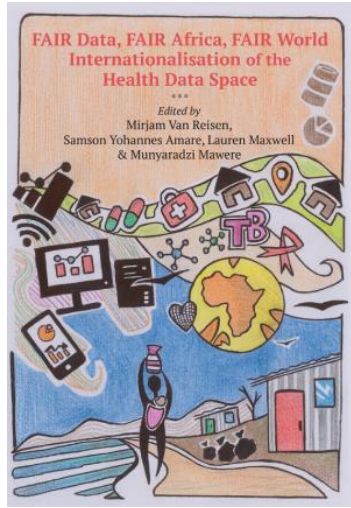
# De Novo FAIRification: A Literature Review

*Zhengyu Lin*

**Chapter in:**
Fair Data Fair Africa Fair World:
Internationalisation of the Health Data Space




Cite as: Lin, Z. (2025). De Novo FAIRification: A Literature Review. https://doi.org/10.5281/zenodo.15382861. In Van Reisen, M., Amare, S. Y., Maxwell, L. & Mawere, M. (Eds.), *FAIR data, FAIR Africa, FAIR world: Internationalisation of the Health Data Space*. (pp. 143–156). Bamenda: Langaa. URL: https://www.researchgate.net/publication/391750151_FAIR_Data_FAIR_Africa_FAIR_World_The_Internationalisation_of_the_Health_Data_Space

The Editorial Note can be found here: https://raee.eu/wp-content/uploads/2025/05/About-the-Authors-and-Editors.pdf

The list of figures can be found here: https://raee.eu/wp-content/uploads/2025/05/List-of-Figures-and-Tables.pdf

# Contents

# Chapter 5

## De Novo FAIRification: A Literature Review

*Zhengyu Lin*

**Abstract**

Examining existing literature is essential for understanding the current state of De Novo FAIRification. This chapter aims to evaluate the body of literature on this topic, highlighting key approaches, challenges, and best practices. The study reveals that research on De Novo FAIRification is still in its early stages, with limited studies available. While the concept holds promise, more extensive research is needed to explore its full potential and practical implementation. The review discusses the limited available publications in extant literature regarding De Novo FAIRification. Currently, De Novo FAIRification has primarily focused on newly developed systems. Unlike post-hoc FAIRification, which requires significant manual effort and time to transform collected data into machine-readable formats, De Novo FAIRification minimises the disruption to ongoing workflows. Moreover, De Novo FAIRification can facilitate better collaboration and data sharing among researchers and institutions. By ensuring that data is findable, accessible, interoperable, and reusable from the outset, De Novo FAIRification supports open science initiatives and promotes transparency in research. This can lead to more robust and reproducible scientific findings, as well as increased innovation through the reuse of data across different disciplines. However, future research must also address its application to legacy systems, which present unique data management challenges. Expanding the scope to include legacy systems is crucial for ensuring that all systems, both old and new, can benefit from strong data quality consistency and provenance traceability. This will enhance the quality of the findability, accessibility, interoperability, and reusability of data across various contexts and applications.

**Keywords:** FAIR data, FAIRification workflows, De Novo FAIRification, FAIR by design, ad hoc FAIRification, FAIR by increment

## Introduction

FAIRification is making data findable, accessible, interoperable and reusable (FAIR) (de Oliveira, Borges, Rodrigues, Campos, & Lopes, 2022). The current method, known as post-hoc FAIRification, involves labour-intensive, semi-manual conversions of collected data into machine-readable formats after data collection (Kersloot et al., 2021). As a result, significant time and budget inefficiencies hinder the effective and timely implementation of FAIR principles in research workflows.

De novo FAIRification places a strong emphasis on automating the FAIR process in real time while gathering data (Groenen et al., 2021). Although De Novo FAIRification has several benefits, including the removal of labour-intensive processes, and time and cost savings, its application in real-world projects is not common (Groenen et al., 2021).

Examining the body of previous literature is crucial to comprehending the current state of De Novo FAIRification. The objective of this chapter is to evaluate the body of existing literature on the subject, highlighting important approaches, difficulties, and best practices that have been previously recorded. In order to provide a thorough grasp of the existing situation and identify any gaps that can be filled by further study, this review attempts to answer the research question: What is available in extant literature regarding De Novo FAIRification?

