Four use cases of FAIR maturity indicators in the life sciences

This manuscript was automatically generated on July 23, 2019.

Authors

- **Serena Bonaretti**
  0000-0003-4264-1773 · sbonaretti · SerenaBonaretti
  Department of Bioinformatics, Maastricht University, The Netherlands

- **Egon Willighagen**
  0000-0001-7542-0286 · egonw · egonwillighagen
  Department of Bioinformatics, Maastricht University, The Netherlands

Abstract

Data reuse is crucial to enhance scientific progress and maximize return on science investments. Given the incremented availability, manual and automatic retrieval of data for new research questions can be challenging. Among the guidelines created to enhance data retrieval, the FAIR (findable, accessible, interoperable, reusable) principles are increasingly adopted at an institutional and funding level. Metrics to assess FAIRness of data repositories are under study and contributions are highly encouraged. In this work, we propose four real use-cases of researchers retrieving data from four different repositories (eNanoMapper, ChEBI, Gene Expression Omnibus, and Array Express) to answer their research questions. For each use case, we harvested data and metadata via application program interface (API) and we calculated FAIR metrics assigning “automatic pass” , “manual pass” and “ not passed” scores. We found [...]. To conclude [...]

Introduction

Data sharing and data reuse are two complementary aspects of modern research activity [cit]. Researchers share their data for a sense of community, to demonstrate integrity of acquired data, and to enhance quality and reproducibility of research work [cit]. In addition, data sharing is supported by the emerging citation system for datasets [1], scientific journals requirements [3], and funding agencies that want to maximize their return on science investments [5]. At the same time, researchers are eager to reuse available data to integrate information coming from different fields [cit], to answer interdisciplinary research questions [4], and to optimize use of fundings [cit]. Although attitudes towards data sharing and reuse are increasingly favorable [12], data discovery and re-use remain difficult in practice [11]. Studies show that 40% of qualitative data sets were never downloaded, and about 25% of data is used just 1-10 times. In addition, data availability decreases 17% per year [10] due to... To happen, data sharing and reuse need appropriate data management, including quality, standardization, ethics, and security [cit].

Data retrieval, the process of identifying and extracting data from a database using a query [2], is one of the main challenges of data reuse. In 2016, the FORCE 11 group proposed guidelines for data reuse in the life sciences and named them FAIR, an acronym for findable, accessible, interoperable, and reusable [3] (principles listed in Table 2). In a short time these guidelines have gained remarkable popularity, and they are currently supported by funding agencies and political entities, such as the European Commission, the National Institutes of Health in the United States, and institutions in Africa and Australia [4]. In addition, many initiatives raised to promote and implement data FAIRness, such as GOFAIR and FAIRsharing. The FAIR principles were specifically conceived as aspirational, and thus they do not specify any technical requirements for implementation, do not represent a standard, and do not imply openness of data [4]. The broad formulation of the FAIR principles created a large spectrum of interpretations and concerns, and raised the need to define data FAIRness evaluators. Some of the authors of the seminal paper proposed a set of FAIR metrics [5], subsequently reformulated as FAIR maturity indicators [6]. At the same time, they invited consortia and communities to create and suggest alternative metrics and tool for evaluation (Table 1). The majority of the proposed tools are manually filled questionnaires that provide a final score for data FAIRness. However, the guidelines stress on the importance of creating “objective, quantitative, machine-interpretable” metrics [5]. Two platforms are currently available: FAIR evaluation services [6] and FAIRshake [7]. The first ... while the second... In addition, there are two studies assessing FAIRness in the databases and repositories. Dunning et al. [8] investigated 37 repositories using a qualitative approach including a traffic-light rating system of FAIRness using a Weber et al. [9] automatically analyzed retrieval of more than a million images from five research data repositories and created absolute and relative metrics.

In this paper, we propose a FAIRness evaluator ... applied to databases used by toxicology and nanomaterial researchers. We created reproducible pipelines using python in Jupyter notebook

Table 1: FAIRness evaluators and studies assessing FAIRness of databases and repositories in the literature. Details of the studies are in Table (Table 2).

