Towards a More Reproducible Biomedical Research Environment: Endorsement and Adoption of the FAIR Principles

Alina Trifan(✉) and José Luís Oliveira

DETI/IEETA, University of Aveiro, Aveiro, Portugal
{alina.trifan,jlo}@ua.pt

Abstract. The FAIR guiding Principles for scientific data management and stewardship are a fundamental enabler for digital transformation and transparent research. They were designed with the purpose of improving data quality, by making it Findable, Accessible, Interoperable and Reusable. While these principles have been endorsed by both data owners and regulators as key data management techniques, their translation into practice in quite novel. The recent publication of FAIR metrics that allow for the evaluation of the degree of FAIRness of a data source, platform or system is a further booster towards their adoption and practical implementation. We present in this paper an overview of the adoption and impact of the FAIR principles in the area of biomedical and life-science research. Moreover, we consider the use case of biomedical data discovery platforms and assess the degree of FAIR compatibility of three such platforms. This assessment is guided by the FAIR metrics.

Keywords: Biomedical and life-science research · Data discovery platforms FAIR principles · Reproducible research · FAIR metrics

1 Introduction

The FAIR guiding principles - FAIR stands for Findable, Accessible, Interoperable and Reusable - were proposed with the ultimate goal of reusing valuable research objects [41]. They represent a set of guidelines for turning data more meaningful and reusable. They emphasize on the necessity to make data discoverable and interoperable not just by humans, but by machines as well. These principles do not provide strict rules or standards to comply with, but rather focus on conventions that enable data interoperability, stewardship and compliance against data and metadata standards, policies and practices. They are not standards to be rigorously followed, but rather permissive guidelines.

The principles are aspirational, in that they do not strictly define how to achieve a state of FAIRness. Depending on the needs or constraints of different research communities, they can be open to interpretation. Independently of this openness, they were designed to assist the interaction between those who want to

A. Roque et al. (Eds.): BIOSTEC 2019, CCIS 1211, pp. 453–470, 2020.
https://doi.org/10.1007/978-3-030-46970-2_22

use community resources and those who provide them. When followed, they are beneficial for both data and system owners and users that seek access to these data and systems. These principles have rapidly been adopted by publishers, funders, and pan-disciplinary infrastructure programmes as key data management issues to be taken into consideration. This can be explained as data management closely relates to interoperability and reproducibility [10].

Generic and research-specific initiatives, such as the European Open Science Cloud[1], the European Elixir infrastructure[2] and the USA National Institutes of Health's Big Data to Knowledge Initiative[3] are some of the current initiatives that endorse the FAIR principles and are committed to provide FAIR ecosystems across multi-disciplinary research areas. Moreover, the European Commission has recently made available a set of recommendations and demands for open data research that are explicitly written in the context of FAIR data[4]. Besides these, several European infrastructures aimed at large scale populational and healthcare research, such as the European Health Data Network[5] and Big Data for Better Outcomes[6] promote these principles as the underlying guide for delivering transparent, reproducible and qualitative research.

Straightforward FAIR-dedicated initiatives such as GOFAIR[7], FAIRsFAIR[8] or FAIRSharing[9] work towards building connected infrastructures of FAIR resources, improve their interoperability and reusability and generally adding value to data by capitalizing on the FAIR principles. They make use of infrastructures that already exist in European countries to create a federated approach for turning the FAIR principles a working standard in science. FAIRDOM[10] alike, a web platform uilt for collecting, managing, storing, and publishing dat alika, models, and operating procedures endorses the FAIR guiding principles as improvements to existing research management practices.

With regard to biomedical and life-science data sources, data interoperability and reusability has been a hot topic over the last decade, strongly correlated with the evolution of the so called Big Data in Healthcare. Despite the incremental increase of the use and storage of electronic health records, research communities still tends to use these data in isolation. Unfortunately more than 80% of the datasets in current practice are effectively unavailable for reuse [18]. This is just one of the factors behind the reproducibility crisis that is manifesting in the biomedical arena [24]. Apart from data still being gathered in silos unavailable outside of the owning institution or country, data privacy concerns and unclear

[1] http://eoscpilot.org.
[2] http://www.elixir-europe.org.
[3] http://commonfound.nih.gov/bd2k/.
[4] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.
[5] www.ehden.eu.
[6] http://bd4bo.eu/.
[7] https://www.go-fair.org.
[8] fairsfair.eu.
[9] https://fairsharing.org/.
[10] https://fair-dom.org/about-fairdom/.

data management approaches are critical barriers for sharing and reusing data. The FAIR principles have been enabling the global debate about better data stewardship in data-driven and open science, and they have triggered funding bodies to discuss their application to biomedical and life-sciences systems. A wide adoption of these principles by the data sources and systems that handle biomedical data has the ability to solve this reproducibility crisis, by ensuring secure interoperability among heterogeneous data sources.

In this paper we propose an overview of the adoption of these principles by the biomedical, life-science and in a more bread sense, health related research communities. We review current approaches of FAIR ecosystems and while such system already perform self-assessments of their methodologies for following the FAIR principles, our exhaustive literature search revealed only a handful of such assessments. We therefore overview FAIR practices that have been thoroughly documented. The FAIR principles are identified as system requirements by several data discovery platforms and biomedical infrastructures, although many of them fall short in really exposing a deep evaluation of their adoption. As such, we dive deeper into the challenges of translating these principles into practice. We chose three biomedical data discovery platforms, as a use case for identifying the methodologies through which they follow the FAIR guidelines.

