FAIRification: A necessary practice for research data management

Dianelis Olivera Batista
Universidad Central "Marta Abreu" de Las Villas, Cuba.
Email: dolivera@uclv.cu

Amed Abel Leiva Mederos
Universidad Central "Marta Abreu" de Las Villas, Cuba.
Email: amed@uclv.edu.cu

María Josefa Peralta González
Universidad Central "Marta Abreu" de Las Villas, Cuba.
Email: mjosefa@uclv.edu.cu

ABSTRACT
The term FAIRification has become widespread among professionals whose work is related to research data management. However, little is known about FAIRification practices. This paper aims to examine FAIRification practices applied to research data from all areas of knowledge. The research is an exploratory study, using documentary analysis and content analysis methods. The results show that the literature on the subject is recent and generally in English. The papers, projects, and scientific articles analyzed show the development of infrastructures and tools but also the need for a culture of research data management. It is concluded that most of the experiences in FAIRification have been directed to the development of workflows, infrastructures, and tools to comply with FAIR principles. There is a predominance of FAIRification of data in the health research domain, with a greater boom after COVID-19.

Keywords: FAIR data, FAIRification, research data management

How to cite: Batista, D. O., Mederos, A. A. L., & González, M. J. P. (2024). FAIRification: A necessary practice for research data management. In M.J. Peralta González (Ed.), Generation and Transfer of Knowledge for Digital Transformation, II International Symposium, SITIC 2023, Santa Clara, 15-17 November 2023, proceedings. *Advanced Notes in Information Science, volume 6* (pp. 41-53). Pro-Metrics: Tallinn, Estonia. DOI: 10.47909/978-9916-9974-5-1.92.

1. INTRODUCTION

Research data management has become established in the scientific and academic domains in recent decades. For Cox and Pinfield (2014), research data management consists of a series of different activities and processes associated with the data lifecycle, involving data design and creation, storage, security, preservation, retrieval, sharing, and reuse.

Wilkinson et al. (2016) propose the findable, accessible, interoperable, and reusable (FAIR) principles, considering them a prerequisite for achieving research data management. These authors point out that contemporary e-Science requires data to be FAIR in the long term, and these objectives are fast becoming expectations of agencies and publishers.

Among the advances in the implementation of these principles, the GO FAIR initiative (GO FAIR, 2017) stands out. Within this initiative, the term "FAIRification" is beginning to be used, whose use, both in English and Spanish (FAIRificación), has spread in recent years. This term is used to refer to the processing of data and metadata to comply with FAIR principles. However, little is known about FAIRification practices. Only the study by Inau et al. (2023) refers to such experiences but in the area of health data management specifically. This work, which is part of the theoretical framework of an ongoing doctoral research, aims to examine FAIRification practices applied to research data from all areas of knowledge.

2. METHODOLOGY

The research is an exploratory study. Documentary analysis and content analysis methods were used. First, a bibliography on the subject was compiled, using the term "FAIRification" as a search strategy. The results were filtered, taking as exclusion criteria those documents that did not allude to practices and experiences. Finally, a total of 14 scientific articles, six papers presented at events, and five research project presentations were analyzed.

3. RESULTS AND DISCUSSION

The literature on FAIRification practices is recent. As can be seen in Table 1, the publications are from 2018 onwards, and an increase is observed from 2020 onwards. It should be noted that one of the project presentations is undated and that perhaps the number of publications is not higher in 2023 because the search was performed in September of the same year.

Table 1. Number of publications per year.

YEAR	PUBLICATIONS
2018	One
2019	One
2020	Three
2021	Six
2022	Seven
2023	Six

Source: Prepared by the authors.

Almost all the documents are in English. Only one paper was published in Spanish where Anglada (2021) describes the requirements of repositories to comply with FAIR principles and the facilities offered by Dataverse. In another paper, Mavraki et al. (2021) define a semantic model for data and metadata to be FAIR within the LifeWatchGreece biodiversity research infrastructure.

Annane et al. (2021) present an ontological model for FAIR data in the meteorological domain. They take into account the characteristics of meteorological data and make it applicable to Météo-France (French National Meteorological Service) but also to any institution working with this type of data. In addition, the authors intend to continue to enrich and refine the proposed model.

