Automating Data Management Plan Evaluation Through the Integration of Machine-Actionable Data Management Plans and Fair Implentation Profiles

Master Thesis Proposal

Student: Bernardo Aceves Partida (12143531)
Main Supervisor: Dr.techn. Mag. Tomasz Miksa
Master Program: Data Science (066 645)

March 2025

1 Context

Research Data Management (RDM)[1] is a fundamental aspect of modern scientific research, ensuring that data is properly documented, stored, and shared for long-term accessibility and reuse. One of the primary tools for organizing research data is the use of Data Management Plans (DMP)[2], a structured document outlining how data will be handled throughout the research lifecycle. Traditional DMPs, however, are static and require manual updates, limiting their adaptability and effectiveness.

To address these limitations, machine-actionable Data Management Plans (maDMPs)[3] have been introduced. Unlike traditional DMPs, maDMPs[4] enable automated updates, integration with external systems, and dynamic evaluation, enhancing research efficiency. Furthermore, the FAIR Implementation Profiles (FIPs)[5] have been developed as structured, community-driven recommendations that specify how FAIR principles[6] (Findable, Accessible, Interoperable, and Reusable) should be implemented in specific domains, enabling standardized yet customizable FAIR implementation strategies across different scientific disciplines. They serve as a formalized consensus among research communities, outlining the best practices for ensuring the FAIRness of data and are also designed to be machine-actionable, facilitating automated processes in data management[7].

Wilkinson et al. introduced the FAIR guiding principles to address the urgent need for improved infrastructure supporting the reuse of scholarly data. These principles, endorsed by stakeholders from academia, industry, funding agencies, and publishers, serve as guidelines to enhance data reusability. Unlike other initiatives focused on human researchers, FAIR emphasizes machine-readability, enabling automated data discovery and reuse. Their work formally established the principles and provided a rationale and real-world implementations within the research community[8]

Since the introduction of the FAIR principles, various institutions, such as the Research Data Alliance (RDA)[9] and the Committee on Data of the International Science Council (CODATA)[10], have promoted their adoption. However, achieving compliance with FAIR guidelines is challenging due to their flexible nature and the absence of strict implementation rules. Wilkinson et al. emphasize that the FAIR principles are guiding concepts rather than fixed standards, allowing for diverse implementations. To support researchers, initiatives such as FAIRsFAIR have proposed structured metrics to assess FAIR compliance. Additionally, tools like F-UJI[11] offer automated evaluations of dataset adherence to FAIR standards.[12]

Despite these advancements, a significant gap remains: there is no clear framework aligning FIPs with maDMPs for automated DMP evaluation. This research aims to bridge that gap by integrating FIPs with maDMPs, enabling a systematic and automated approach to evaluating research data management practices.

2 Problem Statement

Despite the push for FAIR data, many issues persist in research data management (RDM), limiting the efficiency and reusability of datasets; for example:

- **Data is poorly documented, making reuse difficult**.
  Without structured metadata and proper documentation, researchers struggle to interpret and reuse existing datasets, leading to inefficiencies in scientific collaboration and discovery. Well structured metadata could make research data more reusable and interoperable.[8]

- **Licensing is often missing or unclear, creating barriers to sharing**.
  Ambiguous or missing data licenses prevent datasets from being freely accessed and reused, leading to legal uncertainties and compliance issues. Mons et al. (2017)[13] discuss how poorly defined licensing can hinder open data initiatives.

- **Manual DMP evaluations are time-consuming and inconsistent.**
  Assessing the FAIRness of DMPs manually is inefficient and prone to inconsistencies, making large-scale evaluations unfeasible. Devaraju and Huber (2021)[14] propose automated solutions to reduce evaluation time and improve consistency in assessing FAIR research data.

- **Existing solutions (like README files) provide minimal guidance and are non-standardized.**
  README files are commonly used for dataset documentation but lack structured guidance for ensuring FAIR compliance. They do not provide a standardized approach to making data findable, accessible, and reusable. [13]

Integrating FIPs into maDMPs for automated evaluation remains an open challenge. While initiatives like FIP2DMP have begun mapping FIPs to DMP templates, a standardized framework for seamless integration is still lacking.[15]

Also, the creation of the DMP Common Standard Ontology (DCSO)[16] serialization offers a structured way to represent maDMPs using a controlled Resource Description Framework (RDF)[17] vocabulary. This allows for the application of semantic web technologies within the maDMP framework.