The present manuscript is an extension of the article presented by the same authors at HealthInf 2019, the 12th International Conference on Health Informatics, held in Prague, Czech Republic [34]. In the current manuscript the Introduction includes further insight into the importance and endorsement of the FAIR Principles within the biomedical and healthcare research communities. In this paper we broaden the scope of the research question behind it and we not only proposes an open assessment of three biomedical data discovery platforms, but we complement it with an overview of FAIR self-assessments that have been recently published. Moreover, the assessment done is extended with more insights into how the FAIR metrics were applied. The Discussion takes into consideration the scientific advances that the FAIR principles have enabled so far and argues on possible challenges that are still to be overcome from the practical point of view of their implementation.

This paper is structured in 5 more sections. A detailed presentation of the FAIR principles is covered in Sect. 2, followed by an overview of biomedical and healthcare FAIR-endorsing initiatives in Sect. 3. The adoption of the FAIR principles by these platforms is analyzed in Sect. 4. We then review self-assessing publications and we propose an open FAIR evaluation of 3 biomedical platforms. We discuss the importance of this adoption for the biomedical and life-science research communities and the current challenges in Sect. 5. We draw our final remarks in Sect. 6.

2 FAIR Guiding Principles

The FAIR principles were intended as a set of guidelines to be followed in order to enhance the reusability of any type of data. They put specific emphasis on

enhancing the ability of machines to automatically find and (re)use the data, in addition to supporting its (re)use by individuals. The goal is that, through the pursuit of these principles, the quality of a data source becomes a function of its ability to be accurately found and reused. Although they are currently not a strict requirement, nor a standard in biomedical data handling systems, these principles maximize their added-value, by acting as a guidebook for safeguarding transparency, reproducibility, and reusability.

The FAIR principles as initially proposed by [41] are detailed in Table 1. In a nutshell, if a data source is intended to be FAIR, sufficient metadata must be provided to automatically identify its structure, provenance, licensing and potential uses, without having the need to use specialized tools. Moreover, any access protocols should be declared where they do or do not exist. The use of vocabularies and standard ontologies further benefit to the degree of FAIRness of a data set.

Table 1. The FAIR Guiding Principles as originally proposed in [41].

Findable	**F1**. (meta)data are assigned a globally unique and persistent identifier
	F2. data are described with rich metadata (defined by R1 below)
	F3. metadata clearly and explicitly include the identifier of the data it describes
	F4. (meta)data are registered or indexed in a searchable resource
Accessible	**A1**. (meta)data are retrievable by their identifier using a standardized communications protocol
	A1.1 the protocol is open, free, and universally implementable
	A1.2 the protocol allows for an authentication and authorization procedure, where necessary
	A2. metadata are accessible, even when the data are no longer available
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
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

The way these principles should manifest in reality was largely open to interpretation and more recently some of the original authors revisited the principles,

in an attempt to clarify what FAIRness is [18]. They addressed the principles as a community-acceptable set of rules of engagement and a common denominator between those who want to use a community's resources and those who provide them. An important clarification was that FAIR is not a standard and it is not equal to open. The initial release of the FAIR principles were somehow misleading in the sense that accessibility was associated with open access. Instead, in the recent extended explanation of what these principles really mean, the A in FAIR was redefined as "Accessible under well defined conditions". This means that data do not have to be open, but the data access protocol should be open and clearly defined. In fact, data itself should be "as open as possible, as closed as needed".

The recognition that computers must be capable of accessing a data object autonomously was the core to the FAIR principles since the beginning. The recent re-interpretation of these principles maintains their focus on the importance of data being accessible to autonomous machines and further clarifies on the possible degrees of FAIRness. While there is no such notion as unFAIR, the authors discuss the different levels of FAIRness that can be achieved. As such, the addition of rich, FAIR metadata is the most important step towards becoming maximally FAIR. When data objects themselves can be made FAIR and open for reuse, the highest degree of FAIRness can be achieved. When all of these are linked with other FAIR data, the Internet of FAIR data is reached. Ultimately, when a large number of applications and services can link and process FAIR data, the Internet of FAIR Data and Services is attained.

3 Biomedical and Life-Science FAIRness

Massive amounts of data are currently available and being produced at an unprecedented rate in all domains of life sciences worldwide. The large volume and heterogeneity of data demand rigorous data standards and effective data management. This includes modular data processing pipelines, APIs and end-user interfaces to facilitate accurate and reliable data exchange, standardization, integration, and end user access [31]. Along with biomedical and life sciences research, the biopharmaceutical industry R&D is becoming increasingly data-driven and can significantly improve its efficiency and effectiveness by implementing the FAIR guiding principles. Recent powerful analytical tools such as artificial intelligence, data mining and knowledge extraction would be able to access the data required in the learning process. The implementation of FAIR is a differentiating factor to exploit data so that they can be used more effectively but also for catalysing external collaborations and for leveraging public datasets. FAIR data support such collaborations and enable insight generation by facilitating the linking of data sources and enriching them with metadata [42].

In the healthcare context, similarly to the biopharma industry, current practices are highly data-driven. Machine learning algorithms capable of securely learning from massive volumes of patients' deidentified clinical data is an appealing and noninvasive approach toward personalization. Health personalization is

expected to revolutionize current health outcomes. To reach this goal, a scalable big data architecture for the biomedical domain becomes essential, based on data standardization to transform clinical, biomedical and life science data into FAIR data [33]. With a new perspective on the FAIR principles applied to healthcare, Holub et al. [8] argue that biological material and data should be viewed as a unified resource. This approach would facilitate access to complete provenance information, which they consider a prerequisite for reproducibility and meaningful integration of the data. They even proposed an extension of the FAIR Principles for healthcare, to include additional components such as quality aspects and meaningful reuse of the data, incentives to stimulate effective enrichment of data sets and biological material collections, and privacy-respecting approaches for working with the human material and data.