On the other hand, Mangione et al. (2022) analyze the gray literature tools and approaches that emerge when adopting FAIR principles. A total of 477 emerging tools are analyzed and organized into a comprehensive map. Österle and Touré (2022) describe the creation and use of a health network in Switzerland that enables the development of a FAIR ecosystem. Azeroual et al. (2023) refer to a systematic literature review on current research information systems (CRISs) and discuss how FAIRification should work in CRISs based on existing practices.

Regarding projects, ELIXIR-EXCELERATE is presented by Jacobsen et al. (2018). It describes a seven-step FAIRification process, auxiliary tools, and recommendations for data manipulation in the field of rare diseases. The seven steps include (1) defining user driving questions, (2) pre-FAIRification analysis, (3) semantic model definition, (4) data record transformation, (5) metadata definition, (6) FAIR data resource implementation, and (7) user interface or application query.

Another Entellect project, deployed by Elsevier (2019), is a platform that empowers data-driven R&D in the pharmaceutical industry and refers to the need for cultural change. Meanwhile, the FAIRplus project, according to the European Commission (n.d.), aims to improve FAIR levels of data and change the culture of data management. This project has generated a scalable framework for FAIRification of data. It has refined the implementation of FAIR principles in working with public data from Innovative Medicines Initiative projects and internal data from pharmaceutical industry partners.

Later, EOSC-Nordic (2020) is developed, which has guided repositories in the Nordic and Baltic countries to make their data FAIR. The EOSC-Nordic FAIRification initiative team defined a sample size of approximately 100 data repositories, for which they evaluated the implementation of FAIR principles, with the aim of guiding and training repositories towards a higher level of FAIRification.

Aventurier et al. (2022) present the recommendations of the ANR-BRIDGE project for data FAIRification. The goal of this project is to provide guidelines and harmonize research data policies and repository management in a reusable approach for other institutes or contexts, focusing on three priorities: analyze and improve institutional data governance policies, provide and support common guidelines for data producers and managers, and choose FAIR vocabularies and develop tools for repositories with some shared metadata schemas.

One of the articles analyzed describes the FAIR4Health project. According to Alvarez-Romero et al. (2021), FAIR4Health is a project whose main objective is to encourage and promote the application of FAIR principles in data derived from publicly funded health research initiatives.

The main purpose is to be able to share these data and reuse them in the EU health research community. Jacobsen et al. (2020) describe a generic FAIRification workflow. This flow should be applicable to any type of data. The steps are (1) identify the FAIRification target, (2) analyze data, (3) analyze metadata, (4) define a semantic model for data (4a) and metadata (4b), (5) make data (5a) and metadata (5b) linkable, (6) host FAIR data, and (7) evaluate FAIR data. For each step, it describes how data are processed, what expertise is required, and what procedures and tools can be used.

Sinaci et al. (2020) propose a technological architecture for FAIRification. The proposed architecture is based on the use of fast health care interoperability resources. The authors conclude that health care datasets or data resulting from health research can be FAIRified, shared, and reused within the health research community following the proposed workflow and implementation of the technology architecture.

Bernabé et al. (2021) refer to the use of techniques to identify the need for FAIR data, what tasks to perform, resources used, and so on. These authors design a method that uses "goal-oriented models" to support the "objective identification" and "conceptual modeling" steps of FAIRification. First, the motivations for the need for FAIR data are identified. Then, objective models are used to define the scope, identify important concepts, and validate the resulting conceptual model. The method will also describe best practices and activities for conceptual modeling.

Gundersen et al. (2021) decide to advance the application of FAIR principles to produce searchable metadata for genomic clues. To this end, they develop a JSON schema, called FAIRtracks, and integrate it into a novel track search

service, called TrackFind. They demonstrate practical use by importing datasets through TrackFind into existing examples of analytical tools relevant to genomic tracks: EPICO and GSuite HyperBrowser. Thus, they provide a first version of a draft standard for genomic tracking metadata and the accompanying software ecosystem.

Groenen et al. (2021) implement a FAIRification process for recording vascular anomaly data. They describe the five phases of this process in detail: (1) pre-FAIRification, (2) facilitating FAIRification, (3) data collection, (4) generating real-time FAIR data, and (5) using FAIR data. The authors believe that the process can be reused by other rare disease registries and that this work can be a substantial contribution to a FAIR ecosystem of rare disease data.

While dos Santos et al. (2022) show FAIRification experience in rare disease data. In Europe, 24 European Reference Networks (ERNs) are working on rare disease registries in different clinical settings. The FAIRification process differs between the different ERN registries. For example, registries use different software systems and are subject to different legal regulations. To help ERNs make informed decisions and harmonize FAIRification, a management team was created.