Additionally, determining the right benchmarks, metrics, and tests for different research contexts is a challenging task. Initiatives like FAIR-IMPACT are working to refine minimum viable metrics and develop structured assessment tools to better evaluate FAIR data objects.[18]

This research aims to bridge this gap by aligning FIPs with maDMPs and offer a scalable solution for automated DMP evaluation. The DMP Common Standards working group has developed an application profile for maDMPs to facilitate such integrations[19].

3 Goals and Expected Outcome

The goal of this research is to offer a scalable solution for automated DMP evaluation by integrating FIPs with maDMPs. Research has been conducted to develop automated quality indicators for maDMPs to assist in their evaluation.[20]

3.1 Goals

- Create a conceptual framework that defines how to integrate FIPs with maDMPs at a general level, ensuring applicability across different implementations.
- Utilize the integrated framework to facilitate systematic and automated evaluation of the FAIRness of datasets described by DMPs.
- Establish quantifiable metrics and automated tests to evaluate FAIR compliance, ensuring that assessments cover key aspects such as metadata quality, licensing clarity, and interoperability. These metrics should be adaptable across different domains and datasets. [18]

3.2 Success Criteria

To measure the success of this thesis, key success indicators include:

- **Efficiency** – Reduce the time required for DMP evaluation compared to manual review processes.
- **Alignment with Established Success Goals**: Ensure that the framework satisfies the four key success criteria outlined in Arnhold, L.[20, p. 2-5]:
    1. **Completeness:** Evaluate whether all relevant aspects of a maDMP are covered.
    2. **Feasibility:** It should be practical and implementable within existing research workflows, without imposing excessive burdens on researchers or institutions.
    3. **Quality of Actions:** The framework must promote high-quality data management practices, leading to improved data integrity and usability.
    4. **Guideline Compliance:** It should align with relevant guidelines and standards, ensuring that data management plans meet established criteria and expectations.[21]

3.3 Expected Outcomes

- **Conceptual Framework:** A well-defined, implementation- framework detailing the integration of FIPs with maDMPs, serving as a foundation for automated DMP evaluation.
- A mapping model linking FIP components to maDMP fields.

- A structured evaluation system that measures the FAIRness of DMPs and returns a score.
- An automated evaluation tool that generates guidance for improvement. [20, p. 5]

4 Research Questions

Research question 1: To what extent can maDMPs and FIPs be aligned to enable automated evaluation of Data Management Plans?

Aligning maDMPs with FIPs requires structuring metadata in a way that enables automated FAIR compliance assessments. This research will investigate to what extent can maDMP components be mapped to predefined FIP criteria to establish a consistent and machine-readable evaluation framework. The study will evaluate whether this integration enhances DMP assessments compared to current manual review processes.

Research question 2: What is the optimal way to utilize FIPs to evaluate the FAIRness of DMPs during the planning phase, particularly when data has not yet been generated or lacks repository identifiers?

FIPs define community-specific standards for making data FAIR. Since early-stage DMPs describe intended data management actions rather than existing datasets, assessing their FAIRness requires a structured method for comparison. This research question focuses on exploring whether mapping planned DMP decisions to established FIP criteria can serve as a reliable assessment method. By analyzing how closely planned actions align with the best practices outlined in FIPs, we aim to determine whether such an approach can predict the likelihood of future FAIR compliance and provide useful feedback to researchers.

5 Research Methods

The core of this thesis is the creation of a framework for evaluating DMPs by integrating maDMPs with FIPs. This research seeks to offer a way to assess compliance with FAIR principles, that is scalable and efficient by using semantic technologies, benchmarks and Scientific Knowledge Graphs (SKG)[22], improving data sharing and documentation and reducing manual effort.

To achieve this, the author proposes the following approach:

1. **Literature Review:**

   The first step involves a comprehensive review of existing research on FIPs, maDMPs, and FAIR assessment tools. This will help establish the foundation for the framework by understanding how FAIR principles are currently evaluated and identifying gaps in existing methodologies. The review will also examine the challenges of assessing DMPs at different stages of the research lifecycle and explore potential solutions for automating FAIR assessments of DMPs.

2. **Framework Design:**

   Building on the insights from the literature, the next step is to design a conceptual framework that aligns FIP components and maDMP attributes[23]. This will define how DMPs can be assessed before and/or after datasets are created. The goal is to ensure that the framework supports continuous DMP evaluation, providing useful feedback at both early and late stages of a research project.

3. **Implementation:**

   This phase will focus on transforming the conceptual framework into a working prototype, ensuring that the integration of FIPs with maDMPs is functional and scalable. This will involve developing a structured evaluation system that defines FAIR benchmarks and metrics, enabling systematic assessment of DMP adherence to FAIR principles.