The NIH Big Data to Knowledge (BD2K) initiative aims to facilitate digitally enabled biomedical research. Within the BD2K framework, the Commons initiative is intended to establish a virtual environment that will facilitate the use, interoperability, and discoverability of digital research objects. It seeks to promote the widespread use of biomedical digital resources by ensuring that they are FAIR. There are four established working subgroups with the aim of bringing some of the high level concepts established by the BD2K Commons into practice, one of them being solely dedicated to the development of FAIRness metrics [9]. In Europe, the German Network for Bioinformatics Infrastructure collects, curates, and shares life-science data. The work of the center is guided by the FAIR principles. This research initiative developed several different tools as contributions to FAIR data, models, and experimental methods storage and exchange [43]. Similarly, FAICE (FAIR Collaboration and Experiments) allow for comprehensive machine-readable description of an experiment that enables replication as well as modification to reuse other input data or a different execution environment [10].

On the genomics spectrum, publicly available gene expression datasets are growing at an accelerating rate. Such datasets hold value for knowledge discovery, particularly when integrated. Although numerous software platforms and tools have been developed to enable reanalysis and integration of genomics datasets, large-scale reuse is harden by minimal requirements for standardized metadata that are ofthen not met. The ultimate goal of initiatives as the Gene Expression Omnibus is to make such repositories more FAIR [39]. Likewise, to unlock the full potential of genome data and to enhance data interoperability and reusability of genome annotations, the Semantic Annotation Platform with Provenance (SAPP) was developed [12]. As an infrastructure supporting FAIR computational genomics, it can be used to process and analyze existing genome annotations. Because managing FAIR genome annotation data requires a considerable administrative load, SAPP stores the results and their provenance in a Linked Data format, thus enabling the deployment of mining capabilities of the Semantic Web.

Access to consistent, high-quality metadata is critical to finding, understanding, and reusing scientific data. The W3C Semantic Web for Health Care and the

Life Sciences Interest Group identified Resource Description Framework (RDF) vocabularies that could be used to specify common metadata elements and their value sets, thereby enabling the publication of FAIR data [4]. The usage of ontologies adds to transforming data from database schemas into FAIR data. An ontology, combined with Semantic Web technologies, are a strong contributor to the FAIRness of a system by facilitating their reproducibility. One such example is the Radiation Oncology Ontology, a platform that contains classes and properties between classes to represent clinical data and their relationships in the radiation oncology domain following the FAIR principles. The ontology along with Semantic Web technologies show how to efficiently integrate and query data from different sources without a priori knowledge of their structures. When clinical FAIR data sources are combined using the mentioned technologies, new relationships between entities are created and discovered, representing a dynamic body of knowledge that is continuously accessible and increasing [33].

4 FAIRness into Practice

While the FAIR principles have been both identified as key requirements of data management systems and endorsed by multiple research infrastructures and funding bodies over the last years, there are only a handful of scientific publications that address the assessment of the FAIR degree of biomedical, life-science or health related platforms or systems. Dataverse [15], for instance, is an open-source data repository software designed to support public community or institutional research repositories. Open PHACTS[11], a data integration platform for drug discovery, UniProt [25], an online resource for protein sequence and annotation data and the EMIF Catalogue [35], are some of the few FAIR self-assessed data discovery and integration platforms.

Another example is Datasets2Tools, a repository indexing bioinformatics analyses applied to datasets and bioinformatics software tools. It provides a platform for not only the discovery of these resources, but to their compliance with the FAIR principles. Users are enabled to grade digital objects according to their compliance with the FAIR principles. When a user submits a FAIR evaluation, its scores are stored in the database, aggregated with the feedback from all other users, and displayed on the corresponding landing pages. The FAIRness evaluation information is also incorporated within the search ranking system, where users can prioritize and identify resources based on their overall FAIRness score [32].

A detailed FAIR assessment is proposed by Rodriguez et al. [27], who describe the process of migrating the Pathogen-Host Interaction Database (PHI-base) to a form that conforms to each of the FAIR Principles. They detail the technical and architectural decisions, including observations of the difficulty of each step. They examine how multiple FAIR principles can be addressed simultaneously through careful design decisions, including making data FAIR for both humans and machines with minimal duplication of effort. They argue that FAIR data

[11] http://www.openphactsfoundation.org/.

publishing involves more than data reformatting and that the sole use of Semantic Web or Linked Data resources is not sufficient for reaching a high level of FAIRness. They explore the value-added by the FAIR data transformation by testing out the result through integrative questions that could not easily be asked over traditional Web-based data resources [27].

