and collaboration are critical. For this project, Globus will be leveraged to manage the publication and distribution of datasets, ensuring efficient data transfer, and access control, thereby enhancing collaborative research efforts.

Message Queue: Asynchronous service-to-service communication. Message Queues provide a method for asynchronous service-to-service communication in distributed systems. By decoupling services and facilitating message passing, they enable scalable and resilient architecture designs. This proposal will incorporate a Message Queue system to handle communication between different services, improving system responsiveness and reliability. This approach is particularly beneficial in handling peak loads and providing a buffer for service outages or slowdowns.

Design and Methodology

The design of the GeoEDF project is centered around a modular and scalable architecture, which allows for efficient handling and processing of geospatial data. The core components of the project's design include:

Modular Architecture

The project leverages a modular architecture, where each component - from metadata extraction to data publication - is a distinct module. This design allows for easy maintenance, scalability, and integration of new features or tools in the future.

Metadata Extraction and Management

A primary focus of the project is on the extraction and management of metadata. This involves developing scripts and tools to automatically extract relevant information

from geospatial files and to format this data according to the standards set by schema.org. The metadata extraction process is containerized, enhancing its portability and scalability.

Data Publication and Indexing

For data publication, the project utilizes the Globus platform to manage and index the data. This includes setting up data portals and ensuring that the data is easily discoverable and accessible. We also integrate tools for assigning persistent identifiers like DOIs to datasets, further enhancing their discoverability.

Workflow Automation and Message Queuing

To streamline and automate the data publication process, the project incorporates RabbitMQ for message queuing. This setup allows for asynchronous communication and task management across different services, ensuring efficient workflow management.

FAIR Data Principles Compliance

A key aspect of the project is its adherence to the FAIR data principles. This is achieved through careful design considerations that ensure data Findability, Accessibility, Interoperability, and Reusability. The integration of a FAIR evaluator tool helps in assessing our compliance with these principles.

Methodology

Development Approach

The development approach for the GeoEDF project is iterative and agile. This allows for continuous integration of feedback and improvements throughout the project lifecycle. Regular testing and evaluation are integral to this approach, ensuring that each module and feature meets the desired standards and requirements.

Deployment Strategy

For deployment, we use containerization tools like Docker, which provide a consistent and reliable environment for our applications. This strategy ensures that our services are easily deployable and maintainable across different platforms and infrastructures.

User-Centric Design

In developing the front-end components, particularly the resource landing pages and search interfaces, a user-centric design approach is adopted. This ensures that the portal is intuitive and easy to use, enhancing the user experience for researchers and collaborators.

Data Security and Privacy

Considering the sensitivity of some geospatial data, the project incorporates robust data security and privacy measures. This includes secure data transfer mechanisms, authentication protocols, and access control measures to safeguard the data and user privacy.

In conclusion, the design and methodology of the GeoEDF project are geared towards creating a robust, scalable, and user-friendly platform for geospatial data management. Our approach is rooted in best practices of software engineering and data management, ensuring that the project not only meets the current requirements but is also adaptable for future expansions and enhancements.

Conclusion

In this proposal draft, we reviewed the works on the FAIR data principles, examining their significance and application in contemporary research. The review covers the origins, practical workflows, tools, and case studies, and made comparisons with other guiding principles. The information from this review shows the pivotal role of FAIR principles in enhancing data management and research outcomes. Apart from these paper explicitly

cited in the previous sections, I also read the work by Devaraju et al. (2020) which is a report on FAIR data assessment mechanisms at the dataset level, the work by Kaliuzhna and Altemeier (2021) which is a report on a series of workshops on FAIR data principles, the work by wwPDB consortium (2018) which is a parallel comparison about an example of building data bank without using FAIR, the work by Koers et al. (2020) about the recommendation system using FAIR data, the work by Wilkinson, Dumontier, Sansone, da Silva Santos, et al. (2019), the work by Queralt-Rosinach et al. (2022), the work by Inau, Sack, Waltemath, and Zeleke (2021), the work by Wilkinson, Dumontier, Sansone, Bonino da Silva Santos, et al. (2019), the work by RDA FAIR Data Maturity Model Working Group et al. (2020), and the work by Vesteghem et al. (2020).

Besides, there are knowledge not covered in this review, such as how these principles work with the Django framework. Future work will include getting knowledge from these topics as the implementation of the project is important as well.

References

Bahim, C., Casorrán-Amilburu, C., Dekkers, M., Herczog, E., Loozen, N., Repanas, K., ... Stall, S. (2020). The fair data maturity model: An approach to harmonise fair assessments.

