



Two real use cases of FAIR maturity indicators in the life sciences

This manuscript was automatically generated on August 6, 2019.

Authors

- **Serena Bonaretti**

 [0000-0003-4264-1773](https://orcid.org/0000-0003-4264-1773) ·  [sbonaretti](https://github.com/sbonaretti) ·  [SerenaBonaretti](https://twitter.com/SerenaBonaretti)

Department of Bioinformatics - BiGCaT, NUTRIM, Maastricht University, The Netherlands

- **Egon Willighagen**

 [0000-0001-7542-0286](https://orcid.org/0000-0001-7542-0286) ·  [egonw](https://github.com/egonw) ·  [egonwillighagen](https://twitter.com/egonwillighagen)

Department of Bioinformatics - BiGCaT, NUTRIM, 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 two real use-cases of researchers retrieving data from two different repositories (Array Express and Gene Expression Omnibus) to answer their research questions. *[The following part still requires some work]* For each use case, we harvested data and metadata via application program interface (API) and we calculated FAIR metrics [...]. We found [...]. To conclude [...]

Keywords:

FAIR guidelines
FAIR Maturity indicators Life sciences
Jupyter notebook

Introduction

Data sharing and data reuse are two complementary aspects of modern research activity. Researchers share their data for a sense of community, to demonstrate integrity of acquired data, and to enhance quality and reproducibility of research work [1]. In addition, data sharing is supported by the emerging citation system for datasets, scientific journals requirements, and funding agencies that want to maximize their return on science investments [2], [3]. At the same time, researchers are eager to reuse available data to integrate information that answer interdisciplinary research questions and to optimize use of fundings [4]. *[The following part of this paragraph still requires some work]* 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]. (infrastructure) Data retrieval, the process of identifying and extracting data from a database using a query [5], is one of the main challenges of data reuse.

In 2016, the FORCE 11 group proposed guidelines to increase data reuse in the life sciences. These guidelines aimed to make data findable, accessible, interoperable, and reusable, and were summarized in the acronym FAIR [6] (principles fully listed in Table 2). In a short time the FAIR 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 [7]. In addition, academic and institutional initiatives were launched to promote and implement data FAIRness, such as [GOFAIR](#) and [FAIRsharing](#). Although largely adopted, the FAIR principles do not specify any technical requirement as they are deliberately intended as aspirational [7]. The lack of practical specifications generated a large spectrum of interpretations and concerns and raised the need to define measurements of data FAIRness. Some of the authors of the seminal paper proposed a set of FAIR metrics [8], subsequently reformulated as FAIR maturity indicators [9]. At the same time, they invited consortia and communities to suggest and create alternative evaluators. The majority of the proposed tools are online questionnaires that researchers and repository curators can manually fill to assess the FAIRness of their data (Table 1). However, the FAIR metrics guidelines stress on the importance of creating “objective, quantitative, machine-interpretable” evaluators [8]. Following this criterion, two platforms have recently been developed to automatically compute FAIR maturity indicators: [FAIR Evaluation Services](#) and [FAIRshake](#). The first platform offers evaluation of maturity indicators and compliance tests [9], whereas the second platform provides metrics, rubrics and assessments for registered digital resources [10]. Both platforms provide use-cases for FAIRness assessment, however they do not provide systematic analysis of evaluated datasets and repositories. In the literature, two studies report evaluation of FAIRness for large datasets. Dunning et al. [11] used a qualitative approach to investigate 37 repositories and databases. They assessed FAIRness using a traffic-light rating system that ranges from no to full compliance. Weber et al. [12] implemented a computational workflow to analyze the retrieval of more than a million images from five repositories. They proposed metrics specific for images, including time and place of acquisition to assess image provenance. The first study provides valuable concrete guidelines to assess data FAIRness, whereas the second study constitutes a solid example of computational implementation.

[The following paragraph still requires some work] In this paper, we propose a computational approach to calculate FAIR maturity indicators in the life sciences. We followed the recommendations provided by the Maturity Indicator Authoring Group [9] and we created a visualization tool to summarize and compare FAIR maturity indicators across various datasets and repositories. We tested our approach on two real use cases of data retrieval to answer research questions in system biology.

We evaluated FAIRness for two repositories, i.e. ArrayExpress and Gene Expression Omnibus. Finally, we made our work open and reproducible by using the open language python in a Jupyter notebook/

Table 1: Online FAIR evaluators and studies in the literature assessing FAIRness of data repositories.