Several other examples on how the FAIR principles are translated into practice come from research areas that are either limited by the amount of data available or by the rareness of study events. As such, the Immune Epitope Database (IEDB) has the mission to make published experimental data relating to the recognition of immune epitopes easily available to the scientific public. Vita et al. [38] examine how IEDB complies with the FAIR principles and identify broad areas of success, but also areas for improvement, through a systematic inspection. The IEDB does comply with a number of the FAIR principles to a high standard, but at the same time, several areas for improvement were identified [38]. Another example is the area of pharmacovigillance. OpenPVSignal is an ontology aiming to support the semantic enrichment and rigorous communication of pharmacovigilance signal information in a systematic way. It focuses on publishing signal information according to the FAIR data principles, and exploiting automatic reasoning capabilities upon the interlinked signal report data. OpenPVSignal is developed as a reusable, extendable and machine-understandable model based on Semantic Web standards. An evaluation of the model against the FAIR data principles was performed by Natsiavas et al. [19]. Project Tycho, an open-access database comprising million counts of infectious disease cases and deaths reported for over a century by public health surveillance in the United States was recently upgraded to version 2.0. The main changes reflected in this new version were the use of standard vocabularies to encode data, improving thus compliance with FAIR [21].

In the rare diseases spectrum, the Open Source Registry for Rare Diseases (OSSE) provides a software for the management of patient registries. In this area, networking and data exchange for research purposes remains challenging due to interoperability issues and due to the fact that small data chunks are stored locally. A pioneer in this area, the OSSE architecture was adapted so as to follow the FAIR Data Principles. The so called FAIR Data Point [30] was integrated in order to provide a description of metadata in a FAIR manner. This is an important first step towards unified documentation across multiple registries and the implementation of the FAIR Data Principles in the rare disease area [28].

A metadata model, along with a data sharing framework designed to improve findability and reproducibility of experimental data inspired by FAIR principles were proposed by Karim et al. [11]. The developed system is evaluated against competency questions collected from data consumers, and thereby proven to help to interpret and compare data across studies. The authors follow an incremental approach to achieve optimal FAIRness of the data sets and report on the initial degree of FAIRness that was achieved. Their implementation is not complete in terms of coverage of all FAIR principles, but provides a starting point and a good example of practical applications of Semantic Web technologies for FAIR data sharing.

Table 2. The template for creating FAIR Metrics retrieved from https://github.com/FAIRMetrics.

Field	Description
Metric Identifier	FAIR Metrics should, themselves, be FAIR objects, and thus should have globally unique identifiers
Metric Name	A human-readable name for the metric
To which principle does it apply	Metrics should address only one sub-principle, since each FAIR principle is particular to one feature of a digital resource; metrics that address multiple principles are likely to be measuring multiple features, and those should be separated whenever possible
What is being measured	A precise description of the aspect of that digital resource that is going to be evaluated
Why should we measure it	Describe why it is relevant to measure this aspect
What must be provided	What information is required to make this measurement?
How do we measure it	In what way will that information be evaluated?
What is a valid result	What outcome represents "success" versus "failure"?
For which digital resource(s) is this relevant	If possible, a metric should apply to all digital resources; however, some metrics may be applicable only to a subset. In this case, it is necessary to specify the range of resources to which the metric is reasonably applicable
Example of their application across types of digital resource	Whenever possible, provide an existing example of success, and an example of failure

4.1 FAIR Metrics

Along with the narrative analysis of FAIR principles and their adoption, we propose an assessment following the FAIR metrics recently proposed by some of the original authors of the FAIR guiding principles (Table 2).

The increasing ambiguity behind the initially published principles, along with the need of data providers and regulatory bodies to evaluate their translation into practice led to the establishment of the FAIR metrics group[12], with the purpose of defining universal measures of data FAIRness. Nevertheless, these universal metrics can be complemented by resource-specific ones that can reflect the expectations of one or multiple communities.

[12] http://fairmetrics.org.

4.2 FAIR Use Case: Biomedical Discovery Platforms

The integration and reuse of huge amounts of biomedical data currently available in digital format has the ability to impact clinical decisions, pharmaceutical discoveries, disease monitoring and the way population healthcare is provided globally. Storing data for future reuse and reference has been a critical factor in the success of modern biomedical sciences [26]. In order for data to be reused, first it has to be discovered. Finding a dataset for a study can be burdensome due to the need to search individual repositories, read numerous publications and ultimately contact data owners or publication authors on an individual basis. Recent research shows that the time spent by researchers in searching for and identifying multiple useful data sources can take up to 80% of their time dedicated to the project or research question itself [23].

Biomedical data exists in multiple scales, from molecular to patient data. Health systems, genetics and genomics, population and public health are all areas that may benefit from big data integration and its associated technologies [16]. The secondary reuse of citizens' health data and investigation of the real evidence of therapeutics may lead to the achievement of personalized, predictive and preventive medicine [22]. However, in order for researchers to be able to reuse data and conduct integrative studies, they first have to find the right data for their research. Data discovery platforms are one-stop shops that enable clinical researchers to identify datasets of interest without having to perform individual, extensive searches over distributed, heterogeneous health centers.

There are currently many data discovery platforms, developed either as warehouses or simply aggregators of metadata that link to the original data sources. A warehouse platform, the Vanderbilt approach [3] contains both fully de-identified research data and fully identified research that is made available taking into consideration access protocols and governance rules. A cataloguing toolkit is proposed by Maelstrom Research, built upon two main components: a metadata model and a suite of open-source software applications [1]. When combined, the model and software support implementation of study and variable catalogues and provide a powerful search engine to facilitate data discovery. Disease oriented platforms, such as The Ontario Brain Institute's (Brain-CODE) [37] are designed with a very explicit, yet not limited, purpose of supporting researchers in better understanding a specific disease. Brain-CODE addresses the high dimensionality of clinical, neuroimaging and molecular data related with various brain conditions. The platform makes available integrated datasets that can be queried and linked to provincial, national and international databases. Similarly, the breast cancer (B-CAN) platform [40] was designed as a private cancer data center that enables the discovery of cancer-related data and drives research collaborations aimed at better understanding this disease. Still in the spectrum of cancer discovery, the Project Data Sphere was built to voluntarily share, integrate, and analyze historical cancer clinical trial data sets with the final goal of advancing cancer research [6]. In the rare disease spectrum, RD-Connect [5] links genomic data with patient registries, biobanks, and clinical bioinformatics tools in an attempt to provide a FAIR rare disease complete ecosystem.