## Research design

This study adopts a literature review approach to investigate existing research on De Novo FAIRification. The literature review follows a systematic methodology to ensure a comprehensive and unbiased analysis of the available studies. Specifically, two complementary methods are employed: the systematic literature review and the snowballing approach.

A systematic literature review is conducted to ensure a broad yet targeted collection of literature while minimising selection bias. The snowballing approach is then applied to complement the systematic literature review by identifying additional studies through reference

tracking. Backward snowballing explores citations in selected papers, while forward snowballing identifies newer works that cite them, ensuring a more exhaustive review.

**Relevance**

Research on De Novo FAIRification is critical for promoting the widespread adoption of FAIR principles across various domains. Two primary benefits of this research are outlined below:

- Enhancing the Adaptability and Effectiveness of FAIRification Technology: As an innovative method of FAIRification, research on this method will help improve the adaptability and effectiveness of the FAIR principles, promote its successful application in different industries and projects, and provide valuable experience and guidance for future implementation.

- Automating the FAIRification Process for Real-Time or Near Real-Time Transformation: One of the key advantages of De Novo FAIRification is its focus on automating the FAIRification process in real-time or near real-time, ideally beginning at the data collection stage. Unlike post-hoc FAIRification, which requires significant manual effort and time to transform collected data into machine-readable formats, De Novo FAIRification minimises the disruption to ongoing workflows. This automation not only saves time and reduces budgetary constraints but also enables a more efficient and timely application of the FAIR principles, ensuring that data remains compliant and accessible from the moment it is gathered.

**Ethical, regulatory, and data management considerations**

Numerous issues, rules, and regulations relating to data processing are in place globally. Understanding the ideas of data governance and legal frameworks that have a big influence on this topic is essential to understanding how data can be processed. These frameworks' main goal is to create a uniform method for managing data in compliance with applicable rules and regulations (Plug et al., 2022).

The guidelines for protecting natural persons concerning the processing and free transfer of personal data are defined by the General Data Protection Regulation (GDPR) (GDPR, 2016).

According to Art.3 of GDPR, this Regulation covers the processing of personal data in connection with the operations of an establishment of a controller or processor within the Union, regardless of whether the processing takes place inside the Union or outside of it (GDPR, 2016).

- According to Art.4 of GDPR, 'Processing data' means any operation of personal data no matter what operation it is (GDPR, 2016).
- According to Art.4 of GDPR, 'Personal data' means all information about an identifiable natural person (GDPR, 2016).
- According to Art.4 of GDPR, 'Controller' refers to any individual or entity acting alone or collaboratively with others that decides how and why personal data is processed (GDPR, 2016).
- According to Art.4 of GDPR, 'Processor' refers to any individual, organisation, government agency, or other entity that handles personal data (GDPR, 2016).
- According to Art.4 of GDPR, 'Recipient' refers to anyone to whom personal data is supplied (GDPR, 2016).

Below are the key principles of GDPR:

- Lawfulness, fairness, and transparency: According to Art.5 of GDPR, data is handled in a way that is transparent, equitable, and compliant with the law (GDPR, 2016).
- Purpose limitation: According to Art.5 of GDPR, data can only be collected and processed for purposes that are consented to by the data subject (GDPR, 2016).
- Data minimisation: According to Art.5 of GDPR, it is only allowed to require and use the data that are necessary for the given purpose (GDPR, 2016).

- Accuracy: According to Art.5 of GDPR, data should be updated to ensure accuracy. Inaccurate data should be modified (GDPR, 2016).
- Integrity and confidentiality: According to Art.5 of GDPR, data should be processed securely, avoiding unlawful processing and loss of data (GDPR, 2016).
- Storage limitation: According to Art.5 of GDPR, data cannot be stored longer than the purpose needed time, except for the archiving purpose (GDPR, 2016).
- Accountability: According to Art.5 of GDPR, the controller is responsible for ensuring compliance with the principles above (GDPR, 2016).

