FAIR Data by Design: A Case of the DiVA Portal

Phub Namgay[1][a] and Joshua C. Nwokeji[2][b]
[1]*Department of Informatics and Media, Uppsala University, Sweden*
[2]*Computer and Information Science Department, Gannon University, U.S.A.*

Keywords: FAIR Data Principles, FAIR Data, FAIR Data Repositories, DiVA Portal.

Abstract: FAIR Data Principles is a guideline for making data and other digital objects *f*indable, *a*ccessible, *i*nteroperable, and *r*eusable. Thus far, the traction and uptake of the principle are primarily in the domain of bio and natural sciences. The knowledge gap is the application of the FAIR Data Principles in designing data repositories for FAIR data in the academic data ecosystem. This paper provides a critical insight into how the principle can be utilised as a paradigm to design data that embodies the tenets of FAIR Data Principles. We conducted a case study of the DiVA portal, an information repository and finding tool in Sweden, to explicate FAIR data by design. The portal scored high in a qualitative assessment against the 15 facets of FAIR Data Principles, as illustrated by the high density of green cells in the traffic light rating matrix (see Table 1). It indicates the robustness of data in the portal that is easy to share, find, and reuse. This study suggests practitioners operationalise FAIR Data Principles in their data repositories by design through systems and policies underpinned by the principle. It would enrich data governance and management for the back office and data experiences for end users. The study also advances the knowledge base on data management through a granular exposition of FAIR data by design.

1 INTRODUCTION

The heart of a data repository is that end users can effortlessly share, find, and reuse data. The pervasive datafication (Lycett, 2013) of society further intensifies a need for robust data repositories to share and consume data. For example, in academia, students and professors create and share data in information portals and data repositories for reuse by others. In the current study, 'data' refers to digital objects such as publications, theses, data sets, and metadata. Data is also increasingly 'born digital' in the academic data ecosystem. However, the lack of coherent policies on data infrastructure to govern and manage data impedes the processes and practices around data. A paradigm that underpins data repositories and data therein 'by design'—that is, 'by plan'—is crucial for producing and consuming findable and reusable data in a data ecosystem.

FAIR Data Principles is a guideline for managing and stewarding data and other digital objects (Wilkinson et al., 2016). It has witnessed rapid traction in research and practice (van Reisen et al., 2020). Nevertheless, the knowledge gap is an account of the application of FAIR Data Principles in the context of everyday academic data repositories to illustrate the practicality of the principle. A study that examines data repositories by anchoring on FAIR Data Principles is essential, which entails scrutinising digital objects for FAIRness (Jacobsen et al., 2020; Wilkinson et al., 2019). DiVA portal, an information repository and finding tool in Sweden (The DiVA Consortium, 2022), was sampled for a case study to pursue the research question— *Why should data repositories be FAIR by design to facilitate findable and reusable data?*

The exponential growth of data and increased demand for data have informed us that simply sharing data on the data repositories is insufficient today. Reinforcing data infrastructures with data paradigms is essential for sharing and reusing data. DiVA portal was assessed against FAIR Data Principles (Wilkinson et al., 2016). The portal scored high in overall FAIRness, as substantiated by many green cells in a colour-coded matrix using a traffic light

[a] https://orcid.org/0000-0001-6034-7274
[b] https://orcid.org/0000-0003-4643-2418

Namgay, P. and Nwokeji, J.
FAIR Data by Design: A Case of the DiVA Portal.
DOI: 10.5220/0011964300003467
In *Proceedings of the 25th International Conference on Enterprise Information Systems (ICEIS 2023) - Volume 2*, pages 160-167
ISBN: 978-989-758-648-4; ISSN: 2184-4992

rating (Dunning et al., 2017). This study posits that data repositories, such as those used daily in the academic data ecosystem, must consider and operationalise FAIR Data by design to facilitate data governance and management for the back office and a seamless data experience for end users.

In the following sections, first, we provide an overview of FAIR Data Principles. Then, an account of the study design and method is presented to set the tone, followed by the analysis and findings of the study. Next, we discuss the interpretation of the findings, contribution to knowledge and practice, and study limitations. The paper finally ends with a conclusion of the study.

2 OVERVIEW OF FAIR DATA PRINCIPLES

FAIR Data Principles is a guideline for making data and other digital objects findable, accessible, interoperable, and reusable (Wilkinson et al., 2016), as in Figure 1. The principle emphasises data reuse (David et al., 2020) and offers a middle ground for sharing and reusing data on the spectrum of 'close data' and 'open data' (Boeckhout et al., 2018).