Among most established initiatives, Cafe Variome [13] provides a general-purpose, web-based, data discovery tool that can be quickly installed by any genotype–phenotype data owner and turn data discoverable. MONTRA [29], another full-fledged open-source discovery solution, is a rapid-application development framework designed to facilitate the integration and discovery of heterogeneous objects. Both solutions rely on a catalogue for data discovery and include extensive search functionalities and query capabilities.

Linked Data is also explored in discovery platforms, such as YummyData [44] which was designed to improve the findability and reusability of life science datasets provided as Linked Data. It consists of two components, one that periodically polls a curated list of SPARQL endpoints and a second one that monitors them and presents the information measured. Similarly, the Open PHACTS Discovery Platform [7] leverages Linked Data to provide integrated access to pharmacology databases. Still in the spectrum of Linked Data, FAIRSharing is a manually curated searchable portal of three linked registries [17] that cover standards, databases and data policies in the life sciences.

Further contributors to the degree of FAIRness of such systems are APIs that enable machines to discover and interact with FAIR research objects. One such example is the smartAPI[13] [45], which was developed with the aim to make APIs FAIR. It leverages the use of semantic technologies such as ontologies and Linked Data for the annotation, discovery, and reuse of APIs. Considering the diversity, complexity and increasing volume of biomedical research data, Navale et al. argue that cloud based platforms can be leveraged to support several different ingest modes (e.g. machine, software or human entry modes) to make data more FAIR [20].

All these platforms address data discovery from different perspectives, integrating or linking to different types of biomedical data. Another aspect that they share is that they identify the FAIR principles as requirements of their architectures, as well as enablers of data discovery. Although the high majority of these platforms emphasize the importance of providing a way for machines to discover and access the data sets, they are heterogeneous in the way they address the FAIR guidelines. A recent systematic review on biomedical discovery platforms [36] argues that 45% (9 out of 20) of the studies included in the review indicate the FAIR principles as requirements, without providing details on their implementation. For the evaluation that we propose in this paper, we have chosen three of the previously overviewed data discovery platforms that identify the FAIR principles as guidelines for their development. We are keen on understanding their approaches in following the guiding principles. We first overview the scope and methods of these platforms and we present in a narrative form their partial or total compliance with the FAIR principles.

Among the three platforms we chose for this assessment, the Maelstrom Research cataloguing toolkit presented by [1] is built upon two main components: a metadata model and a suite of open-source software applications. The model sets out specific fields to describe study profiles, characteristics of the

[13] www.smart-api.info.

subpopulations of participants, timing and design of data collection events and variables collected at each data collection event. The model and software support implementation of study and variable catalogues and provide a powerful search engine to facilitate data discovery. Developed as an open source and generic tool to be used by a broad range of initiatives, the Maelstrom Research cataloguing toolkit serves several national and international initiatives. The FAIR principles have been identified from early on as a requirement of its architecture. With respect to Findability, each dataset is complemented by rich metadata. To ensure quality and standardization of the metadata documented across networks, standard operating procedures were implemented. In what concerns Accessibility, when completed, study and variable-specific metadata are made publicly available on the Maelstrom Research website. Using information found in peer-reviewed journals or on institutional websites, the study outline is documented and validated by study investigators. Thus, the linkage with other FAIR metadata is achieved. Where possible, data dictionaries or codebooks are obtained, which contributes to the data interoperability.

Many life science datasets are nowadays represented via Linked Data technologies in a common format (the Resource Description Framework). This makes them accessible via standard APIs (SPARQL endpoints), which can be understood as one of the FAIR requirements. While this is an important step toward developing an interoperable bioinformatics data landscape it also creates a new set of obstacles as it is often difficult for researchers to find the datasets they need. YummyData provides researchers the ability to discover and assess datasets from different providers [44]. This assessment can be done in terms of metrics such as service stability or metadata richness. YummyData consists of two components: one that periodically polls a curated list of SPARQL endpoints monitoring the states of their Linked Data implementations and content and another one that presents the information measured for the endpoints and provides a forum for discussion and feedback. It was designed with the purpose to improve the findability and reusability of life science datasets provided as Linked Data and to foster its adoption. Apart from making data available to software agents via an API, the adoption of Linked Data principles has the potential to make data FAIR.

FAIRSharing, originally named Biosharing, is a manually curated searchable portal of three linked registries [17]. These resources cover standards, databases and data policies in the life sciences broadly encompassing the biological environmental and biomedical sciences. The manifest of the initiative is that FAIRSharing makes these resources findable and accessible - the core of the FAIR principle. Every record is designed to be interlinked providing a detailed description not only on the resource itself but also on its relations with other life science infrastructures. FAIRSharing is working with an increasing number of journals and other registries and its focus is to ensures that data standards, biological databases and data policies are registered, informative and discoverable. Thus, it is considered a pivotal resource for the implementation of the ELIXIR-supported FAIR principles.

