FAIRness in Biomedical Data Discovery

Alina Trifan and José Luís Oliveira
IEETA/DETI, University of Aveiro, Portugal

Keywords: Biomedical Data Discovery, FAIR Guidelines, Data Discovery Platforms, FAIR Metrics.

Abstract: The FAIR Guiding Principles are a recent, yet powerful set of recommendations for turning data Findable, Accessible, Interoperable and Reusable. They were designed with the purpose of improving data quality and reusability. Over the last couple of years they have been adopted more and more by both data owners and funders as key data management approaches. Despite their increasing popularity and endorsement by multiple research initiatives from some of the most diverse areas, there are still only a few examples on how these principles have been translated into practice. In this work we propose an open evaluation of their adoption by biomedical data discovery platforms. We first overview current biomedical data discovery platforms that introduce the FAIR guiding principles as requirements of their functioning. We then employ the more recent FAIR metrics for evaluating the degree to which these biomedical data discovery platforms follow the FAIR principles. Moreover, we assess their impact on enabling data interoperability and secondary reuse.

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 (Wilkinson et al., 2016). They represent a set of guidelines for turning data more meaningful and reusable and they emphasize the need of making data discoverable and interoperable by machines. 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 use community resources and those who provide them. When followed, they are beneficial for both data owners and users that seek access to the data. 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 (Jansen et al., 2017).

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 endorsing the FAIR principles and commiting 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].

With regard to biomedical 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, the biomedical community still tends to use these data in isolation. Unfortunately more than 80% of the datasets in current practice are effectively unavailable for reuse (Mons et al., 2017). This is just one of the factors behind the reproducibility crisis that is manifesting in the biomedical

[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

Trifan, A. and Oliveira, J.
FAIRness in Biomedical Data Discovery.
DOI: 10.5220/0007576401590166
In *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2019)*, pages 159-166
ISBN: 978-989-758-353-7

arena (Prinz et al., 2011). Apart from data still being gathered in silos unavailable outside of the owning institution or country, data privacy concerns and unclear 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 systems. A wide adoption of these principles by the data sources and systems that handle biomedical data has the ability to leverage this reproducibility crisis, by ensuring interoperability among heterogeneous data sources.

In this paper we propose an overview of the adoption of these principles by biomedical data discovery platforms, which are considered important enablers of secondary research. In the process of reusing biomedical data for secondary research, discovery platforms play an important part as they provide support for identifying data sources that can answer a given translational research question. We review current biomedical data platforms in order to understand their level of FAIRness. While some of these platforms already performed self-assesments of their methodologies for following the FAIR principles, our exhaustive literature search revealed only a handful of such assessments. Otherwise the FAIR principles are identified as system requirements of several discovery platforms, without a deep evaluation of their adoption. We therefore chose three such platforms and identify the means through which they answer to the FAIR guidelines.

This paper is structured in 5 more sections. A detailed presentation of the FAIR principles is covered in Section 2, followed by an overview of biomedical data discovery platforms in Section 3. The adoption of the FAIR principles by these platforms is analyzed in Section 4. We discuss the importance of this adoption for the biomedical research community in Section 5 and we draw the final remarks in Section 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 (Wilkinson et al., 2016) 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.

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 (Mons et al., 2017). 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 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 accesible 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.

Table 1: The FAIR Guiding Principles as originally proposed in (Wilkinson et al., 2016).

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.

3 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 (Razick et al., 2014). 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 (Press, 2016).

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 (Martin-Sanchez and Verspoor, 2014). 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 (Phan et al., 2012). 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 (Danciu et al., 2014) 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 (Bergeron et al., 2018). 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) (Vaccarino et al., 2018) 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 (Wen et al., 2017) 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 (Green et al., 2015). In the rare disease spectrum, RD-Connect (Gainotti et al., 2018) 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 (Lancaster et al., 2015) 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 (Silva et al., 2018), 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 (Yamamoto et al., 2018) 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 (Groth et al., 2014) leverages Linked Data to provide integrated access to pharmacology databases. Still in the spectrum of Linked Data, BioSharing is a manually curated searchable portal of three linked registries (McQuilton et al., 2016) that cover standards, databases and data policies in the life sciences.

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. For this evaluation, we have chosen three of the previously overviewed data discovey platforms for understanding their approaches in following the FAIR 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.

4 ADOPTION OF THE FAIR PRINCIPLES

In lifesciences, initiatives such as GOFAIR[5] 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. Dataverse (Magazine, 2011), for instance, is an open-source data repository software designed to support public community or institutional research repositories. Another example is FAIRDOM[6], a web platform built for collecting, managing, storing, and publishing data, models, and operating procedures. Both solutions follow the FAIR guiding principles in an attempt to improve research management practices. Open PHACTS[7], a data integration platform for drug discovery, UniProt (Pundir et al., 2017), an online resource for protein sequence and annotation data and the EMIF Catalogue (Trifan and Oliveira, 2018), are some of the few FAIR self-assessed data discovery and integration platforms.

Among the three platforms we chose for this assessment, the Maelstrom Research cataloguing toolkit presented by (Bergeron et al., 2018) 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 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 code-

[5]http://go-fair.org

[6]https://fair-dom.org/about-fairdom/

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

books 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 (Yamamoto et al., 2018). 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.

BioSharing is a manually curated searchable portal of three linked registries (McQuilton et al., 2016). 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 BioSharing 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. BioSharing 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.

4.1 FAIR Metrics

Along with the narrative analysis of their FAIR approaches, 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[8], 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.

The second part of our 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[9]:

- 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.

[8] http://fairmetrics.org

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

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.

- 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 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.

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

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.