Queiroz et al. (2022) present actions to generate FAIR data and matadata for COVID-19 research. The article presents a workflow of actions taken to generate FAIR metadata for COVID-19 research. In addition, tools for (semi) automating metadata processing are evaluated whenever possible. Although defined for a particular use case, it is expected that this workflow can be applied to other epidemic investigations and in other domains.

Both Alharbi et al. (2022) and Alharbi et al. (2023) propose methodologies and structures to facilitate

FAIRification decision making in the pharmaceutical R&D industry. In turn, these authors identify challenges facing the FAIRification process and formulate a cost-benefit assessment. In this way, they aim to provide adequate decision making regarding the research data in the pharmaceutical sector that should be prioritized when implementing FAIR principles.

Ribeiro et al. (2023) conducted a case study of the application of FAIR principles for data sharing, use and reuse in musicology. Datasets were located and consulted at the FAIR preprocessing level from sites in the area of Musicology and in Google Scholar. As a result, a semantic model based on vocabularies to describe electronic resources could be proposed. In addition, points of convergence between FAIRification processes and data preparation for use as related open data were perceived.

Touré et al. (2023) follow up on the work presented in the paper by Österle and Touré (2022) concerning the Swiss Personalized Health Network. The Resource Description Framework schema is implemented along with a data ecosystem encompassing data integration, validation tools, analysis aids, training, and documentation to represent metadata and health data in a consistent manner. This will enable the achievement of national data interoperability goals. In this way, researchers in Switzerland have access to FAIR health data for further use.

In the article by Parciak et al. (2023), the implementation of a FAIR-mode data processing automation framework in a hospital research center is presented. The authors demonstrate the implementation of the proposed framework by describing its use in the Medical Data Integration Center. The prototype implementation also includes a

metadata schema for data provenance and a process validation concept.

Welter et al. (2023) develop a flexible, multilevel, domain-independent FAIRification framework that provides practical guidance for improving FAIRification for existing and future clinical and molecular datasets. The framework is validated in collaboration with several major public–private partnership projects, demonstrating and delivering improvements across all FAIR aspects and across a variety of datasets and their contexts. The authors believe that the reproducibility and applicability of the approach to FAIRification tasks is established.

In recent years, there has been a predominance of publications focused on FAIRification in the field of health research. The authors of this study concur with Inau et al. (2023), who discuss concepts, approaches, and implementation experiences in health FAIR initiatives, that successful FAIRification of data has informed the management and prognosis of various diseases. In addition, the available literature indicates that more efforts have been made to FAIRify disease data since COVID-19.

4. CONCLUSIONS

Most of the experiences in FAIRification have been directed to the development of workflows, infrastructures, and tools to make data and metadata compliant with FAIR principles. FAIRification predominates in the health research domain. A boom has been seen after the COVID-19 pandemic, as the need to share data to address the disease sparked greater interest in these issues. Future studies should be conducted to see how FAIRification practices evolve, as they are very new.

REFERENCES

Alharbi, E., Gadiya, Y., Henderson, D., Zaliani, A., Delfin-Rossaro, A., Cambon-Thomsen, A., Kohler, M., Witt, G., Welter, D., Juty, N., Jay, C., Engkvist, O., Goble, C., Reilly, D. S., Satagopam, V., Ioannidis, V., Gu, W., & Gribbon, P. (2022). Selection of data sets for FAIRification in drug discovery and development: Which, why, and how? *Drug Discovery Today*, *27*(8), 2080–2085. https://doi.org/10.1016/j.drudis.2022.05.010

Alharbi, E., Skeva, R., Juty, N., Jay, C., & Goble, C. (2023). A FAIR-Decide framework for pharmaceutical R&D: FAIR data cost–benefit assessment. *Drug Discovery Today*, *28*(4), Article 103510. https://doi.org/10.1016/j.drudis.2023.103510

Alvarez-Romero, C., Martínez-García, A., Román-Villarán, E., & Parra-Calderón, C. L. (2021). FAIR4Health: Reutilizar y Compartir Datos de Salud Aplicando los Principios FAIR en un Modelo Híbrido de Nube y Datos Distribuidos. *Informática + Salud*, *143*, 16–20.