Our biomedical discovery use-case assessment follows the previously identified FAIRness metrics, applied to each of the 13 items of the FAIR guiding principles. For each of the principles, we outline next the questions that we tried to answer in the evaluation and the name of the metric, within brackets. The following information is a summary of the FAIR metrics description proposed by some of the original authors of the guiding principles[14]:

- F1 (Identifier uniqueness) Whether there is a scheme to uniquely identify the digital resource.
- F1 (Identifier persistence) Whether there is a policy that describes what the provider will do in the event an identifier scheme becomes deprecated.
- F2 (Machine-readability of metadata) The availability of machine-readable metadata that describes a digital resource.
- F3 (Resource identifier in metadata) Whether the metadata document contains the globally unique and persistent identifier for the digital resource.
- F4 (Indexed in a searchable resource) The degree to which the digital resource can be found using web-based search engines.
- A1.1 (Access Protocol) The nature and use limitations of the access protocol.
- A1.2 (Access authorization) Specification of a protocol to access restricted content.
- A2 (Metadata longevity) The existence of metadata even in the absence/removal of data.
- I1 (Use a knowledge representation language) The use of a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2 (Use FAIR Vocabularies) The metadata values and qualified relations should themselves be FAIR, for example, terms from open, community-accepted vocabularies published in an appropriate knowledge-exchange format.
- I3 (Use qualified references) Relationships within (meta)data, and between local and third-party data, have explicit and 'useful' semantic meaning.
- R1.1 (Accessible Usage License) The existence of a license document, for both (independently) the data and its associated metadata, and the ability to retrieve those documents.
- R1.2 (Detailed Provenance) That there is provenance information associated with the data, covering at least two primary types of provenance information: who/what/when produced the data (i.e. for citation) and why/how was the data produced (i.e. to understand context and relevance of the data).
- R1.3 (Meets Community Standards) Certification, from a recognized body, of the resource meeting community standards.

This evaluation allowed us to identify the FAIR requirements already satisfied and the ones that are not undressed, or unclear. Our findings show a high level of FAIRness achieved by the three platforms, mainly favored by the rich metadata with which each of these platform complement the actual data sources. In all cases the metadata can be accessed both by humans and machines through a

[14] https://github.com/FAIRMetrics/Metrics/blob/master/ALL.pdf.

unique and persistent identifier, mostly in the form of an URI. Moreover, the use of FAIR standards and vocabularies contributes to their degree of FAIRness. This is complemented in two of the platforms by the ability to link to other FAIR metadata, which speaks for the data interoperability and reusability. Still related to reusability, the use of Linked Data by two of the platforms is one of its strong enablers. Last but not least, all of the platforms support machine discoverability and access, by providing dedicated APIs. The main unclear aspect was the access protocol, which was not trivial to identify. Another weak point was the lack of quantifiable certification that the resources meet community standards. We present our summarized assessment in Table 3.

With respect to Findability, we can argue that FAIRsharing exposes a higher quality of FAIRness as all resources are identified by truly globally and unique identifiers, following a schema similar to the Digital Object Identifiers used in the case of scientific publications. All three portals have clearly defined and easily findable data description sections, that link to the origin of the data in question. They all support search capabilities for the identification of the resources they hold. Additionally, Google Dataset Search engine indexes FAIRsharing resources. Regarding Accessibility, the only requirement that we were not able to retrieve based on the publications behind these platforms is the one related to the prevalence of the metadata when original date is no longer available. Interoperability wise, apart from using FAIR compliant vocabularies, both FAIRSharing and the Maelstrom Catalogue link to Pubmed publications, when existing. YummyData uses LinkedData technologies, which contributes to its higher interoperability.

Table 3. Assessment of the FAIRness of each of the three discovery platforms based on the FAIRness metrics. X represents a satisfied requirement and - means that no proof to support the requirement was found. This table was originally published in [35].

Platform	F1	F2	F3	F4	A1.1	A1.2	A2	I1	I2	I3	R1.1	R1.2	R1.3
Maelstrom catalogue	X	X	X	X	X	–	–	X	X	X	–	X	–
YummyData	X	X	X	X	X	X	–	X	X	–	X	X	–
FAIRsharing	X	X	X	X	X	X	–	X	X	X	X	X	–

These open applications have benefitted from strong alignment with the FAIR principles, which have facilitated their adoption by many different research bodies. As an example, YummyData and FAIRSharing have been identified by Wise et al. [42] as important FAIR tools, that enable other research projects to transition towards a more FAIR development and data re(use).

5 Discussion

Researchers need tools and support to manage, search and reuse data as part of their research work. In the biomedical area, data discovery platforms, either in the shape of data warehouses or metadata integrators that link to original

data silos support the researcher in the process of finding the right data for a given research topic. However, finding the right data is not sufficient for conducting a study. Data should be not only qualitative and accessible under clear and well-defined protocols, but it should also be interoperable and reusable in order to maximize the research outcomes. The FAIR guiding principles are recommendations on the steps to follow in order to increase the meaningfulness and impact of data and are strongly related to data management. FAIR compliant biomedical data discovery platforms have the ability to support biomedical researchers throughout all the steps from finding the right data source to reusing it for secondary research. This can ultimately lead to better health and healthcare outcomes. Ultimately, these principles give an important contribution to the reproducibility of research.

Big biomedical data creates exciting opportunities for discovery, but are often seen as make difficult for capturing analyses and outputs in forms that are FAIR [14]. The FAIR guiding principles have been widely endorsed by publishers, funders, data owners and innovation networks across multiple research areas. However, up until recently, they did not strictly define how to achieve a state of FAIRness and this ambiguity led to some qualitatively different self-assessments of FAIRness. A new template for evaluating the FAIRness of a data set or a data handling system, recently proposed by some of the original authors of the principles, offers a benchmark for a standardized evaluation of such self-assessments. In this paper we have applied them to three different biomedical data discovery platforms in order to estimate their FAIRness. Moreover, we sought to understand the impact that the adoption of these guidelines has in the quality of the output produced by these platforms and to what degree ensuring data reusability and interoperability turns data more prone to be reused for secondary research.