**Theoretical framework**

FAIR principles stand for Findability, Accessibility, Interoperability, and Reuse of digital assets. The guiding principles emphasise machine-actionability. This is because humans rely more and more on computing assistance to handle mass data (FAIR Principles, 2022).

- Findable:

Findable means that both humans and computers can find metadata and data easily. In this process, machine-readable metadata is the core part (FAIR Principles, 2022).

Data that offer details about other data—typically a dataset—are known as metadata. These data assist in understanding the significance of the data in the datasets when combined with documentation. Since metadata offers a precise definition of the data that may be used to interpret them, they also facilitate interoperability, or the capacity to interchange and interpret data.

Global unique and persistent identifiers can ensure the unambiguity of the meaning of data or metadata (FAIR Principles, 2022). Rich metadata enhances the possibility of computers to approach and accomplish tasks, even without the data's identifier. This means that metadata should be extensive. It requires researchers to be generous and give metadata as much as possible (FAIR Principles, 2022).

Also, usually the dataset and the metadata are offered in separate files. The connection of these files should be clarified explicitly. In the metadata file, the annotation should be made clear by using the unique and persistent identifier of the dataset (FAIR Principles, 2022).

Additionally, the (meta)data should be discoverable, which means they need to be registered or indexed in a searchable resource. Otherwise, other people cannot access them (FAIR Principles, 2022).

- Accessible:

When people find the requested data, they should be able to access that data (FAIR Principles, 2022). People retrieve data via the internet. Therefore, a standardised communication protocol is necessary (FAIR Principles, 2022). Data can be either public or private, but the key lies in specifying accessibility conditions in a machine-readable format. This enables machines to automatically enforce the requirements or alert users to specific access rules. For instance, when selecting a data repository, access permissions can be managed through user accounts, ensuring both data security and privacy while adhering to the 'Accessibility' principle of FAIR (FAIR Principles, 2022).

Even though the dataset tends to disappear over time, metadata should be persistent. Also, compared to datasets, it is cheaper and easier to store metadata which ensures the feasibility of the persistent metadata (FAIR Principles, 2022).

- Interoperable:

Data should be interpretable by both humans and machines, avoiding the need for specialised tools or ad hoc translations. To ensure interoperability, the chosen language must have a formal specification with clearly defined syntax and grammar, be openly accessible for others to learn, and be versatile enough to apply across multiple scenarios (FAIR Principles, 2022). When we describe data or metadata, the vocabulary we use should also adhere to FAIR principles (FAIR Principles, 2022).

- Reusable:

To achieve the utmost objective of data reusability, precise and relevant attributes need to be attached to metadata and the usage license should be defined in an explicit way (FAIR Principles, 2022). The (meta)data should also adhere to the community standards of specific domains (FAIR Principles, 2022).

**Related literature**

De novo FAIRification emphasises the automation of the FAIR process in real-time during data collection. This innovative approach not only eliminates the labour-intensive post-hoc FAIRification operations, involving repeated, semi-manual transformation of collected data into RDF formats after data collection but also results in significant time and budget savings (Groenen et al., 2021).

The first project that successfully implemented the De Novo FAIRification is in the rare illness registry domain (Groenen et al., 2021). The researchers in this project created a technique for the De Novo FAIRification, ensuring that data is automatically made FAIR when the data is entered into an Electronic Data Capture (EDC) system (Groenen et al., 2021).

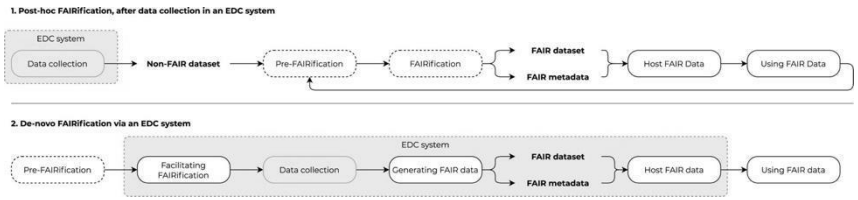



**Figure 1. The difference between post-hoc FAIRification and De Novo FAIRification**

Source: Kersloot et al., 2021

Figure 1 indicates the difference using the example of an EDC system.