Authors	Questionnaire / Platform	Manual Assessment	Automatic Assessment			Data / Code Repository
			Code / Language	Metadata Format	Protocol / Library	
FAIRness evaluators						
Wilkinsons et al. [8]	-	x	-	-	-	GitHub
Australian Research Data Commons	FAIR self-assessment tool	x	-	-	-	-
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. [9]	FAIR evaluation services	-	Ruby on Rails	JSON, Microformat, JSSON-LD, RDFa	nanopublications	GitHub
Clark et al. [10]	FAIRshake	-	Django and python	RDF	Extract	GitHub
Studies assessing FAIRness of repositories						
Dunning et al. [11]	-	x	-	-	-	Institutional repository
Weber et al. [12]	-	-	python	DataCite	OAI-PMH	GitLab
Our approach	-	x (partially)	Jupyter notebook with python	XML, JSON	request?	GitHub

Materials and methods

Use cases in the life sciences

We asked researchers in our department for cases where they looked for datasets in a scientific repository to answer their research questions. We selected two cases that involved different repositories. For each use case, name used throughout the paper, research question, and investigated repository 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 the search engine of ArrayExpress, a repository for microarray gene expression data based at the European Bioinformatics Institute (EBI), United Kingdom [13];
- *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 the search engine of Gene Expression Omnibus (GEO), a repository containing gene expression and other functional genomics data hosted at the National Center for Biotechnology Information (NCBI), United States [14].

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 Merriam-Webster online dictionary, *data* are "information in digital form that can be transmitted or processed" [15] whereas *metadata* are "data that provide information about other data" [16].

Following these definitions, we considered the dataset that researchers analyzed to answer to the research question as *data*, and the additional information provided in the database about *data* as *metadata*. In addition, we subdivided *metadata* (M) in groups 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): Identifier of the dataset in the repository;
- M(I3): Reference to other metadata;
- M(R1): Information about the dataset, other than the search keywords;
- M(R1.1): Data license;
- M(R1.2): Data provenance: publication title, author names, and one author's email address.

In all cases, we assumed that *data* and *metadata* were hosted in the same repository.

Calculating FAIR maturity indicators

Because the FAIR guidelines stress on the importance of *data* and *metadata* being "machine-readable" [cit], we collected information about datasets and repositories via application programming interface (API) wherever possible. We queried three different sources:

- Data repositories ([ArrayExpress](#) and [Gene Expression Omnibus](#)): We programmatically queried each repository using the same keywords researchers had used in their manual query when looking for a dataset. From the obtained metadata, we retrieved information to calculate maturity indicators for the principles F2, F3, I1, I3, R1, and R12;
- Registry of repository: We queried [re3data.org](#), a registry containing information about more than 2000 data repositories from various disciplines [cit]. We used the retrieved information to

computed the maturity indicators for the principles F1, A2, and R12;

- Searchable resource: We queried [Google Dataset Search](#), an emerging search engine specific for datasets, to quantify the principle F4.

The output of queries consisted of information structured in `xml`, which we vectorized for easier string finding. Details about the computation of each specific maturity indicator are in Table 2 and in our Jupyter notebook (interactive on binder). To each maturity indicator, we assigned value 1 if the criterion was satisfied, 0 in the opposite case. The maturity indicator for F2, which we calculated as the ratio between the number of keywords in the dataset metadata over the total number of keywords used by the researcher in the manual query.

We did not evaluate M(I2) because of the and M(1.3) because it requires community consensus, which still has to be formally reached.

Table 2: FAIR principles and corresponding evaluation criteria proposed by the Maturity Indicator Authoring Group [9], a qualitative analysis by Dunning et al. [11], a computational implementation by Weber et al. [12], and our approach. In our approach, *dataset* metadata refers to metadata retrieved from ArrayExpress and Gene Expression Omnibus, whereas *registry* metadata to metadata retrieved from re3data. Acronyms: GUID = Globally Unique Identifier, A = automatic information retrieval, M = manual information retrieval.