This analysis revealed that the adoption of the FAIR principles is an ongoing process within the biomedical community. However, the FAIR-compliance of a resource or system can be distinct from its impact. The platforms discussed exposed a high level of FAIRness and an increased concern for enabling data discovery by machines. While FAIR is not equal to Linked Data, Semantic Web technologies along with formal ontologies fulfill the FAIR requirements and can contribute to the FAIRness of a discovery platform.

With digital patient data increasing at an exponential rate and having understood the importance of reusing these data for secondary research purposes, it is highly important to ensure its interoperability and reusability. Recently developed data search engines such as DataMed [2] and Google DataSet Search[15] are powered by machine readable metadata and are a powerful stimulus into turning datasets more FAIR. The assessment of data FAIRness is a key element for providing a common ground for data quality to be understood by both data owners and data users. If up until recently the open interpretation of the FAIR guiding principles could lead to assessment biases, the recently published FAIR metrics support more than ever the implementation of the common ground. For this, the biomedical research community should continue to challenge and refine

[15] https://toolbox.google.com/datasetsearch.

their implementation choices in order to achieve a desirable Internet of FAIR Data and Services.

6 Conclusions

The FAIR principles demand well-defined qualities and properties from data resources but at the same time they allow a great deal of freedom with respect to how they should be implemented. In this work we reviewed current approaches taken in the areas of biomedical and life-science research for putting them into practice and, as a use case, we further evaluated the approaches followed by three different biomedical data discovery platforms in providing FAIR data and services by following the recently published FAIR metrics. These fresh examples highlighted the increasing impact of the FAIR principles among the biomedical and life-science research community. By acting in accordance with the FAIR metrics we, as a community, can reach an agreed basis for the assessment of data quality and not only add value to data, but ensure reproducibility, transparency and ultimately facilitate research collaborations.

Acknowledgements. This work has received support from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968 and from the Integrated Programme of SR&TD SOCA (Ref. CENTRO-01-0145-FEDER-000010). The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

References

1. Bergeron, J., Doiron, D., Marcon, Y., Ferretti, V., Fortier, I.: Fostering population-based cohort data discovery: the Maelstrom research cataloguing toolkit. PLoS ONE **13**(7), e0200926 (2018)
2. Chen, X., et al.: Datamed-an open source discovery index for finding biomedical datasets. J. Am. Med. Inform. Assoc. **25**(3), 300–308 (2018)
3. Danciu, I., et al.: Secondary use of clinical data: the Vanderbilt approach. J. Biomed. Inform. **52**, 28–35 (2014)
4. Dumontier, M., et al.: The health care and life sciences community profile for dataset descriptions. PeerJ **4**, e2331 (2016)
5. Gainotti, S., et al.: The RD-Connect Registry & Biobank Finder: a tool for sharing aggregated data and metadata among rare disease researchers. Eur. J. Hum. Genet. **26**(5), 631 (2018)
6. Green, A.K., et al.: The project data sphere initiative: accelerating cancer research by sharing data. Oncologist **20**(5), 464–e20 (2015)
7. Groth, P., Loizou, A., Gray, A.J., Goble, C., Harland, L., Pettifer, S.: API-centric linked data integration: the open PHACTS discovery platform case study. Web Semant.: Sci. Serv. Agents World Wide Web **29**, 12–18 (2014)
8. Holub, P., et al.: Enhancing reuse of data and biological material in medical research: from FAIR to FAIR-health. Biopreserv. Biobank. **16**(2), 97–105 (2018)
9. Jagodnik, K.M., et al.: Developing a framework for digital objects in the big data to knowledge (BD2K) commons: report from the commons framework pilots workshop. J. Biomed. Inform. **71**, 49–57 (2017)