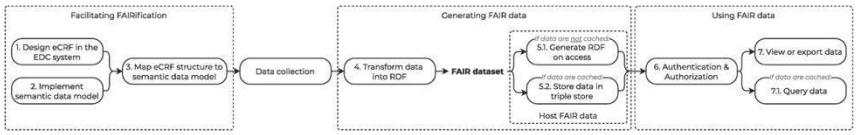



**Figure 2. Overview of the method**

Source: Kersloot et al., 2021

Figure 2 shows the method used in this project. The below part explains this method in depth.

- Design electronic Case Report Form (eCRF) in the Electronic Data Capture (EDC) system: The eCRFs are designed in the EDC system (Kersloot et al., 2021).
- Implement semantic data model: Based on the European Joint Programme on Rare Diseases (EJPRD), the Common Data Elements (CDEs) and the relationships between them were defined in a semantic data model (Kersloot et al., 2021).
- Map eCRF structure to the semantic data model: To apply the semantic data model to the eCRF data, each eCRF and its questions must be mapped to the model (Kersloot et al., 2021).
- Data transformation: The data transformation application created by EDC vendor Castor EDC performs the RDF transformation (Kersloot et al., 2021). To obtain the eCRFs from the EDC system, the data transformation application makes use of the Application Programming Interface (API) of the EDC system (Kersloot et al., 2021).
- Host FAIR data: There are two methods of storing data to support queries. One is generating access. For this method, when a user attempts to access the semantically modelled data, the Resource Description Framework (RDF) will be produced (Kersloot et al., 2021). Also, data retrieval is slower since each time the RDF is accessed, all patient data needs to be loaded and converted (Kersloot et al., 2021). The other method is cache. For this method, the data may be retrieved quickly since the creation procedure has already been completed (Kersloot et al., 2021).
- Perform authentication and authorisation: The user must log in to view the data if the semantically modelled data is not publicly released (Kersloot et al., 2021). The user will only see or get data if they can access the EDC system's eCRFs for the designated hospital or institute (Kersloot et al., 2021).

- View, export, or query data: The EDC system checks authentication when users try to access the data and then sends required information back (Kersloot et al., 2021).

**Results**

The results of the literature review using the snowballing approach and systematic method show insufficient research on the implementation of De Novo FAIRification.