FAIR principle [6]	Wilkinson et al. [9]	Dunning et al. [11]	Weber et al. [12]	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	"doi" icon is enabled in www.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 <i>dataset</i> 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	<i>Dataset</i> metadata contains dataset ID (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 found in google.com or duckduckgo.com	Pass	Dataset title found in Google Dataset Search (M)
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	Accomplished if protocol is HTTP	Q_{ret}	Accomplished if protocol is HTTP (A)
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	Accomplished if protocol is HTTP	Q_{ret}	Accomplished if protocol is HTTP (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	"data availability policy" is filled in <i>registry</i> metadata (A)
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	<i>Dataset</i> metadata is structured (e.g. xml) (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	<i>Dataset</i> metadata includes reference to other dataset IDs (M)
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}	<i>Dataset</i> metadata contain 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}	"datalicenseurl" and "datalicenseurl" are filled in <i>registry</i> metadata (A)

FAIR principle [6]	Wilkinson et al. [9]	Dunning et al. [11]	Weber et al. [12]	Our approach
R1.2. (meta)data are associated with detailed provenance	N/A	Documentation on how data was created	N/A	"authors", "email" and "title" are filled in <i>dataset</i> metadata (A)
R1.3. (meta)data meet domain-relevant community standards	N/A	N/A	N/A	N/A

Visualizing FAIR maturity indicators

To summarize and compare the outputs of our calculation, we created a customized balloon plots using the R library ggplot2 [\[17\]](#). In the graph, each row corresponds to a user-case and each column to a FAIR maturity indicator. The size of each shape is the value of a specific FAIR maturity indicator for a particular dataset. Squares represent maturity indicators determined manually, circles depict maturity indicators established automatically, and crosses illustrate the maturity indicators we did not computer. Finally, colors represent the group of principles in the acronym: blue for findable, red for accessible, green for interoperable, and orange for reusable.

Results

For both use cases, metadata contained all keywords used in the manual search (F2), dataset unique identifiers (F3), and additional information for data reuse (R1). In addition, they were structured in `xml` format (I1) and were released with a clear usage license (R11). The protocol used to retrieve all information was HTTP, which is standardized (A1), open, free and universally implementable (A11), and allows for authentication where needed (A12). In both cases, dataset metadata were not assigned a persistent identifiers (F1) and did not reference to other metadata (I3). Finally, the dataset of the use case *Parkinson_AE* was listed in Google Dataset Search (F4) and had detailed provenance (R12), whereas the dataset *NBIA_GEO* did not. Comparative summary of results is in Figure 1, whereas details of findings are in Table 3.