Anglada, L. (2021). *Implementação de serviços nacionais e estratégias institucionais para a Gestão de Dados de Investigação. Foro Portugués de Gestión de Datos de Investigación, Coimbra-Barcelona*. https://forumgdi.rcaap.pt/wp-content/uploads/2021/11/Keynote_2111RDMiRDR_ForoPortuguesDatosCoimbraV1.pdf

Annane, A., Kamel, M., Trojahn, C., Aussenac-Gilles, N., Comparot, C., & Baehr, C. (2021). *Towards the FAIRification of meteorological data: A meteorological semantic model. metadata and semantic research 15th international conference, MTSR 2021* (pp. 81–93). Springer International Publishing.

Aventurier, P., Fortuno, S., Szabo, D., Alaux, M., Bartelemy, C., Bonnet, P., Catherine, H., Dzalé, E., Deboin, M.-C., Decker, L., Desconnets, J.-C., Doux, G., Mougin, C., Perez, J., & Sabot, F. (2022, June 16). *Data FAIRification in a cross-institutional governance framework: recommendations from the ANR-BRIDGE project*. Research Data Alliance. 19th Plenary Meeting, Séoul, South Korea. Zenodo. https://doi.org/10.5281/zenodo.6652405

Azeroual, O., Schöpfel, J., Pölönen, J., & Nikiforova, A. (2023). FAIRification of CRIS: A review. In: F. Coenen, et al. (Eds.), *Knowledge discovery, knowledge engineering and knowledge management. IC3K 2022. Communications in computer and information science* (vol. 1842). Springer.

Bernabé, C. H., Jacobsen, A., Queralt Rosinach, N., Bonino da Silva Santos, L. O., Silva Souza, V. E., Mons, B., & Roos, M. (2021). *Goal-models to support communication, planning and guiding of FAIRification*. Zenodo. https://doi.org/10.5281/zenodo.5784628

European Commission. (n.d.). *'FAIRification' of valuable data sets to share the wealth*. https://cordis.europa.eu/article/id/444085-fairification-of-valuable-data-sets-to-share-the-wealth

Cox, A. M., & Pinfield, S. (2014). Research data management and libraries: Current activities and future priorities. *Journal of Librarianship and Information Science*, *46*(4), 299–316. https://doi.org/10.1177/0961000613492542

dos Santos, B., Bernabé, C. H., Zhang, S., Abaza, H., Benis, N., Cámara, A., Cornet, R., Cornec, L., Peter, Schaefer, F., van, Swertz, M. A., Wilkinson, M. D., Jacobsen, A., & Roos, M. (2022). Towards FAIRifcation of sensitive and fragmented rare disease patient data: challenges and solutions in European reference network registries. *Orphanet Journal of Rare Diseases*, *17*, 436. https://doi.org/10.1186/s13023-022-02558-5

Elsevier. (2019). *Realizing the potential of FAIR data for pharmaceutical R&D*. https://www.elsevier.com/__data/assets/pdf_file/0007/961972/Potential-of-FAIR-Data-in-Pharma_whitepaper_PLS_WEB.pdf

EOSC-Nordic. (2020). *EOSC-NORDIC FAIRification study testing F-UJI*. https://eoscsecretariat.eu/system/files/fairsfair_eosc_nordic_adoption_story.pdf

GO FAIR. (2017). *Progress towards the European Open Science Cloud: GO FAIR Office established*. https://www.go-fair.org/2017/12/07/progress-towards-european-open-science-cloud-go-fair-office-established/

Groenen, K. H. J., Jacobsen, A., Kersloot, M. G., dos Santos, B., Enckevort, E., Kaliyaperumal, R., Arts, D. L., 't Hoen, P. A. C., Cornet, R., Roos, M., & Schultze, L. (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

Gundersen, S., Boddu, S., Capella-Gutierrez, S., Drabløs, F., Fernández, J. M., Kompova, R., Titov, D., Zerbino, D., & Hovig, E. (2021). Recommendations for the FAIRification of genomic track metadata [version 1; peer review: 2 approved]. *F1000Research*, *10*, 268. https://doi.org/10.12688/f1000research.28449.1

Inau, E. T., Sack, J., Waltemath, D., & Zeleke, A. A. (2023). Initiatives, concepts, and implementation practices of the findable, accessible, interoperable, and reusable data principles in health data stewardship: Scoping review. *Journal of Medical Internet Research, 25*, e45013. https://doi.org/10.2196/45013