10. Jansen, C., Beier, M., Witt, M., Frey, S., Krefting, D.: Towards reproducible research in a biomedical collaboration platform following the FAIR guiding principles. In: Companion Proceedings of the 10th International Conference on Utility and Cloud Computing, pp. 3–8. ACM (2017)
11. Karim, M.R., et al.: Towards a FAIR sharing of scientific experiments: improving discoverability and reusability of dielectric measurements of biological tissues. In: SWAT4LS (2017)
12. Koehorst, J.J., van Dam, J.C., Saccenti, E., Martins dos Santos, V.A., Suarez-Diez, M., Schaap, P.J.: SAPP: functional genome annotation and analysis through a semantic framework using FAIR principles. Bioinformatics **34**(8), 1401–1403 (2017)
13. Lancaster, O., et al.: Cafe Variome: general-purpose software for making genotype-phenotype data discoverable in restricted or open access contexts. Hum. Mutat. **36**(10), 957–964 (2015)
14. Madduri, R., et al.: Reproducible big data science: a case study in continuous FAIRness. PLoS ONE **14**(4), e0213013 (2019)
15. Magazine, D.L.: The dataverse network®: an open-source application for sharing, discovering and preserving data. D-lib Mag. **17**(1), 2 (2011)
16. Martin-Sanchez, F., Verspoor, K.: Big data in medicine is driving big changes. Yearb. Med. Inform. **9**(1), 14 (2014)
17. McQuilton, P., et al.: BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. Database **2016** (2016)
18. Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L.O.B., Wilkinson, M.D.: Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the European Open Science Cloud. Inf. Serv. **37**(1), 49–56 (2017)
19. Natsiavas, P., Boyce, R.D., Jaulent, M.C., Koutkias, V.: OpenPVSignal: advancing information search, sharing and reuse on pharmacovigilance signals via FAIR principles and semantic web technologies. Front. Pharmacol. **9**, 609 (2018)
20. Navale, V., McAuliffe, M.: Long-term preservation of biomedical research data. F1000Research **7** (2018)
21. van Panhuis, W.G., Cross, A., Burke, D.S.: Project Tycho 2.0: a repository to improve the integration and reuse of data for global population health. J. Am. Med. Inform. Assoc. **25**(12), 1608–1617 (2018)
22. Phan, J.H., Quo, C.F., Cheng, C., Wang, M.D.: Multiscale integration of -omic, imaging, and clinical data in biomedical informatics. IEEE Rev. Biomed. Eng. **5**, 74–87 (2012)
23. Press, G.: Cleaning big data: most time-consuming, least enjoyable data science task, survey says. Forbes, 23 March 2016
24. Prinz, F., Schlange, T., Asadullah, K.: Believe it or not: how much can we rely on published data on potential drug targets? Nat. Rev. Drug Discov. **10**(9), 712 (2011)
25. Pundir, S., Martin, M.J., O'Donovan, C.: UniProt protein knowledgebase. In: Wu, C.H., Arighi, C.N., Ross, K.E. (eds.) Protein Bioinformatics. MMB, vol. 1558, pp. 41–55. Springer, New York (2017). https://doi.org/10.1007/978-1-4939-6783-4_2
26. Razick, S., Močnik, R., Thomas, L.F., Ryeng, E., Drabløs, F., Sætrom, P.: The eGenVar data management system-cataloguing and sharing sensitive data and metadata for the life sciences. Database **2014** (2014)
27. Rodríguez-Iglesias, A., et al.: Publishing FAIR data: an exemplar methodology utilizing PHI-base. Front. Plant Sci. **7**, 641 (2016)
28. Schaaf, J., et al.: OSSE goes FAIR-implementation of the FAIR data principles for an open-source registry for rare diseases. Stud. Health Technol. Inform. **253**, 209–213 (2018)

29. Silva, L.B., Trifan, A., Oliveira, J.L.: Montra: an agile architecture for data publishing and discovery. Comput. Methods Programs Biomed. **160**, 33–42 (2018)
30. da Silva Santos, L., et al.: FAIR data points supporting big data interoperability. In: Enterprise Interoperability in the Digitized and Networked Factory of the Future. ISTE, London pp. 270–279 (2016)
31. Stathias, V., et al.: Sustainable data and metadata management at the BD2K-lincs data coordination and integration center. Sci. Data **5**, 180117 (2018)
32. Torre, D., et al.: Datasets2Tools, repository and search engine for bioinformatics datasets, tools and canned analyses. Sci. Data **5**, 180023 (2018)
33. Traverso, A., van Soest, J., Wee, L., Dekker, A.: The radiation oncology ontology (ROO): publishing linked data in radiation oncology using semantic web and ontology techniques. Med. Phys. **45**(10), e854–e862 (2018)
34. Trifan, A., Oliveira, J.: FAIRness in biomedical data discovery, pp. 159–166, January 2019. https://doi.org/10.5220/0007576401590166
35. Trifan, A., Oliveira, J.L.: A FAIR marketplace for biomedical data custodians and clinical researchers. In: 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), pp. 188–193. IEEE (2018)
36. Trifan, A., Oliveira, J.L.: Patient data discovery platforms as enablers of biomedical and translational research: a systematic review. J. Biomed. Inform. **93**, 103154 (2019)
37. Vaccarino, A.L., et al.: Brain-CODE: a secure neuroinformatics platform for management, federation, sharing and analysis of multi-dimensional neuroscience data. Front. Neuroinform. **12**, 28 (2018)
38. Vita, R., Overton, J.A., Mungall, C.J., Sette, A., Peters, B.: FAIR principles and the IEDB: short-term improvements and a long-term vision of obo-foundry mediated machine-actionable interoperability. Database **2018** (2018)
39. Wang, Z., Lachmann, A., Ma'ayan, A.: Mining data and metadata from the gene expression omnibus. Biophys. Rev. **11**(1), 103–110 (2018). https://doi.org/10.1007/s12551-018-0490-8
40. Wen, C.H., et al.: B-CAN: a resource sharing platform to improve the operation, visualization and integrated analysis of TCGA breast cancer data. Oncotarget **8**(65), 108778 (2017)
41. Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. Sci. Data **3** (2016)
42. Wise, J., et al.: Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug Discov. Today **24**(4), 933–938 (2019)
43. Wittig, U., Rey, M., Weidemann, A., Mueller, W.: Data management and data enrichment for systems biology projects. J. Biotechnol. **261**, 229–237 (2017)
44. Yamamoto, Y., Yamaguchi, A., Splendiani, A.: YummyData: providing high-quality open life science data. Database **2018** (2018)
45. Zaveri, A., et al.: smartAPI: towards a more intelligent network of web APIs. In: Blomqvist, E., Maynard, D., Gangemi, A., Hoekstra, R., Hitzler, P., Hartig, O. (eds.) ESWC 2017. LNCS, vol. 10250, pp. 154–169. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58451-5_11