4. **Testing & Validation:**

   The last phase will assess whether the implemented framework aligns with FAIR principles and effectively supports automated DMP evaluation. This will involve applying FAIR assessment metrics to measure its impact on both planned and existing research data management projects(early and final stages). The system's performance will be validated by running evaluations on real-world maDMP datasets and comparing the results to manual expert assessments to determine its accuracy and reliability. Additionally, if needed, a custom FIP could be developed for a specific domain, for example, for the TU Wien. This could to test the framework's adaptability in defining community-specific FAIR guidelines

6 State of the Art

The evaluation of DMPs has considerably evolved, changing from static, manually reviewed documents to dynamic, machine-actionable formats. maDMPs enable automated processing, necessitating further research into their integration with structured FAIR assessment tools.[24] While existing approaches have predominantly focused on assessing datasets, there has been less emphasis on structured plans for data management. Additionally, the review process of DMPs requires specific expert knowledge in data stewardship, which is scarce, making automated solutions increasingly necessary. Miksa et al. (2019)[4] discuss how machine-actionable frameworks can help address this challenge by streamlining DMP assessments while reducing dependency on human experts.

FIPs have emerged as a standard for domain-specific FAIR best practices. These profiles provide community-driven guidelines on making data FAIR but are not yet well-integrated with the automation capabilities of maDMPs.[25] The GO FAIR initiative highlights how FIPs allow communities to define structured best practices for FAIR data management, yet their potential to assist in automated DMP evaluation remains underexplored.[5].

Recent developments in automated FAIR assessment tools, such as F-UJI and FAIR Evaluator, have improved the ability to measure dataset compliance with FAIR principles. However, these tools only focus on assessing existing datasets and do not address the need for structured evaluation frameworks that can assess FAIR compliance in early research stages, particularly when data has not yet been generated or deposited in repositories. [26] As a result, there is a gap in ensuring that research projects incorporate FAIR principles from the very beginning. By aligning FIPs with maDMPs, this research aims to fill this gap by enabling automated evaluation and structured guidance for researchers.

7 Relevance to the Curriculum

The scope of this master's thesis covers several key aspects of the Data Science curriculum. The author considers the content of the following courses to be highly relevant for the success of this research.

- 194.044 Data Stewardship
- 194.048 Data-Intensive Computing
- 184.772 Description Logics and Ontologies
- 192.116 Knowledge Graphs
- 188.995 Data-oriented Programming Paradigms
- 188.399 Introduction to Semantic Systems
- 184.780 Advanced Database Systems
- 188.992 Experiment Design for Data Science

Data stewardship plays a central role in this research, as it focuses on ensuring that research data is properly managed, shared, and evaluated for compliance with FAIR principles.[27] By integrating automation techniques with data stewardship best practices, this thesis contributes to the broader goal of improving research data accessibility and reusability.

8 References

[1] Science Europe. Research data management. URL https://scienceeurope.org/our-priorities/open-science/research-data-management/. Accessed: 4 Mar. 2025.

[2] TU Wien. Research data management: Dmp information and tips. URL https://www.tuwien.at/en/research/rti-support/research-data/rdm-infos-tips/dmp/. Accessed: 7 Mar. 2025.

[3] Stephanie Simms, Sarah Jones, Daniel Mietchen, and Tomasz Miksa. Machine-actionable data management plans (madmps). *Research Ideas and Outcomes*, 3:e13086, 2017. doi:10.3897/rio.3.e13086. URL https://doi.org/10.3897/rio.3.e13086.

[4] T. et al. Miksa. Ten principles for machine-actionable data management plans. *PLOS Computational Biology*, 15(3):e1006750, 2019. doi:10.1371/journal.pcbi.1006750.

[5] Erik Schultes, Barbara Magagna, Kristina Maria Hettne, Robert Pergl, Marek Suchánek, and Tobias Kuhn. Reusable fair implementation profiles as accelerators of fair convergence, 2020. URL https://osf.io/preprints/osf/2p85g_v1.

[6] GO FAIR. Fair principles, . URL https://www.go-fair.org/fair-principles/.

[7] GO FAIR. Metadata for machines, . URL https://www.go-fair.org/how-to-go-fair/metadata-for-machines/.

[8] Mark D. Wilkinson et al. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018, 2016. doi:10.1038/sdata.2016.18. URL https://www.nature.com/articles/sdata201618.

[9] Research Data Alliance. About the rda. URL https://www.rd-alliance.org/about-the-rda/.

[10] International Science Council. Committee on data (codata). URL https://council.science/member/committee-on-data-codata/. Accessed: 15 Mar. 2025.