FAIR Data Principles have gained traction and uptake in the research, government, and practice (Jacobsen et al., 2020; van Reisen et al., 2020; Wilkinson et al., 2019). The H2020 Program Guideline underlines FAIR Data Principles by highlighting that data should be "as open as possible, as closed as necessary" (Celina, 2017; Landi et al., 2020). The inequality in implementing FAIR Data Principles across different world geographies is a concern as high uptake is reported in the Western and Northern hemispheres (van Reisen et al., 2020), predominantly in the domain of bio and natural sciences (Dunning et al., 2017; van Reisen et al., 2020).

The facets of FAIR Data Principles are concise and thus susceptible to diverse interpretations (David et al., 2020; Jacobsen et al., 2020). However, the interpretations should still be within the spirit of the original facets or established behavioural guidelines of the principle (Mons et al., 2017), even if the requirements for FAIR data vary across disciplines (Dunning et al., 2017). The unqualified application of FAIR Data Principles also risks creating additional issues for data stewardship (Boeckhout et al., 2018).

FAIR Data Principles has not been seriously used in everyday data repositories or digital artefacts, such as those used daily in academia. The discourse and literature in this area are scant. Thus, it is in order to explore how data repositories in the academic data ecosystem operationalise FAIR Data Principles and explicate FAIR data by design to ease the production and consumption of data.

FAIR Data Principles

To be findable:
- F1. (Meta)data are assigned a globally unique and persistent identifier
- F2. Data are described with rich metadata
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a searchable resource

To be accessible:
- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
  - A1.1 The protocol is open, free, and universally implementable
  - A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available

To be interoperable:
- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
- I2. (Meta)data use vocabularies that follow FAIR principles
- I3. (Meta)data include qualified references to other (meta)data

To be reusable:
- R1. (Meta)data are richly described with a plurality of accurate and relevant attributes
  - R1.1. (Meta)data are released with a clear and accessible data usage license
  - R1.2. (Meta)data are associated with detailed provenance
  - R1.3. (Meta)data meet domain-relevant community standards

Figure 1: FAIR Data Principles and its facets (Wilkinson et al., 2016).

3 STUDY DESIGN AND METHOD

3.1 Study Context

DiVA portal is an institutional information repository or publishing database established for research, digital materials, and thesis publications by the 50 member universities, colleges, institutes, and museums of the DiVA consortium in Sweden (The DiVA Consortium, 2022). It is a common tool for finding and reusing academic resources, non-academic materials, and data sets. DiVA portal currently stores about 20 different types of publications, including data sets. It also provides a service for the long-term archiving of scholarly output. The portal promotes Open Science, and digital objects therein are globally accessible.

3.2 Study Method

A case study was conducted to examine a data repository against the four foundational principles and 15 facets of FAIR Data Principles, as in Figure 1. DiVA portal was sampled as a case to answer the

research question. The portal is a widely used information system in Sweden. Furthermore, with the hope of disseminating research, findings, and knowledge, the portal is a dynamic system with continual new data every day. It is also apparent from using the portal that it would be compelling to explore how the portal embodies the facets of FAIR Data Principles. We also gathered data about the portal, through email correspondence, from a university library and the data office of one of the universities in the DiVA consortium. Their experiential insights and perspectives on FAIR data were crucial for the in-depth analysis of the portal.

4 ANALYSIS AND FINDINGS

The following subsections illustrate the manual assessment of the DiVA portal against the 15 facets of FAIR Data Principles (see Dunning et al., 2017). In the interest of space, we layered the figures as the text behind the inset is not central in the current study.

4.1 Findable Data

The first task in locating data of interest is to *find* it in a data repository. Metadata, defined as 'data about data', should facilitate finding data, and machine-actionable metadata is desired for the automatic data search.

F1. *(Meta)data are assigned a globally unique and persistent identifier*—DiVA portal uses unique persistent identifier (PID), such as Uniform Resource Name (URN), Open Archive Initiative OAI, DiVA id, and Digital Object Identifier (DOI) to identify data on the portal uniquely. The portal automatically mints a URN: NBN (Uniform Resource Name for National Bibliography Number) while registering any digital objects on the portal, as shown in Figure 2.