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	–
Biosharing	X	X	X	X	X	X	–	X	X	X	X	X	–

secondary research. This can ultimately lead to better health and healthcare outcomes. Ultimately, these principles give an important contribution to the reproducibility of research.

The FAIR guiding principles have been widely endorsed by publishers, funders, data owners and innovation networks across multiple research areas. 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. 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 guidinig 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 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 evaluated the approaches followed by three different biomedical data discovery platforms in providing FAIR data and services. This evaluation was strictly done based on the analysis of the scientific publications describing these platforms. As future work, we intend to extend this assessment by exploring these platforms hands-on, in an attempt to address specific driving medical questions. Nevertheless, these fresh examples highlighted the increasing impact of the FAIR principles among the biomedical research community. Moreover, by acting in accordance with the FAIR metrics we, as a community, can reach an agreed basis for the assessment of data quality.

REFERENCES

Bergeron, J., Doiron, D., Marcon, Y., Ferretti, V., and Fortier, I. (2018). Fostering population-based cohort data discovery: The Maelstrom Research cataloguing toolkit. *PloS one*, 13(7):e0200926.

Danciu, I., Cowan, J. D., Basford, M., Wang, X., Saip, A., Osgood, S., Shirey-Rice, J., Kirby, J., and Harris, P. A. (2014). Secondary use of clinical data: the Vanderbilt approach. *Journal of biomedical informatics*, 52:28–35.

Gainotti, S., Torreri, P., Wang, C. M., Reihs, R., Mueller, H., Heslop, E., Roos, M., Badowska, D. M., Paulis, F., Kodra, Y., et al. (2018). The RD-Connect Registry & Biobank Finder: a tool for sharing aggregated data and metadata among rare disease researchers. *European Journal of Human Genetics*, 26(5):631.

Green, A. K., Reeder-Hayes, K. E., Corty, R. W., Basch, E., Milowsky, M. I., Dusetzina, S. B., Bennett, A. V., and Wood, W. A. (2015). The project data sphere initia-

tive: accelerating cancer research by sharing data. *The oncologist*, 20(5):464–e20.

Groth, P., Loizou, A., Gray, A. J., Goble, C., Harland, L., and Pettifer, S. (2014). Api-centric linked data integration: the open PHACTS discovery platform case study. *Web Semantics: Science, Services and Agents on the World Wide Web*, 29:12–18.

Jansen, C., Beier, M., Witt, M., Frey, S., and Krefting, D. (2017). Towards reproducible research in a biomedical collaboration platform following the FAIR guiding principles. In *Companion Proceedings of the10th International Conference on Utility and Cloud Computing*, pages 3–8. ACM.

Lancaster, O., Beck, T., Atlan, D., Swertz, M., Thangavelu, D., Veal, C., Dalgleish, R., and Brookes, A. J. (2015). Cafe Variome: General-purpose software for making genotype–phenotype data discoverable in restricted or open access contexts. *Human mutation*, 36(10):957–964.

Magazine, D.-L. (2011). The dataverse network®: an open-source application for sharing, discovering and preserving data. *D-lib Magazine*, 17(1/2).

Martin-Sanchez, F. and Verspoor, K. (2014). Big data in medicine is driving big changes. *Yearbook of medical informatics*, 9(1):14.

McQuilton, P., Gonzalez-Beltran, A., Rocca-Serra, P., Thurston, M., Lister, A., Maguire, E., and Sansone, S.-A. (2016). Biosharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. *Database*, 2016.

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., and 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.

Phan, J. H., Quo, C. F., Cheng, C., and Wang, M. D. (2012). Multiscale integration of-omic, imaging, and clinical data in biomedical informatics. *IEEE reviews in biomedical engineering*, 5:74–87.

Press, G. (2016). Cleaning big data: Most time-consuming, least enjoyable data science task, survey says. *Forbes, March*, 23.

Prinz, F., Schlange, T., and Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nature reviews Drug discovery*, 10(9):712.

Pundir, S., Martin, M. J., and O'Donovan, C. (2017). Uniprot protein knowledgebase. *Protein Bioinformatics: From Protein Modifications and Networks to Proteomics*, pages 41–55.

Razick, S., Močnik, R., Thomas, L. F., Ryeng, E., Drabløs, F., and Sætrom, P. (2014). The eGenVar data management system-cataloguing and sharing sensitive data and metadata for the life sciences. *Database*, 2014.

Silva, L. B., Trifan, A., and Oliveira, J. L. (2018). Montra: An agile architecture for data publishing and discovery. *Computer methods and programs in biomedicine*, 160:33–42.

Trifan, A. and Oliveira, J. L. (2018). A FAIR marketplace for biomedical data custodians and clinical researchers. In *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*, pages 188–193. IEEE.

Vaccarino, A. L., Dharsee, M., Strother, S. C., Aldridge, D., Arnott, S. R., Behan, B., Dafnas, C., Dong, F., Edgecombe, K., El-Badrawi, R., et al. (2018). Brain-code: A secure neuroinformatics platform for management, federation, sharing and analysis of multi-dimensional neuroscience data. *Frontiers in neuroinformatics*, 12:28.

Wen, C.-H., Ou, S.-M., Guo, X.-B., Liu, C.-F., Shen, Y.-B., You, N., Cai, W.-H., Shen, W.-J., Wang, X.-Q., and Tan, H.-Z. (2017). B-CAN: a resource sharing platform to improve the operation, visualization and integrated analysis of TCGA breast cancer data. *Oncotarget*, 8(65):108778.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3.

Yamamoto, Y., Yamaguchi, A., and Splendiani, A. (2018). YummyData: providing high-quality open life science data. *Database*, 2018.