Authors	Tool	Manual Assessment	Automatic Assessment			Data / Code Repository
			Code / Language	Metadata Format	Protocol / Library	
FAIRness evaluators						
Wilkinsons et al. [5]	-	x	-	-	-	GitHub
Australian Research Data Commons	FAIR self-assessment tool	x	-	-	-	-

Authors	Tool	Manual Assessment	Automatic Assessment			Data / Code Repository
Commonwealth Scientific and Industrial Research Organization	5 star data rating tool	x	-	-	-	-
Data Archiving and Networked Services	FAIR enough? and FAIR data assessment tool	x	-	-	-	-
GOFAIR consortium	FAIR ImplementationMatrix	x	-	-	-	Open Science Framework
EUDAT2020	How FAIR are your data?	x	-	-	-	Zenodo
Wilkinsons et al. [6]	FAIR evaluation services	-	Ruby on Rails	JSON, Microformat, JSSON-LD, RDFa	nanopublications	GitHub
Clark et al. [7]	FAIRshake	-	Django and python	RDF	Extruct	GitHub
Studies assessing FAIRness of databases and repositories						
Dunning et al. [8]	-	x	-	-	-	Institutional repository
Weber et al. [9]	-	-	python	DataCite	OAI-OMH	GitLab
Our approach	-	x (partially)	Jupyter notebook with python	XML, JSON	request?	GitHub

Materials and methods

Use cases in the life sciences

We evaluated FAIRness for four use-cases where researchers retrieve information from a repository or a database to answer their research question. The first and the second use-cases are real research questions asked by researchers in our department, whereas the third and fourth use-cases are plausible research questions. Use-case name used throughout the paper, research question, and repository or database investigated are:

- Parkinsons_AE: What are the differentially expressed genes between normal subjects and subjects with Parkinson's diseases in the brain frontal lobe? To answer this question, the researcher looked for a dataset in ArrayExpress, a repository for microarray gene expression data based at the European Bioinformatics Institute (EBI), United Kingdom [10];
- NBIA_GEO: What is the function of mutation of WDR45 protein in the brain? In this case, the researcher looked for a dataset to analyze in Gene Expression Onmibus (GEO), a repository containing gene expression and other functional genomics data hosted at the National Center for Biotechnology Information (NCBI), United States [11];
- TiO2_ENM: What are the concentrations of titanium dioxide that kill cells? The fictitious researcher looked for an answer in eNanoMapper, a database containing physical, chemical, and biological identity of nanomaterials [12];
- Caffeine_ChEBI: What are the biological roles of caffeine? In this case, the fictitious researcher looked data in ChEBI (Chemical Entities of Biological Interest), a database containing chemical entities of biological interest hosted at EBI [13].

What is data and what is metadata?

The FAIR guidelines recursively use the terminology *data*, *metadata*, and *(meta)data* .
For our computational implementation, we needed precise definitions of these terms. Accordingly to the Merrian-Webster online dictionary, *data* are "information in digital form that can be transmitted or processed" [14] whereas *metadata* are "data that provides information about other data" [15].
Following these definitions, we considered the answer to the research question as *data*, and the extra information provided in the database about *data* as *metadata*. In addition, we divided *metadata* (M) in subcategories according to the requirements of the FAIR guidelines (indicated with their enumeration):

- M(F2): Information that allows researchers to find the dataset s/he looks for. It coincides with the keywords used in the search;
- M(F3): Data identifier in the repository;
- M(I3): Reference to other metadata;
- M(R1): Further information about data content, other than the search keywords;
- M(R1.1): Data license;
- M(R1.2): Data provenance: author name, publication title, and one author's email address

Metadata corresponding to guidelines F2 and R1 change with the research question, while the remaining metadata are independent from the research question. We did not define M(1.3) as it requires community consensus. In all cases, we assumed that *data* and *metadata* were hosted in the same repository.

Calculating FAIR maturity indicators

As the FAIR principles stress on the importance of *data* and *metadata* being "machine-readable", we collected information about datasets and repositories via application programming interface (API) wherever possible.
We asked the researchers to show us how they searched for their dataset using repositories user interfaces in the browser, and we reproduced their search via API. For the two plausible we did the same. We investigated using three methods:

- API of the repository: We queried the repository to get the dataset. Once found the dataset of interest, we retrieved presence of keywords in metadata (F2), presence of identifier in metadata (F3), metadata are retrievable using a standard communication protocol (A1), reference to other metadata (I3), plurality of metadat (R1), and provenance (R1.2).
- API of re3data.org: We looked for information about persistent identifier (F1 - manual?), license (R1.1)
- API of Google Dataset Search: To see if the dataset is indexed in searchable research (F4)
- Email to repository curators: We asked for information that we could not retrieve via API or to confirm information we found in the online documentation, i.e. policy describing metadata accessibility when data are no longer available (A2), and structure of metadata representation (I1).

Finally, we did not calculate maturity indicators concerning use of FAIR vocabularies (I2), and meet domain-relevant community standards (R1.3).