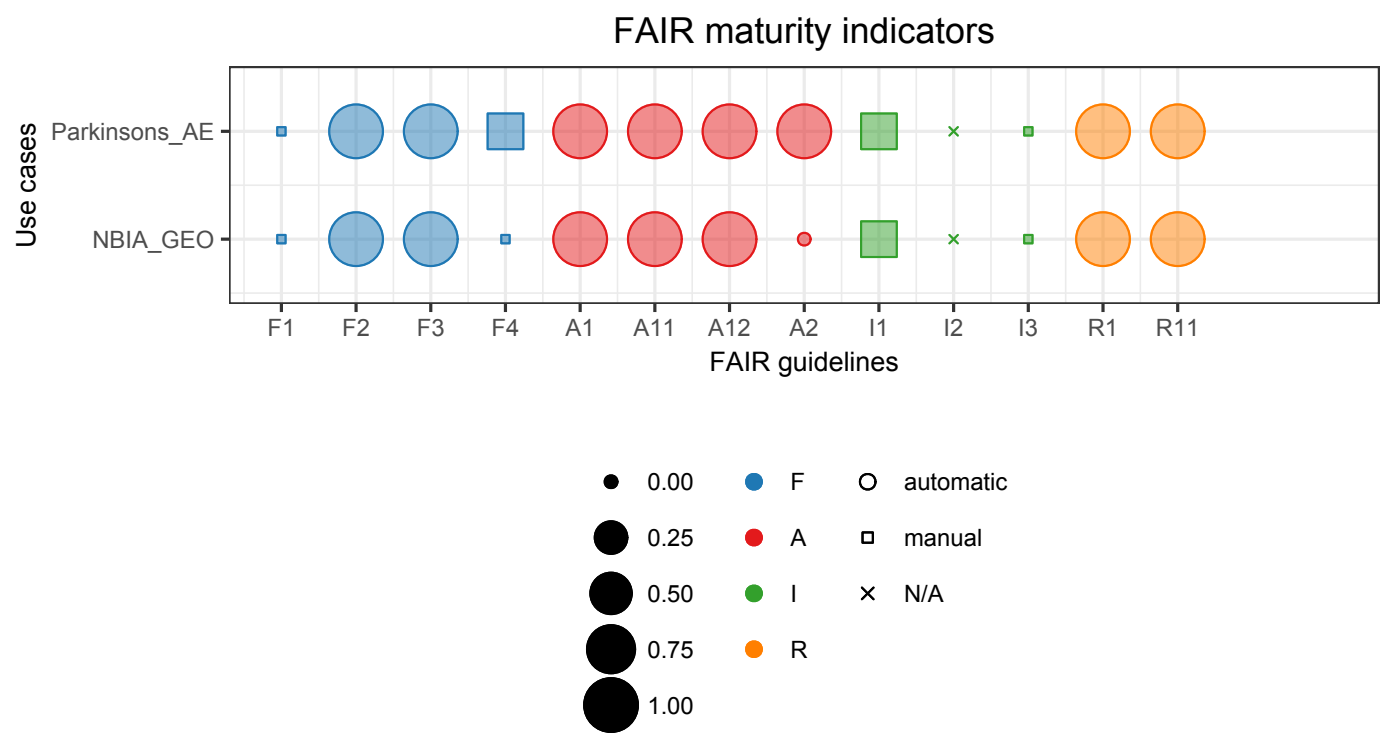


Figure 1: Comparative summary of FAIR maturity indicators for the two use cases evaluated in this work. Shape size corresponds to the numerical value of mutual indicators, colors represent FAIR categories, and shapes illustrate the way we retrieved information (N/A = not available). The graph can be fully reproduced from our [jupyter notebook](#) on GitHub and interactively in [binder](#).

Table 3: Comparison of API systems and FAIR maturity indicators for the two uses cases analyzed in this work. For each maturity indicator, we indicate the outcome in natural language and in numbers (1 for pass and 0 for fail).

Use case	Parkinsons_AE	NBIA_GEO
Repository / Database	Array Express	Gene Expression Omnibus
Search output on browser	link	link
API		
Type	REST	REST
Documentation	link	link
Output format	XML	XML
FAIR maturity indicators		
F1 (Persistent identifier)	No (0)	No (0)
F2 (Findable metadata)	Parkinson's disease, normal, homo sapiens, transcription profiling by array, raw data, frontal lobe, male, female (1)	nbia, homo sapiens, expression profiling by array (1)
F3 (Unique identifier)	219251 (1)	200070433 (1)
F4 (Google Dataset Search)	Yes (1)	No (0)

Use case	Parkinsons_AE	NBIA_GEO
A1 (Communication protocol)	request status code = 200 (1)	request status code = 200 (1)
A11 (Open and free protocol)	Yes (1)	Yes (1)
A12 (Communication protocol)	Yes (1)	Yes (1)
A2 (Metadata always accessible)	Yes: https://www.ebi.ac.uk/arrayexpress/help/data_availability.html (1)	No (0)
I1 (Language representation)	XML (1)	XML (1)
I2 (FAIR vocabularies)	Not evaluated (None)	Not evaluated (None)
I3 (Reference to other metadata)	No (0)	No (0)
R1 (Metadata for reuse)	56 metadata fields (1)	58 metadata fields (1)
R1.1 (License)	name: other url: https://www.ebi.ac.uk/arrayexpress/help/data_availability.html (1)	name: other url: http://www.ncbi.nlm.nih.gov/geo/info/disclaimer.html (1)
R1.2 (Provenance)	Authors: Garcia-Esparcia P, Schlüter A, Carmona M, Moreno J, Ansoleaga B, Torrejón-Escribano B, Gustincich S, Pujol A, Ferrer I Email: aschluter@idibell.org Title: Functional genomics reveals dysregulation of cortical olfactory receptors in parkinson disease: novel putative chemoreceptors in the human brain (1)	No (0)
R1.3 (Community standards)	Not evaluated (None)	Not evaluated (None)

Discussion

Heavily work in progress

In this study, we created an approach to calculate FAIR maturity indicators in the life science. We tested our approach on two use cases that are actual research questions asked in our department.

Discussion of the four groups.

We present a workflow to compute FAIR maturity indicators for data repositories in the life sciences.

for two use-cases of dataset retrieval in the life sciences.

Discussion of results

Comparison with other papers - create graph for dunning? - Comments on the findings

How we calculated the metrics *Criteria* We developed criteria to assess FAIRness based on the guidelines proposed by the group and from the practical implementations by Dunning and Weber.