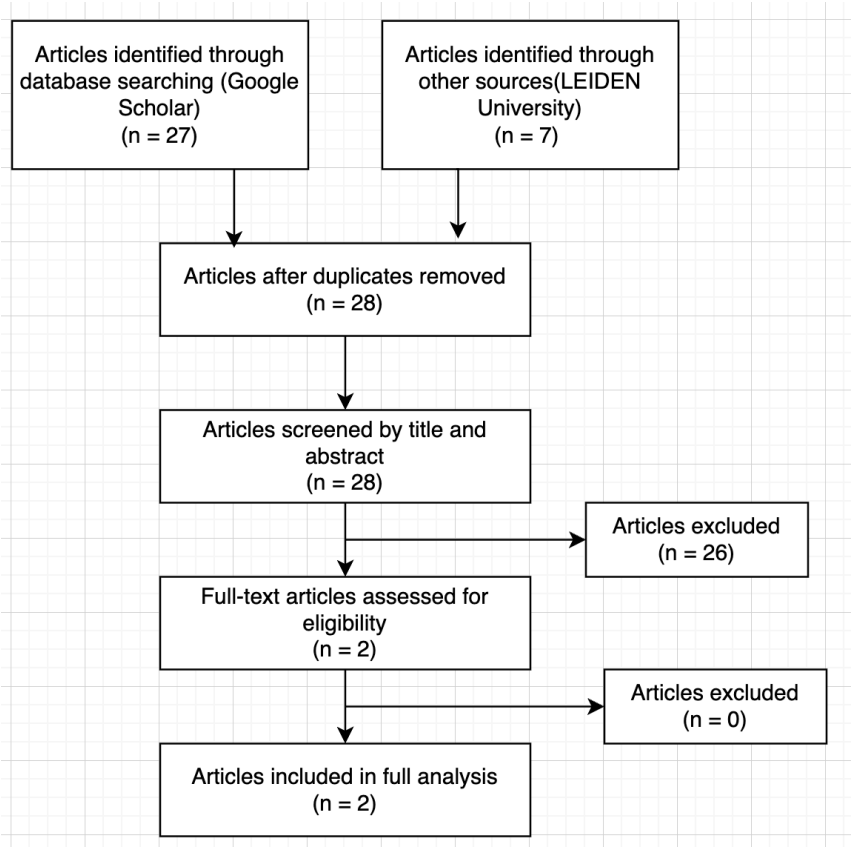



**Figure 3. Academic searching result of the De Novo FAIRification**

As illustrated in Figure 3, only two scientific papers – both investigating the same project – investigated the implementation of De Novo FAIRification. This result confirms that the first project successfully adopting De Novo FAIRification was in 2021 (Groenen et al., 2021). After that, no subsequent project utilises this technique.

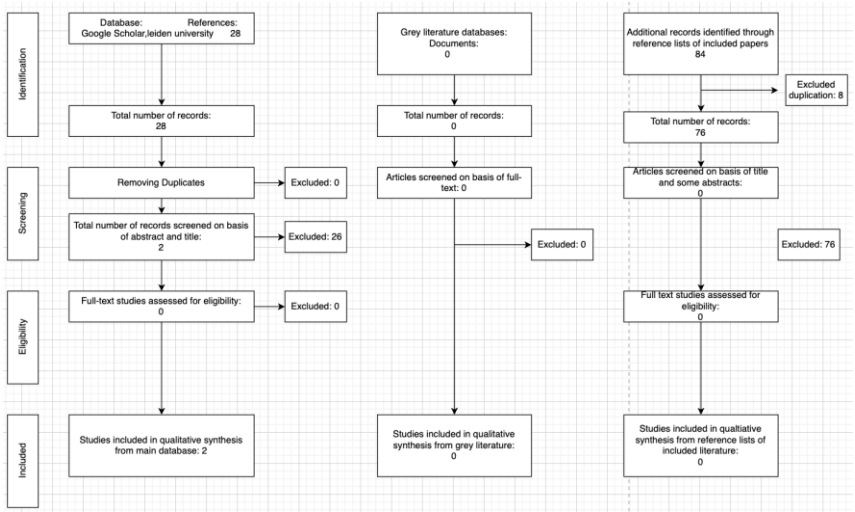



**Figure 4. Complete literature review result of the De Novo FAIRification**

As shown in Figure 4, none of the references in the two publications mentioned above examine the application of De Novo FAIRification further demonstrating the shortage of scholarly research on the subject before 2021. The absence of De Novo FAIRification in the grey literature database also indicates this scarcity.

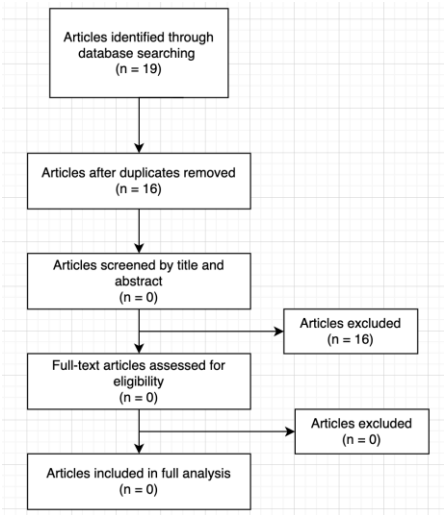



**Figure 5. Result of progressive literature review extension**

Furthermore, as shown in Figure 5, while 16 papers have cited these two foundational studies, none of them extend research on the implementation of De Novo FAIRification.

Moreover, there are limitations in the scope of De Novo FAIRification implementation. The 2021 project on De Novo FAIRification focuses solely on newly developed information systems.

## Discussion

Currently, there is a significant lack of research on the implementation of De Novo FAIRification, and only one project has attempted it. More research should be conducted on De Novo FAIRification.