F2. *Data are described with rich metadata*— All the digital objects on the DiVA portal are described with rich metadata for meaningful reuse, as shown in Figure 2. Most of the information in metadata is captured while registering data on the portal.

F3. *Metadata clearly and explicitly include the identifier of the data it describe*—The PID of data, that is, URN, is captured in the metadata, as shown in Figure 3. Metadata is viewable through the portal's 'Export' feature, which exports metadata in 13 different formats. For example, the Electronic Theses and Dissertations-Metadata Standard (ETD-MS) is a metadata standard for electronic theses and dissertations based on Dublin Core.

F4. *(Meta)data are registered or indexed in a searchable resource*—Data and its associated metadata are registered in a searchable Web-based resource, as shown in Figure 3 (see Figure 2 for PID). It is also consistent with linking data and metadata (Jacobsen et al., 2020).

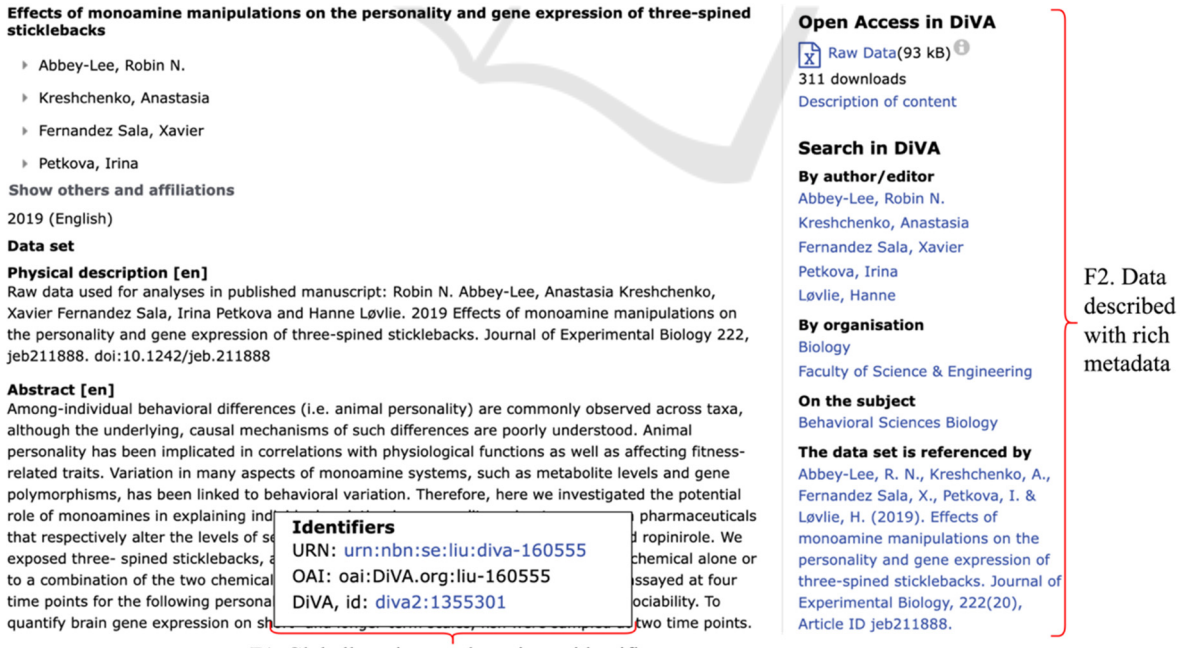


Figure 2: Facet F1 and F2 of FAIR Data Principles.

Figure 3: Facet F3 and F4 of FAIR Data Principles.