We developed a mixed manual and automatic approach to assess FAIR metrics. Dunning et al. used a manual approach, while Weber et al. used a fully computational approach. We opted for a mixed approach because the guidelines stress on the importance of the fact that information should be retrieved by machines, therefore a complete manual approach is not desirable. On the other side, a fully computational approach is currently not possible because not all information are retrievable via API and in fact, Weber et al. calculated 2/3 of metrics. Specifically, we used a mixed automatic and manual implementation to retrieve all the possible information with the hope that information now retrieved manually will in the close future be possible to retrieve them automatically. Both the manual and the automatic retrieval of information required a large amount of time and the knowledge of data structure and schema. Querying data repository required knowledge of metadata structure and of data field (i.e. xml tags). Specifically, we had to know in advance some keywords, such as "author", "email". Similarly, we had to know the fields in re3data, such as "data availability policy" for A2 "datalicensename" and "datalicenseurl" for R1.1. In addition, we had to implement a manual change. For example, when assessing criteria F2 (keywords are in metadata), we had to change the keyword "true" that we used for data retrieval into "rawdata". To simplify our computations, we put all tags lower case and vectorized for quick search.

We also simulated data retrieval. This is different from Wilkison because their metrics start with a globally unique identifier (GUID). However, when researchers look for data they do not know already the identifiers, they do not simulate what researchers do. In addition, this workflow would preclude the analysis of our two repositories as they do not provide GUIDs (F1).

We extracted information about the repositories from re3data. These information were about DOI, availability policy of metadata when data are not available, and data license. We selected only one registry (re3data) and one search engine (Google Dataset search). In the first case, alternative can be FAIRshare, however it still does not provide an open API. On the other side, we could have chosen a generic search engine, like google, but we considered pertinent to look for a dataset specifically in a dataset search engine. To compute the metrics we follow

Similarly to Dunning et al. and to Weber et al. we did not computer two principles. The first is the principle I2, i.e. presence of a fair metadata vocabulary. The difficulty arises from the fact that The

second is R13, which is that metadata follow community standards at those have not been formally established yet.

comparison with dunning: mainly the criteria especially difference between metadata for findable and reusable. comparison with Weber: re3data, comparison with Wilkinson: assumes that the GUID is already known thus skipping the finding of the dataset - this is not compatible with our findable criteria

Similarity with all cases:

Practical implementation Before calculating the metrics we decided to simulate dataset retrieval via API. We assumed that a researcher does not know a priori the. Sometimes you don't get only 1 datasets

Difficulties with the three repositories To answer our criteria, we queried three different kinds of repositories: Data repositories, re3data and Google Dataset Search.

We chose re3data as it provides an API for queries. Other registries, such as FAIRshare, do not provide open API yet.

Similarly, Google Dataset Search does not provide any API, so we did a manual search.

Manual vs automatic retrieval

Comparison to the other works

- 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

Limitations of current implementation - Limitations: A1, A11, A12: retrieve information only via HTTP, so they are all true. If data is on a local excel file or database, then we need a different implementation - Repositories not databases

Visualizing the FAIR maturity indicators

We visualized t This visualization allowed us to summarize FAIR evaluation and compare the results of various dataset and implementations. We chose not to create a final score in accordance (to uniform) with the recommendations for the FAIR guidelines that want to keep suggestions and not to assign a score. The summary of metrics is provided by the fact that we exploited shapes, colors, and sizes to put all possible information. On the other side, the fact that each row represents a dataset allows for comparison among datasets. The platform FAIRshake provides visualizations that can dynamically expand to fit scores calculated using different metrics. Scores are in a range that goes from blue (satisfactory) to red (unsatisfactory). Differently, we chose a static approach because these the FAIRshake insignias do not allow for comparison has implemented visualizations called "insignias", where they use color gradients

- 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)

- NO final score because

Reproducibility - 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

Difficulty / Limitations

Some limitations must be acknowledged. First, our approach requires an a-priori knowledge of metadata structure.

Different repositories use different data structure. Automatization occurs after a lot of manual investigation

- 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
- 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)
Our implementation requires specific knowledge of the database structure and thus it is difficult to directly generalize it to various databases.
Different repositories use different html tags to define their
- 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)

General considerations All the information needed are retrieved from metadata, not from data. We did not investigate databases, but only repositories

Conclusion: - it would be great if all repos had similar schema - it would be great if we could access everything via API

Acknowledgments

This work received funding from from the European Union's Horizon 2020 research and innovation programme via NanoSolveIT Project under grant agreement No 814572 and via RiskGONE Project under grant agreement No 814425. We would like to thank Nasim B. Sangani, Gwen Keulen, and Friederike Ehrhart for the use cases, Tobias Weber for the insightful discussion about data retrieval, and Lauren Dupuis for revising our manuscript.