Moreover, there are limitations in the scope of De Novo FAIRification implementation. The 2021 project on De Novo FAIRification focuses solely on newly developed information systems. There is no research on how to apply this technique to historical systems. This lack of research ignores the unique needs of historical systems in data management and sharing. Therefore, there is an urgent need to research how to implement De Novo FAIRification for legacy systems so that these systems can follow the FAIR principles and thus improve the findability, accessibility, interoperability, and reusability of their data.

## Conclusion

This study employed a literature review to assess the current state of research on De Novo FAIRification. The findings indicate that research on this topic is still in its early stages, with very few studies available in the literature. While the concept shows promise, there is a clear need for more extensive research to explore its full potential and practical implementation.

Automation of FAIRification in De Novo workflows not only saves time and reduces budgetary constraints but also enables a more efficient and timely application of the FAIR principles, ensuring that data remains compliant and accessible from the moment it is gathered. The integration of De Novo FAIRification into existing

workflows can streamline data management processes, reducing the need for extensive manual interventions. This approach ensures that data is FAIR-compliant from its inception, eliminating the need for retroactive adjustments. By automating the FAIRification process, organisations can achieve higher data quality and consistency, which is essential for reliable and reproducible research.

Moreover, the current focus of De Novo FAIRification has primarily been on newly developed systems. However, it is crucial that future research also considers its application to legacy systems, which present unique challenges in data management. Expanding the scope of De Novo FAIRification to include legacy systems will be essential to ensure that all systems, old and new, can benefit from the advantages of the FAIR principles. This will enhance the findability, accessibility, interoperability, and reusability of data across a broader range of contexts and applications.

## Acknowledgements

This chapter is based on: Lin, Z. (2025). *A Scalable Approach for De Novo FAIRification in Legacy Systems: Enabling Real-Time RDF Transformation, Semantic Integration, and Automated Data Upload*. Leiden University.

## Authors' Contributions

**Zhengyu Lin** conceptualised the research, conducted the research and wrote the article.

## Ethical Considerations

This research was conducted as part of a master thesis research project.

Tilburg University, Research Ethics and Data Management Committee of Tilburg School of Humanities and Digital Sciences REDC#2020/013, June 1, 2020-May 31, 2024, on Social Dynamics of Digital Innovation in remote non-western communities. Uganda National Council for Science and Technology, Reference IS18ES, July 23, 2019-July 23, 2023.

## References


de Oliveira, N. Q., Borges, V., Rodrigues, H. F., Campos, M. L. M., & Lopes, G. R. (2022). A Practical Approach of Actions for FAIRification Workflows. *Communications in Computer and Information Science*, *1537 CCIS*, 94–105. Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-3-030-98876-0_8

GO FAIR. (2022, January 21). FAIR principles. Retrieved June 26, 2024, from https://www.go-fair.org/fair-principles

General Data Protection Regulation (GDPR). (2016). Retrieved January 21, 2025, from https://gdpr-info.eu/

Groenen, K. H. J., Jacobsen, A., Kersloot, M. G., dos Santos Vieira, B., van Enckevort, E., Kaliyaperumal, R., Arts, D. L., 't Hoen, P. A. C., Cornet, R., Roos, M., & Kool, L. S. (2021). The De Novo FAIRification process of a registry for vascular anomalies. *Orphanet Journal of Rare Diseases, 16*(1), 376. https://doi.org/10.1186/s13023-021-02004-y

Kersloot, M. G., Jacobsen, A., Groenen, K. H. J., dos Santos Vieira, B., Kaliyaperumal, R., Abu-Hanna, A., … Arts, D. L. (2021). De-novo FAIRification via an Electronic Data Capture system by automated transformation of filled electronic Case Report Forms into machine-readable data. *Journal of Biomedical Informatics*, *122*. https://doi.org/10.1016/j.jbi.2021.103897

Lin, Z. (2025). *A scalable approach for De Novo FAIRification in legacy systems: Enabling real-time RDF transformation, semantic integration, and automated data upload* [Doctoral dissertation, Leiden University]. Leiden University Repository.

Plug, R., Liang, Y., Aktau, A., Basajja, M., Oladipo, F., & Van Reisen, M. (2022). Terminology for a FAIR Framework for the Virus Outbreak Data Network-Africa. *Data Intelligence*, *4*(4). https://doi.org/10.1162/dint_a_00167