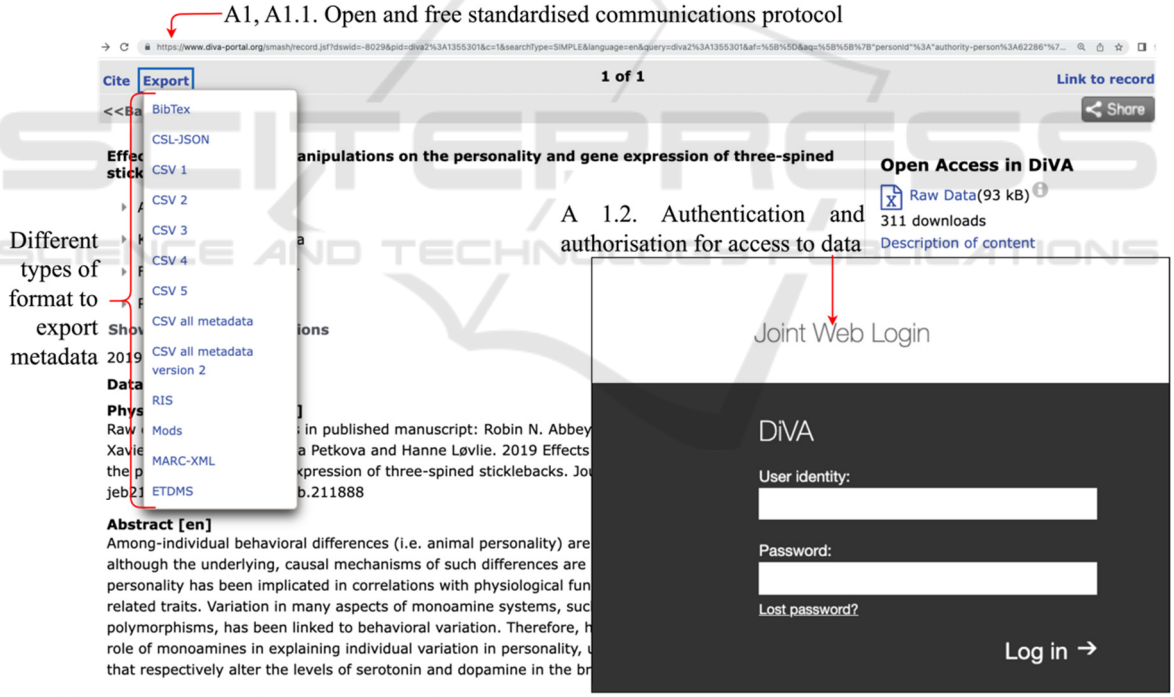


Figure 4: Facet A1, A1.1, A 1.2 (inset), and A2 of FAIR Data Principles.

4.2 Accessible Data

Once the data is found, a user needs to know how it can be *accessed* using authentication and authorisation mechanisms (Landi et al., 2020), such as gatekeeper, open-access, or a mix of both. This information is necessary to ensure data can be located and reused, not simply found and viewed on a data repository, through well-defined data 'access' control.

A1. *(Meta)data are retrievable by their identifier using a standardised communications protocol*— DiVA portal uses HTTPS communication protocol to deliver data on the Web, as in Figure 4. The associated metadata is also retrievable from the Web. It is worth noting that data can be found yet not accessible.

A1.1 *The protocol is open, free, and universally implementable*—DiVA portal is open and accessible through an HTTPS connection, as in Figure 4.

A1.2 *The protocol allows for an authentication and authorisation procedure, where necessary*—A user must log in using the university or institute login credential or Single sign-on to share any data on the DiVA portal, as in Figure 4. It resonates with the controlled access model (Landi et al., 2020) for a regulatory and ethical check. Any data to be shared on the portal must be bibliographically approved.

A2. *Metadata are accessible, even when the data are no longer available*— Metadata of data on the DiVA portal is accessible even if data ceases to exist. Personnel from the Scholarly Communication Division of one of the universities in the DiVA consortium noted: "*Yes, metadata, i.e. the description of the data, will still be available in DiVA even if the data no longer exists. Once registered in DiVA, the record with the metadata will receive a permanent identifier and link. The only way the metadata can disappear from DiVA is if someone hides or deletes the record itself*."

4.3 Interoperable Data

Data is often *interoperated* or integrated with other data, workflows, and applications. It can also be an input to workflow systems.

I1. *(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation*— Different metadata schemas are used for organising knowledge. DiVA portal uses the Metadata Object Description Schema (MODS), a bibliographic element set schema, as shown in Figure 5. The information captured in the portal for registering data is based on the DataCite metadata schema. The portal also provides the feature to retrieve digital e-archive files in Metadata Encoding and Transmission Standard (METS) format.

I2. *(Meta)data use vocabularies that follow FAIR principles*— This facet was challenging to assess. Dunning et al. (2017) reported a similar experience. However, personnel from the data office of one of the universities in the DiVA consortium shared: "*DiVA developers are doing some other FAIR-related work in DiVA, like making it possible to register licences in the form*."