To calculate the maturity indicators, we built on the already available implementations. As a searchable resource (F4), we used re3data.org, a registry containing metadata of more than 2000 data repositories from various disciplines. re3data.org also provides information about licenses used if the repository provides unique and persistent identifiers (F1).

We assigned 1 when the principle was completely satisfied, 0 when it failed and 0.5 when we entered manual information. We assigned decimal number for the principle F2, where we divided the number of found keywords in the metadata over the number or keywords used by the researcher.

Table 2: FAIR maturity indicators. In our approach, (A) and (M) represent automatic and manual retrieval of information. Acronyms: GUID = Globally Unique IDentifier.

FAIR principle	Wilkinson et al. [6]	Dunning et al. [8]	Weber et al. [9]	Our approach
F1: (meta)data are assigned a globally unique and persistent identifier	The GUID matches a scheme that is globally unique and persistent in FAIRsharing	Persistent identifier is DOI or similar	Pass	Retrieval from re3data.org (M)
F2: data are described with rich metadata (defined by R1 below)	Metadata contains "structured" elements (micrograph, JSON) or linked data (JSON-LD, RDFa)	Title, creator, date, contributors, keywords, temporal and spatial coverage	Q_{geo}, Q_{chrono}	Search keywords are in metadata (A)
F3: metadata clearly and explicitly include the identifier of the data it describes	Metadata contains both its own GUID and the data GUID	DOI of data is in metadata	Pass	Metadata contains dataset ID found in query output (A)
F4: (meta)data are registered or indexed in a searchable resource	The digital resource can be found using web-based search engines	Dataset title searched in google.com or duckduckgo.com	Pass	Dataset title searched in google.com (A)
A.1 (meta)data are retrievable by their identifier using a standardized communications protocol	N/A	HTTP request returns 200	Q_{ret}	HTTP request returns 200 (A)
A1.1 the protocol is open, free, and universally implementable	The resolution protocol is universally implementable with an open protocol	Included in A.1	Q_{ret}	Included in A.1 (A)

FAIR principle	Wilkinson et al. [6]	Dunning et al. [8]	Weber et al. [9]	Our approach
A1.2 the protocol allows for an authentication and authorization procedure, where necessary	The resolution protocol supports authentication and authorization for access to restricted content	Included in A.1	Q_{ret}	Included in A.1 (A)
A2. metadata are accessible, even when the data are no longer available	There is a policy for metadata	Repository has a clear policy statement	N/A	Retrieval from re3data.org (M)
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation	If hash-style metadata (e.g. JSON) or Linked Data are found, pass	Metadata is structured (e.g. Dublin Core)	Pass	Metadata is structured (M)
I2. (meta)data use vocabularies that follow FAIR principles	(meta)data uses vocabularies that are, themselves, FAIR	N/A	N/A	N/A
I3. (meta)data include qualified references to other (meta)data	Metadata contain links that are not from the same source (domain/host)	Links to publications and terms definitions	N/A	Metadata includes reference to other dataset IDs (A)
R1. meta(data) are richly described with a plurality of accurate and relevant attributes	N/A	Metadata provide information on how to reuse a dataset	Q_{geo}, Q_{chrono}	Metadata has more information than search keywords (F2) (A)
R1.1. (meta)data are released with a clear and accessible data usage license	Metadata contains a pointer to the data license	Metadata license is present	Q_{lic}	Retrieval from re3data.org (A)
R1.2. (meta)data are associated with detailed provenance	N/A	Documentation on how data was created	N/A	Author name, email address, publication (A)
R1.3. (meta)data meet domain-relevant community standards	N/A	N/A	N/A	N/A

Results

Table 1

- What passed/failed
- Comparison of outcomes in Figure 1. Data are still work in progress and are not final