Jacobsen, A., Kaliyaperumal, R., Bonino, L. O., Mons, B., Schultes, E., Roos, M., & Thompson, M. (2020). A generic workflow for the data FAIRification process. *Data Intelligence*, *2*(1–2), 56–65. https://doi.org/10.1162/dint_a_00028

Jacobsen, A., Thompson, M., Hanauer, M., Beltran, S., Gray, A., Juty, N., Ehrhart, F., Evelo, CT. T., & Ross, M. (2018). ELIXIR-EXCELERATE: Fast-track ELIXIR implementation and drive early user exploitation across the life sciences. *ELIXIR Infrastructure for Rare Disease Research*. https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5bd475987&appId=PPGMS

Mangione, D., Candela, L., & Castelli, D. (2022). *A taxonomy of tools and approaches for FAIRification*. 18th Italian Research Conference on Digital Libraries, Padova, Italy. https://ceur-ws.org/Vol-3160/paper6.pdf

Mavraki, D., Chatzinikolaoy, E., & Panteri, E. (2021). *FAIRification process for RIs: The case of the LifeWatchGreece Research Infrastructure*. LifeWatch ERIC Workshop "e-Science for NIS Research". https://www.lifewatch.eu/wp-content/uploads/2022/01/IJINISW_S1_Mavraki.pdf

Österle, S., & Touré, V. (2022). *FAIRification of health-related data in the Swiss Personalized Health Network*. MILA Seminar, University of Greifswald. https://www.medizin.uni-greifswald.de/medizininformatik/fileadmin/user_upload/poster_presentation_articles/by_externals/2022-01-24-MILA_seminar_SO_VT.pdf

Parciak, M., Suhr, M., Schmidt, C., Bönisch, C., Löhnhardt, B., Kesztyüs, D., & Kesztyüs, T. (2023). FAIRness through automation: development of an automated medical data integration infrastructure for FAIR health data in a maximum care university hospital. *BMC Medical Informatics and Decision Making*, *23*, 94. https://doi.org/10.1186/s12911-023-02195-3

Queiroz, N., Borges, V., Rodrigues, H. F., Machado, M. L., & Rabello, G. (2022). A practical approach of actions for FAIRification workflows. *ArXiv, abs/2201.07866*.

Ribeiro, C. J. S., de Sousa, C. S., Tertulino, C. Í., do Amaral, I. S., dos Santos, M. J. S., Cóquero, S. D. M. S., & de Ulhôa, M. T. (2023). Adopción de los principios FAIR (proceso de FAIRificación) en conjuntos de datos: interconexión de noticias de publicaciones periódicas musicales del siglo XIX. *Ibersid: Revista De Sistemas De información Y documentación*, *17*(2), 123–124. https://doi.org/10.54886/ibersid.v17i2.4949

Sinaci, A. A., Núñez-Benjumea, F. J., Gencturk, M., Jauer, M. L., Deserno, T., Chronaki, C., Cangioli, G., Cavero-Barca, C., Rodríguez-Pérez, J. M., Pérez-Pérez, M. M., Laleci Erturkmen, G. B., Hernández-Pérez, T., Méndez-Rodríguez, E., & Parra-Calderón, C. L. (2020). *From raw data to FAIR data: The FAIRification workflow for health research*. *Methods of Information in Medicine*, *59*, e21–e32.

Touré, V., Krauss, P., Gnodtke, K., Buchhorn, J., Unni, D., Horki, P., Raisaro, J. L., Kalt, K., Teixeira, D., Crameri, K., & Österle, S. (2023). FAIRification of health-related data using semantic web technologies in the Swiss Personalized Health Network. *Science Data*, *10*(127). https://doi.org/10.1038/s41597-023-02028-y

Welter, D., Juty, N., Rocca-Serra, P., Xu, F., Henderson, D., Gu, W., Strubel, J., Giessmann, R. T., Emam, I., Gadiya, Y., Abbassi-Daloii, T., Alharbi, E., Gray, A. J. G., Courtot, M., Gribbon, P., Ioannidis, V., Reilly, D. S., Lynch, N., Boiten, J.-W., … Burdett, T. (2023). FAIR in action—A flexible framework to guide FAIRification. *Scientific Data*, *10*(1). https://doi.org/10.1038/s41597-023-02167-2

Wilkinson, M. D., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), Article 160018. https://doi.org/10.1038/sdata.2016.18