I3. *(Meta)data include qualified references to other (meta)data*—'Qualified references' refers to informational links that would help disambiguate the terms used (Dunning et al., 2017). Data shared on the DiVA portal has a qualified reference to the publication, as in Figure 5. The PID of the research publication is linked to its data set through 'The data set is referenced by' and 'The publication has references to' features in the DiVA Portal.

4.4 Reusable Data

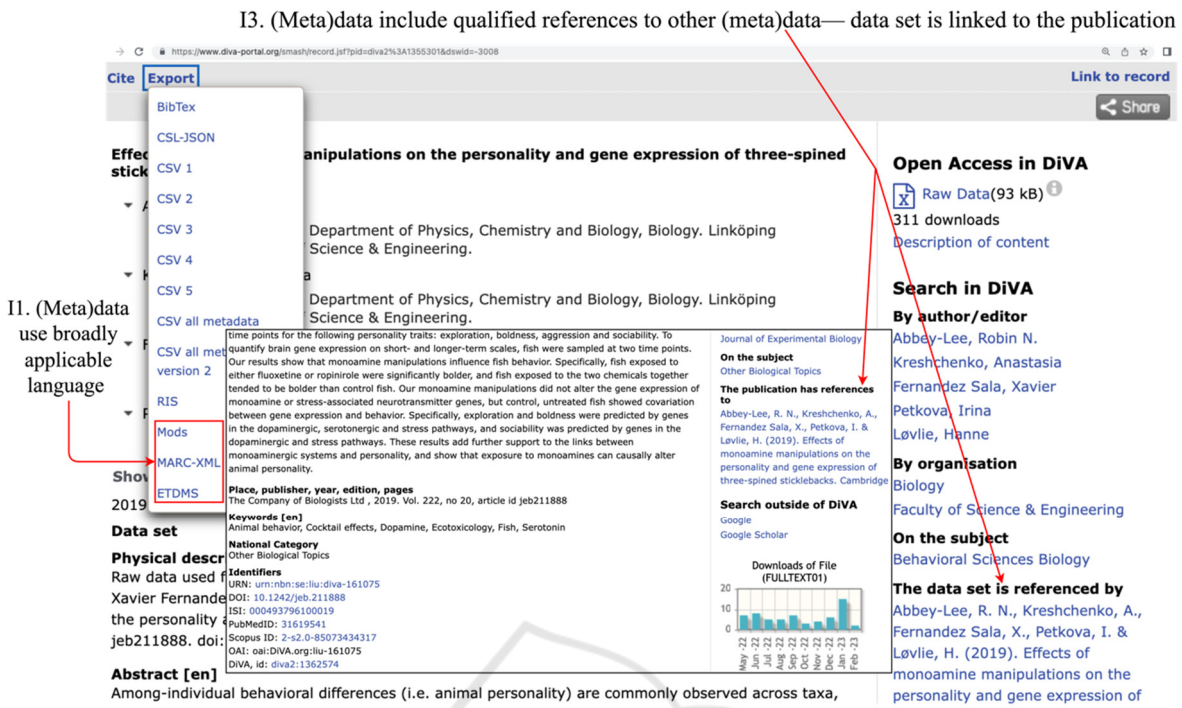
The goal of sharing data on various data repositories is to *reuse* it in order to create an innovative solution or solve problems.

R1. *Meta(data) are richly described with a plurality of accurate and relevant attributes*—Data on the DiVA portal is described with rich metadata. For example, in the context of a data set, the ETD-MS includes essentials such as PID, authors, and provenance, as shown in Figure 6. The metadata for publications includes richer information on the business, technical, and operational aspects, such as identifiers, conferences, journals, and publishers, as in Figure 5 (see inset).

R1.1. *(Meta)data are released with a clear and accessible data usage license*—DiVA portal mentions the publishing conditions. The portal clearly states that the copyright of any data belongs to the author, as shown in Figure 6. Data on the portal is open access. Likewise, the portal recommends authors share their work with a licence such as Creative Commons.

R1.2. *(Meta)data are associated with detailed provenance*—Metadata captures information on the provenance of the data. For example, information such as author, publisher, and identifier is captured across different metadata schemas, as shown in Figure 6.

R1.3. *(Meta)data meet domain-relevant community standards*—Unlike other facets, which are observable by brief investigations such as facet 'F1', some facets require detailed subject knowledge, such as assessing whether data and its associated metadata meet community or discipline-specific standards. Examining whether the data shared on the DiVA portal explicitly meets domain-specific metadata and data standards is difficult. Dunning et al. (2017) found some facets challenging to assess and gave the lowest rating—'unclear' in traffic light rating.