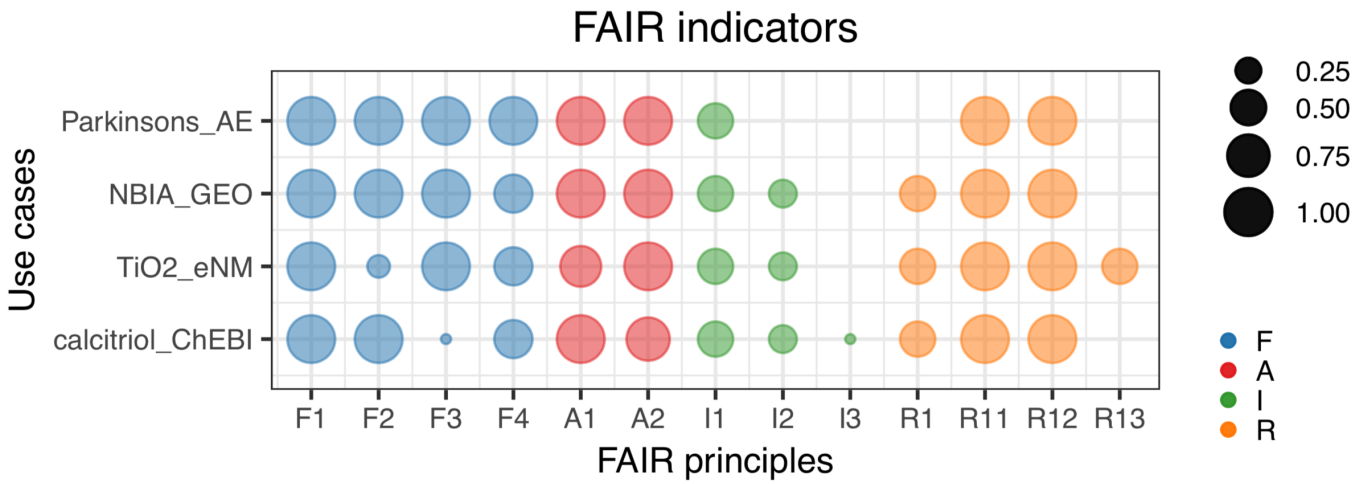

Figure 1: Outcome comparison

Table 3: Use cases.

Use case	Parkinsons_AE	NBIA_GEO	TiO2_ENM	Caffeine_ChEBI
Repository / Database	Array Express	Gene Expression Omnibus	eNanoMapper	ChEBI
Search output on browser	link	link	link	link
API				
API type	REST	REST	REST	SOAP
API documentation	link	link	link	link
Output format	XML, JSON	XML	RDF	XML
Data schema	?	?	?	?
FAIR guidelines				
F1 (Persistent identifier)	No	No		
F2 (Findable metadata)	Parkinson's disease, normal, homo sapiens, transcription profiling by array, raw data, frontal lobe, male, female	nbia, homo sapiens, expression profiling by array	Publication and protocol	
F3 (Unique identifier)	219251	200070433		
F4 (Google search)	?	?		
A1 (Communication protocol)	HTTP	HTTP		
A2 (Metadata always accessible)	?	?		
I1 (Language representation)	help/programmatic_access.html#Format_XML_results			
I2 (FAIR vocabularies)	Not evaluated	Not evaluated		
I3 (Reference to other metadata)	?	?		

Use case	Parkinsons_AE	NBIA_GEO	TiO2_ENM	Caffeine_ChEBI
R1 (Reusable metadata)	?	?		
R1.1 (License)	help/data_availability.html	http://www.ncbi.nlm.nih.gov/geo/info/disclaimer.html		
R1.2 (Provenance)	Garcia-Esparcia P et al.			
R1.3 (Community standards)	Not evaluated			

Discussion

Main points to discuss (in random order):

- We had to adapt the code based on API type and response schema. Our implementation requires specific knowledge of the database structure and thus it is difficult to directly generalize it to various databases
- Comparison to Wilkinson: Our metrics do not start with metadata GUID (general user identifier) (see gen2) but with the researcher's question. Using GUID implies that the researcher has already found the dataset of interest
- Compare to Weber and Dunning
- Database APIs do not allow to retrieve all the information that the user interface allows (example 1: ChEBI does not allow to retrieve information about reactions; example 2: Array Express has some metadata in tables that must be downloaded locally before being queried
- We considered use cases where all queries provided one final dataset. In real practice, researchers often need to compare subset of retrieved datasets manually because there are not enough information to discriminate them computationally (the information is present, but not machine-readable)
- Comments on the findings
- We chose to use Jupyter notebooks for reproducibility of our results. However, databases change but they do not provide versions. Therefore, we can just declare the time stamps when our query was done. In addition, Jupyter notebooks are both machine and human readable, and easier to export to other domains that do not use specifically programming languages designed for the web
- We chose to plot our results instead of providing a final score to avoid negative connotations (see FAIR metrics vs. maturity indicators). However, we wanted to be able to compare our results, so we used balloon plots, usually used for categorical data visualization and comparison. (FAIR shake uses visualizations too but they are not comparable)

Acknowledgments

The NanoSolveIT project is funded by the European Union's xxx under grant agreement no. xxx. The RiskGONE project is funded by the European Union's xxx under grant agreement no. xxx.

We would like to thank Tobias Weber for the insighful discussion about data retrieval, Nasim B. Sangani and xxx for the use case in the department.