Bizer, C., Heath, T., & Berners-Lee, T. (2008). Linked data: Principles and state of the art. In *World wide web conference* (Vol. 1, p. 40).

Clarke, D. J. B., Wang, L., Jones, A., Wojciechowicz, M. L., Torre, D., Jagodnik, K. M., ... Ma'ayan, A. (2019). Fairshake: Toolkit to evaluate the fairness of research digital resources. *Cell systems*. Retrieved from `https://api.semanticscholar.org/CorpusID:207896394`

Devaraju, A., Mokrane, M., Cepinskas, L., Huber, R., Herterich, P., de Vries, J., ... Diepenbroek, M. (2021). From conceptualization to implementation: Fair assessment of research data objects. *Data Science Journal*, *20*(1), 1–14.

Devaraju, A., Mokrane, M., Cepinskas, L., Huber, R. A., Herterich, P., de Vries, J., ... Diepenbroek, M. (2020). M4.9 report on fair data assessment mechanisms to develop pragmatic concepts for fairness evaluation at the dataset level.. Retrieved from `https://api.semanticscholar.org/CorpusID:236816987`

Hasnain, A., & Rebholz-Schuhmann, D. (2018). Assessing fair data principles against the 5-star open data principles. In *The semantic web: Eswc 2018 satellite events: Eswc 2018 satellite events, heraklion, crete, greece, june 3-7, 2018, revised selected papers 15* (pp. 469–477).

Inau, E. T., Sack, J., Waltemath, D., & Zeleke, A. A. (2021). 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*(2), e22505.

Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., ... others (2020). *Fair principles: interpretations and implementation considerations*

(Vol. 2) (No. 1-2). MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info ....

Jacobsen, A., Kaliyaperumal, R., da Silva Santos, L. O. B., Mons, B., Schultes, E., Roos, M., & Thompson, M. (2020). A generic workflow for the data fairification process. *Data Intelligence*, *2*(1-2), 56–65.

Kaliuzhna, N., & Altemeier, F. (2021). Towards fair principles for research information: Report on a series of workshops.. Retrieved from https://api.semanticscholar.org/CorpusID:238810298

Koers, H., Bangert, D., Hermans, E., van Horik, R., de Jong, M., & Mokrane, M. (2020). Recommendations for services in a fair data ecosystem. *Patterns*, *1*. Retrieved from https://api.semanticscholar.org/CorpusID:214073040

Queralt-Rosinach, N., Kaliyaperumal, R., Bernabé, C. H., Long, Q., Joosten, S. A., van der Wijk, H. J., ... others (2022). Applying the fair principles to data in a hospital: challenges and opportunities in a pandemic. *Journal of biomedical semantics*, *13*(1), 12.

RDA FAIR Data Maturity Model Working Group, B., et al. (2020). Fair data maturity model: specification and guidelines. *Res. Data Alliance*, *10*.

Sun, C., Emonet, V., & Dumontier, M. (2022). A comprehensive comparison of automated fairness evaluation tools. In *Swat4hcls.* Retrieved from https://api.semanticscholar.org/CorpusID:248396324

Thompson, M., Burger, K., Kaliyaperumal, R., Roos, M., & da Silva Santos, L. O. B. (2020). Making fair easy with fair tools: From creolization to convergence. *Data Intelligence*, *2*, 87-95. Retrieved from https://api.semanticscholar.org/CorpusID:207828580

Vesteghem, C., Brøndum, R. F., Sønderkær, M., Sommer, M., Schmitz, A., Bødker, J. S., ... Bøgsted, M. (2020). Implementing the fair data principles in precision oncology: review of supporting initiatives. *Briefings in bioinformatics*, *21*(3), 936–945.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... others (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1–9.

Wilkinson, M. D., Dumontier, M., Sansone, S.-A., Bonino da Silva Santos, L. O., Prieto, M., Batista, D., ... others (2019). Evaluating fair maturity through a scalable, automated, community-governed framework. *Scientific data*, *6*(1), 174.

Wilkinson, M. D., Dumontier, M., Sansone, S.-A., da Silva Santos, L. O. B., Prieto, M., Batista, D., ... Schultes, E. A. (2019). Evaluating fair maturity through a scalable, automated, community-governed framework. *Scientific Data*, *6*. Retrieved from `https://api.semanticscholar.org/CorpusID:190876394`

wwPDB consortium. (2018). Protein data bank: the single global archive for 3d macromolecular structure data. *Nucleic Acids Research*, *47*, D520 - D528. Retrieved from `https://api.semanticscholar.org/CorpusID:53027084`