I2. (Meta)data use vocabulary that follow FAIR Data Principles – *Challenging to assess; Dunning et al. (2017) reported a similar experience.*

Figure 5: Facet I1, I2, and I3 of FAIR Data Principles.

R 1.3. (Meta) data meet domain-relevant community standards – *It is challenging to assess whether (meta)data meets domain-specific standards for all the data shared on the portal. Dunning et al. (2017) reported a similar experience.*

Figure 6: Facet R1, R1.1, R1.2, and R1.3 of FAIR Data Principles.

5 DISCUSSION

The following subsections discuss the interpretation of the findings, contributions to knowledge, and limitations of the study.

5.1 FAIR Data by Design

A fundamental quality of data repositories is the ease of searching, finding, and reusing data. It is evident from the findings that data repositories underpinned by FAIR Data Principles would enrich the processes and practices of sharing and consuming data via coherent PIDs, standards, metadata, and provenance. Thus far, the principle has not been explicitly applied to study data repositories used daily in academia (see van Reisen et al., 2020). We posit that everyday data, such as data in data repositories or information portals of the academic data ecosystem, should be FAIR by design to ease finding and reusing data for end users.

The operationalisation of FAIR Data Principles would enrich the governance and management of data. Data repositories could mandate data to be FAIR explicitly through policies and frameworks or implicitly by implementing facets of FAIR Data Principles in information systems by design—that is, data repositories implement FAIRification features (Jacobsen et al., 2020). It will encourage those who want to share and reuse data to consider FAIR Data Principles, and eventually, the principle will become second nature in routine data practices.

FAIR Data by design necessitates addressing some challenges. Some vague facets of FAIR Data Principles are difficult to assess in practice, even with adequate knowledge (see Dunning et al. 2017). How does one define 'rich' metadata for digital objects? A situation might demand specificity and granularity in the metadata of data. In contrast, it is not an issue with facets that only require a brief examination of visible attributes in data, such as PIDs or indexed in a searchable resource. In sum, data practitioners should operationalise FAIR Data Principles according to one's requirements in data.

5.2 FAIR Data Repositories

A data repository that simply gathers, integrates, and stores data is insufficient in today's complex sociotechnical systems. It is essential to have data repositories that bear the hallmark of FAIR Data Principles for dealing with a wide assortment of data. Furthermore, as FAIR Data Principles gains traction and uptake in various areas (van Reisen et al., 2020), it will likely inform the paradigm to share and reuse data. Hence, know-how and skills to operationalise FAIR Data Principles, especially by manual means, are crucial for all stakeholders (David et al., 2020). Moreover, a proactive approach to FAIRify data repositories will pay dividends through seamless data governance and management in a data ecosystem.

Table 1 illustrates the traffic-light rating matrix (Dunning et al., 2017) for a quick summary of the FAIRness score of the DiVA portal. One could strive to maximise the score; nonetheless, it should not be a goal in and of itself. The score does not have an intrinsic value. Rather, the score should be used meaningfully to gain insights and knowledge about a data repository to facilitate a smooth data experience for the back office and end users. The assessment of a data repository for FAIRness is not a binary test of 'Pass/FAIR' or 'FAIL/Not FAIR'. If a data repository scores low in FAIRness, it provides an avenue to improve a data repository. The manifestation of the characteristics of FAIR in a particular context is open to subjective interpretation (Mons et al., 2017). Sometimes a repository may perform poorly in the FAIRness due to various factors, such as data being accessible on the Web or other mediums, but insufficiently described (van Reisen et al., 2020) or only using basic README files. It is imperative to consider the context of what needs to be a FAIR data repository.

Table 1: Traffic Light Rating Matrix of the FAIRness score of the DiVA portal.

FAIR Data Principles	15 Facets of FAIR Data Principles Green [G]: compliant, orange [O]: just about/maybe not, red [R]: not compliant, and blue [B]: unclear			
Findable	F1 [G]	F2 [G]	F3 [G]	F4 [G]
Accessible	A1 [G]	A1.1 [G]	A1.2 [G]	A2 [G]
Interoperable	I1 [G]	I2 [O]	I3 [G]	
Reusable	R1 [G]	R1.1 [G]	R1.2 [G]	R1.3 [O]

5.3 Contribution to Knowledge and Limitations of the Study

This study advances the knowledge base of enterprise data management from a fresh perspective, namely the granular exposition of FAIR data by design through a case study. In addition, the study contributes to the literature on FAIR Data Principles by expanding its application in other fields (van Reisen et al., 2020). This study suggests practitioners

implement FAIR Data Principles in their data repositories through data policies underpinned by the principle and embed facets of the principle in the system by design. It would enrich data governance and management for the back office and create seamless data experiences for end users.