We created this manuscript using [manubot](#) [18].

References

1. Attitudes and norms affecting scientists' data reuse

Renata Gonçalves Curty, Kevin Crowston, Alison Specht, Bruce W. Grant, Elizabeth D. Dalton
PLOS ONE (2017-12-27) <https://doi.org/gcrjn3>
DOI: [10.1371/journal.pone.0189288](https://doi.org/10.1371/journal.pone.0189288) · PMID: [29281658](https://pubmed.ncbi.nlm.nih.gov/29281658/) · PMCID: [PMC5744933](https://pubmed.ncbi.nlm.nih.gov/PMC5744933/)

2. 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](https://doi.org/10.1038/d41586-019-01715-4) · PMID: [31164773](https://pubmed.ncbi.nlm.nih.gov/31164773/)

3. Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data

Heather A. Piwowar
PLoS ONE (2011-07-13) <https://doi.org/cqvpdd>
DOI: [10.1371/journal.pone.0018657](https://doi.org/10.1371/journal.pone.0018657) · PMID: [21765886](https://pubmed.ncbi.nlm.nih.gov/21765886/) · PMCID: [PMC3135593](https://pubmed.ncbi.nlm.nih.gov/PMC3135593/)

4. Exploring visual representations to support data re-use for interdisciplinary science

Andrea Wiggins, Alyson Young, Melissa A. Kenney
Proceedings of the Association for Information Science and Technology (2018) <https://doi.org/gf5shc>
DOI: [10.1002/pra2.2018.14505501060](https://doi.org/10.1002/pra2.2018.14505501060)

5. What is Data Retrieval? - Definition from Techopedia

Techopedia.com
<https://www.techopedia.com/definition/30140/data-retrieval>

6. 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](https://doi.org/10.1038/sdata.2016.18) · PMID: [26978244](https://pubmed.ncbi.nlm.nih.gov/26978244/) · PMCID: [PMC4792175](https://pubmed.ncbi.nlm.nih.gov/PMC4792175/)

7. 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/gfkrvv>
DOI: [10.3233/isu-170824](https://doi.org/10.3233/isu-170824)

8. 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/gfkrvt>
DOI: [10.1038/sdata.2018.118](https://doi.org/10.1038/sdata.2018.118) · PMID: [29944145](https://pubmed.ncbi.nlm.nih.gov/29944145/) · PMCID: [PMC6018520](https://pubmed.ncbi.nlm.nih.gov/PMC6018520/)

9. 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](https://doi.org/10.1101/649202)

10. 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](https://doi.org/10.1101/657676)

11. 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](https://doi.org/10.2218/ijdc.v12i2.567)

12. How FAIR Can you Get? Image Retrieval as a Use Case to Calculate FAIR Metrics

Tobias Weber, Dieter Kranzlmüller

2018 IEEE 14th International Conference on e-Science (e-Science) (2018-10) <https://doi.org/gf4bm9>

DOI: [10.1109/escience.2018.00027](https://doi.org/10.1109/escience.2018.00027)

13. 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](https://doi.org/10.1093/nar/gkg091) · PMID: [12519949](https://pubmed.ncbi.nlm.nih.gov/12519949/) · PMCID: [PMC165538](https://pubmed.ncbi.nlm.nih.gov/PMC165538/)

14. 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](https://doi.org/10.1093/nar/30.1.207) · PMID: [11752295](https://pubmed.ncbi.nlm.nih.gov/11752295/) · PMCID: [PMC99122](https://pubmed.ncbi.nlm.nih.gov/PMC99122/)

15. Definition of DATA <https://www.merriam-webster.com/dictionary/data>

16. Definition of METADATA <https://www.merriam-webster.com/dictionary/metadata>

17. ggplot2

Hadley Wickham

Springer New York (2009) <https://doi.org/djmzjq>

DOI: [10.1007/978-0-387-98141-3](https://doi.org/10.1007/978-0-387-98141-3)

18. Open collaborative writing with Manubot

Daniel S. Himmelstein, Vincent Rubineti, David R. Slochower, Dongbo Hu, Venkat S. Malladi, Casey S. Greene, Anthony Gitter

PLOS Computational Biology (2019-06-24) <https://doi.org/c7np>

DOI: [10.1371/journal.pcbi.1007128](https://doi.org/10.1371/journal.pcbi.1007128) · PMID: [31233491](https://pubmed.ncbi.nlm.nih.gov/31233491/)