[11] FAIRsFAIR. F-uji automated fair data assessment tool. URL https://www.fairsfair.eu/f-uji-automated-fair-data-assessment-tool.

[12] Martin Boeckhout, Gerhard A Zielhuis, and Annelien L Bredenoord. The fair guiding principles for data stewardship: fair enough? *European journal of human genetics*, 26(7):931–936, 2018. doi:https://doi.org/10.1038/s41431-018-0160-0.

[13] Barend Mons, Cameron Neylon, Jan Velterop, Michel Dumontier, Luiz Olavo Bonino da Silva Santos, and Mark D. Wilkinson. Cloudy, increasingly fair; revisiting the fair data guiding principles for the european open science cloud. *Information Services and Use*, 37(1):49–56, 2017. doi:10.3233/ISU-170824. URL https://doi.org/10.3233/ISU-170824.

[14] Anusuriya Devaraju and Robert Huber. An automated solution for measuring the progress toward fair research data. *Patterns*, 2(11):100370, 2021. doi:10.1016/j.patter.2021.100370. URL https://pubmed.ncbi.nlm.nih.gov/34820651/.

[15] Kristina Maria Hettne, Barbara Magagna, Alessa An Gambardella, Marek Suchánek, Fieke Schoots, and Erik Schultes. Fip2dmp: Linking data management plans with fair implementation profiles. *FAIR Connect*, 1(1):23–27, 2023. doi:10.3233/FC-221515. URL https://journals.sagepub.com/doi/abs/10.3233/FC-221515.

[16] RDA DMP Common Standards WG. Dmp common standard ontology (dcso), . URL https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard/tree/master/ontologies.

[17] Jeff Z. Pan. *Resource Description Framework*, pages 71–90. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-540-92673-3. doi:10.1007/978-3-540-92673-3_3. URL https://doi.org/10.1007/978-3-540-92673-3_3.

[18] FAIR-IMPACT. Metrics for data. URL https://fair-impact.eu/metrics-data.

[19] Tomasz Miksa, Paul Walk, Peter Neish, Simon Oblasser, Hollydawn Murray, Tom Renner, Marie-Christine Jacquemot-Perbal, João Cardoso, Trond Kvamme, Maria Praetzellis, Marek Suchánek, Rob Hooft, Benjamin Faure, Hanne Moa, Adil Hasan, and Sarah Jones. Application profile for machine-actionable data management plans. *Data Science Journal*, 20:1–24, 2021. doi:10.5334/dsj-2021-032. URL https://doi.org/10.5334/dsj-2021-032.

[20] Lukas Arnhold. Automated quality indicators for machine-actionable data management plans. Master's thesis, TU Wien, 2024. URL https://repositum.tuwien.at/handle/20.500.12708/200466.

[21] RDA DMP Common Standards WG. Rda dmp common standard for machine-actionable data management plans, . URL https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard.

[22] Scientific Knowledge Graphs Workshop. Scientific knowledge graphs workshop 2020, 2020. URL https://sci-k.github.io/SKG2020/.

[23] G. et al. Guizzardi. Towards ontological foundations for conceptual modeling: The unified foundational ontology (ufo) story. *Applied Ontology*, 10(3-4):259–271, 2015. doi:10.3233/AO-150157.

[24] GO FAIR. Fairsfair implementation story: Leveraging machine-actionable dmps to build rdm infrastructures, . URL https://www.go-fair.org/2022/05/12/fairsfair-implementation-story-leveraging-machine-actionable-dmps-to-build-rdm-

[25] Kristina Maria Hettne, Barbara Magagna, Alessa An Gambardella, Marek Suchánek, Fieke Schoots, and Erik Schultes. Fip2dmp: Linking data management plans with fair implementation profiles. *FAIR Connect*, 1(1):23–27, 2023. doi:10.3233/FC-221515. URL https://journals.sagepub.com/doi/abs/10.3233/FC-221515.

[26] Miksa Tomasz M. Slifka J. Knaisl V. Ekaputra F.J. Kovacevic F. Ningtyas A.M. El-Ebshihy A. T., Suchánek and R. Pergl. Towards a toolbox for automated assessment of machine-actionable data management plans. *Data Science Journal*, 21:28, 2023. doi:10.5334/dsj-2023-02. URL https://doi.org/10.5334/dsj-2023-028.

[27] S. et al. Blumesberger. Fair data austria – aligning the implementation of fair tools and services. *Mitteilungen der VÖB*, 74(2):102–113, 2021. doi:10.31263/voebm.v74i2.6379.