Dunning et al. (2017) highlighted the concise nature of the facets of FAIR Data Principles. The possibility of various interpretations is inevitable. Hence, the reproducibility of this study should yield results within an acceptable level of variation of the frame of the principle. This study is qualitative, and we suggest quantitative or mixed methods studies to further develop the concept of FAIR data by design.

6 CONCLUSION

A case study of the DiVA portal in Sweden was conducted to explicate FAIR data by design. This study advances the knowledge base of data management through an in-depth exposition of operationalising FAIR Data Principles in designing information repositories for FAIR data. This study suggests practitioners consider implementing FAIR Data Principles as a data paradigm that underpins their data repositories through policies that underscore the principle and implement facets of the principle in the system by design. Data sources that implement the principle facilitate the production and consumption of findable and reusable data for the end users. In addition, this study acquaints practitioners with a manual assessment of data repositories of interest against FAIR Data Principles and the challenges of realising the principle in practice. In sum, FAIR data by design is a way forward to govern and manage data in a dynamic data ecosystem.

REFERENCES

Boeckhout, M., Zielhuis, G. A., & Bredenoord, A. L. (2018). The FAIR guiding principles for data stewardship: fair enough? *European Journal of Human Genetics, 26*(7), 931-936. Retrieved from https://doi.org/10.1038/s41431-018-0160-0

Celina, R. (2017). H2020 Programme-Guidelines on FAIR Data Management in Horizon 2020. Retrieved from https://policycommons.net/artifacts/1940350/h2020-programme/2692119/

David, R., Mabile, L., Specht, A., Stryeck, S., Thomsen, M., Yahia, M., & Bailo, D. (2020). FAIRness Literacy: The Achilles' heel of applying FAIR principles. *CODATA Data Science Journal, 19*(32), 1-11. Retrieved from http://doi.org/10.5334/dsj-2020-032

Dunning, A., De Smaele, M., & Böhmer, J. (2017). Are the FAIR Data Principles fair? *International Journal of Digital Curation, 12*(2), 177-195. Retrieved from https://doi.org/10.2218/ijdc.v12i2.567

Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., & Evelo, C. T. (2020). FAIR principles: interpretations and implementation considerations. *Data Intelligence, 2*(1-2), 10-29. Retrieved from https://doi.org/10.1162/dint_r_00024

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. Retrieved from https://doi.org/10.1162/dint_a_00028

Landi, A., Thompson, M., Giannuzzi, V., Bonifazi, F., Labastida, I., da Silva Santos, L. O. B., & Roos, M. (2020). The "A" of FAIR–as open as possible, as closed as necessary. *Data Intelligence, 2*(1-2), 47-55. Retrieved from https://doi.org/10.1162/dint_a_00027

Lycett, M. (2013). 'Datafication': Making sense of (big) data in a complex world. *European Journal of Information Systems, 22*(4). Retrieved from https://doi.org/10.1057/ejis.2013.10

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use, 37*(1), 49-56. doi:10.3233/ISU-170824

The DiVA Consortium. (2022). About DiVA. Retrieved from https://www.info.diva-portal.org/about-diva/

van Reisen, M., Stokmans, M., Basajja, M., Ong'ayo, A. O., Kirkpatrick, C., & Mons, B. (2020). Towards the tipping point for FAIR implementation. *Data Intelligence, 2*(1-2), 264-275. Retrieved from https://doi.org/10.1162/dint_a_00049

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., & Bourne, P. E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*(1), 1-9. doi:10.1038/sdata.2016.18

Wilkinson, M. D., Dumontier, M., Sansone, S.-A., Bonino da Silva Santos, L. O., Prieto, M., Batista, D., & Crosas, M. (2019). Evaluating FAIR maturity through a scalable, automated, community-governed framework. *Scientific Data, 6*(1), 1-12. Retrieved from https://doi.org/10.1038/s41597-019-0184-5