This manuscript was created with manubot [16]

References

1. **Credit data generators for data reuse**
Heather H. Pierce, Anurupa Dev, Emily Statham, Barbara E. Bierer
Nature (2019-06) https://doi.org/gf3j9t
DOI: 10.1038/d41586-019-01715-4 · PMID: 31164773

2. **What is Data Retrieval? - Definition from Techopedia**
Techopedia.com
http://www.techopedia.com/definition/30140/data-retrieval

3. **The FAIR Guiding Principles for scientific data management and stewardship**
Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, … Barend Mons
Scientific Data (2016-03-15) https://doi.org/bdd4
DOI: 10.1038/sdata.2016.18 · PMID: 26978244 · PMCID: PMC4792175

4. **Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud**
Barend Mons, Cameron Neylon, Jan Velterop, Michel Dumontier, Luiz Olavo Bonino da Silva Santos, Mark D. Wilkinson
Information Services & Use (2017-03-07) https://doi.org/gfkrrv
DOI: 10.3233/isu-170824

5. **A design framework and exemplar metrics for FAIRness**
Mark D. Wilkinson, Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Olavo Bonino da Silva Santos, Michel Dumontier
Scientific Data (2018-06-26) https://doi.org/gfkrrt
DOI: 10.1038/sdata.2018.118 · PMID: 29944145 · PMCID: PMC6018520

6. **Evaluating FAIR Maturity Through a Scalable, Automated, Community-Governed Framework**
Mark D Wilkinson, Michel Dumontier, Susanna-Assunta Sansone, Luiz Olavo Bonino da Silva Santos, Mario Prieto, Dominique Batista, Peter McQuilton, Tobias Kuhn, Philippe Rocca-Serra, Mercè Crosas, Erik Schultes
Cold Spring Harbor Laboratory (2019-05-28) https://doi.org/gf492b
DOI: 10.1101/649202

7. **FAIRshake: toolkit to evaluate the findability, accessibility, interoperability, and reusability of research digital resources**
Daniel J. B. Clarke, Lily Wang, Alex Jones, Megan L. Wojciechowicz, Denis Torre, Kathleen M. Jagodnik, Sherry L. Jenkins, Peter McQuilton, Zachary Flamholz, Moshe C. Silverstein, … Avi Ma'ayan
Cold Spring Harbor Laboratory (2019-06-03) https://doi.org/gf4cm8
DOI: 10.1101/657676

8. **Are the FAIR Data Principles fair?**
Alastair Dunning, Madeleine De Smaele, Jasmin Böhmer
International Journal of Digital Curation (1970-01-01) https://doi.org/gf4bnb
DOI: 10.2218/ijdc.v12i2.567

9. **How FAIR Can you Get? Image Retrieval as a Use Case to Calculate FAIR Metrics**
Tobias Weber, Dieter Kranzlmuller
2018 IEEE 14th International Conference on e-Science (e-Science) (2018-10) https://doi.org/gf4bm9
DOI: 10.1109/escience.2018.00027

10. **ArrayExpress–a public repository for microarray gene expression data at the EBI**
A. Brazma
Nucleic Acids Research (2003-01-01) https://doi.org/fvff5t
DOI: 10.1093/nar/gkg091 · PMID: 12519949 · PMCID: PMC165538

11. **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**
R. Edgar
Nucleic Acids Research (2002-01-01) https://doi.org/fttpkn
DOI: 10.1093/nar/30.1.207 · PMID: 11752295 · PMCID: PMC99122

12. **The eNanoMapper database for nanomaterial safety information**
Nina Jeliazkova, Charalampos Chomenidis, Philip Doganis, Bengt Fadeel, Roland Grafström, Barry Hardy, Janna Hastings, Markus Hegi, Vedrin Jeliazkov, Nikolay Kochev, … Egon Willighagen
Beilstein Journal of Nanotechnology (2015-07-27) https://doi.org/f3p2xj
DOI: 10.3762/bjnano.6.165 · PMID: 26425413 · PMCID: PMC4578352

13. **ChEBI: a database and ontology for chemical entities of biological interest**
K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcantara, M. Darsow, M. Guedj, M. Ashburner
Nucleic Acids Research (2007-12-23) https://doi.org/dzw32t
DOI: 10.1093/nar/gkm791 · PMID: 17932057 · PMCID: PMC2238832

14. **Definition of DATA**https://www.merriam-webster.com/dictionary/data

15. **Definition of METADATA**https://www.merriam-webster.com/dictionary/metadata

16. **manubot/rootstock GitHub repository**
Daniel Himmelstein
GitHub (2019) https://github.com/manubot/rootstock