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

This manuscript was automatically generated on August 5, 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 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 [...]

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 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 [7]. In addition, many initiatives were launched 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 [7]. The broad formulation of the FAIR principles created a large spectrum of interpretations and concerns, and raised the need to define evaluators 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 evaluate 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 for automatic evaluation of 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 assess FAIRness using a traffic-light rating system that ranges from full to no 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 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: Available tools and studies to assess FAIRness. FAIRness evaluators and studies assessing FAIRness of databases and repositories in the literature. Details of the studies are in Table (Table 2).

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	Extruct	GitHub
Studies assessing FAIRness of databases and 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 evaluated FAIRness for two real use-cases from our department where researchers retrieve information from a repository to answer their research question. 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 [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 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 [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 Merrian-Webster online dictionary, *data* are "information in digital form that can be transmitted or processed" [15] whereas *metadata* are "data that provides information about other data" [16].
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

Because the FAIR guidelines 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 retrieved information from three different sources:

- Data repositories (ArrayExpress and Gene Expression Omnibus): We programmatically queried each repository using the same keywords that researchers had used in their manual query. 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. 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 xml structured information, which we vectorized for easier string finding. Details about the computation of each specific maturity indicator are in Table 2 and in the Jupyter notebook (interactive on binder). To each maturity indicator, we assigned value 1 if the criterion was satisfied, 0 in the opposite case. The only exception is the maturity indicator related to F2, which we calculated as the ratio between the number of keywords in the dataset metadata and the total number of keywords used by the researcher in the manual query.

Table 2: FAIR principles and evaluation criteria used by the promoters, a qualitative analysis, a quantitative calculation, 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 *dataset* 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	*Dataset* 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 (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	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 completed in *registry* 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	*Dataset* 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	*Dataset* 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}	*Dataset* metadata contain more information than search keywords (F2) (A)

FAIR principle [6]	Wilkinson et al. [9]	Dunning et al. [11]	Weber et al. [12]	Our approach
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}	"datalicensename" and "datalicenseulr" are filled in *registry* metadata (A)
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 *dataset* 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 balloon plots using the R library ggplot2 [17]. In the graph, each row corresponds to a user-case and each column to 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 are for maturity indicators established automatically, and crosses for not computed maturity indicators. Finally, colors represent a group of principles of the acronym: blue for findable, red for accessible, green for interoperable, and orange for reusable.

Results

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

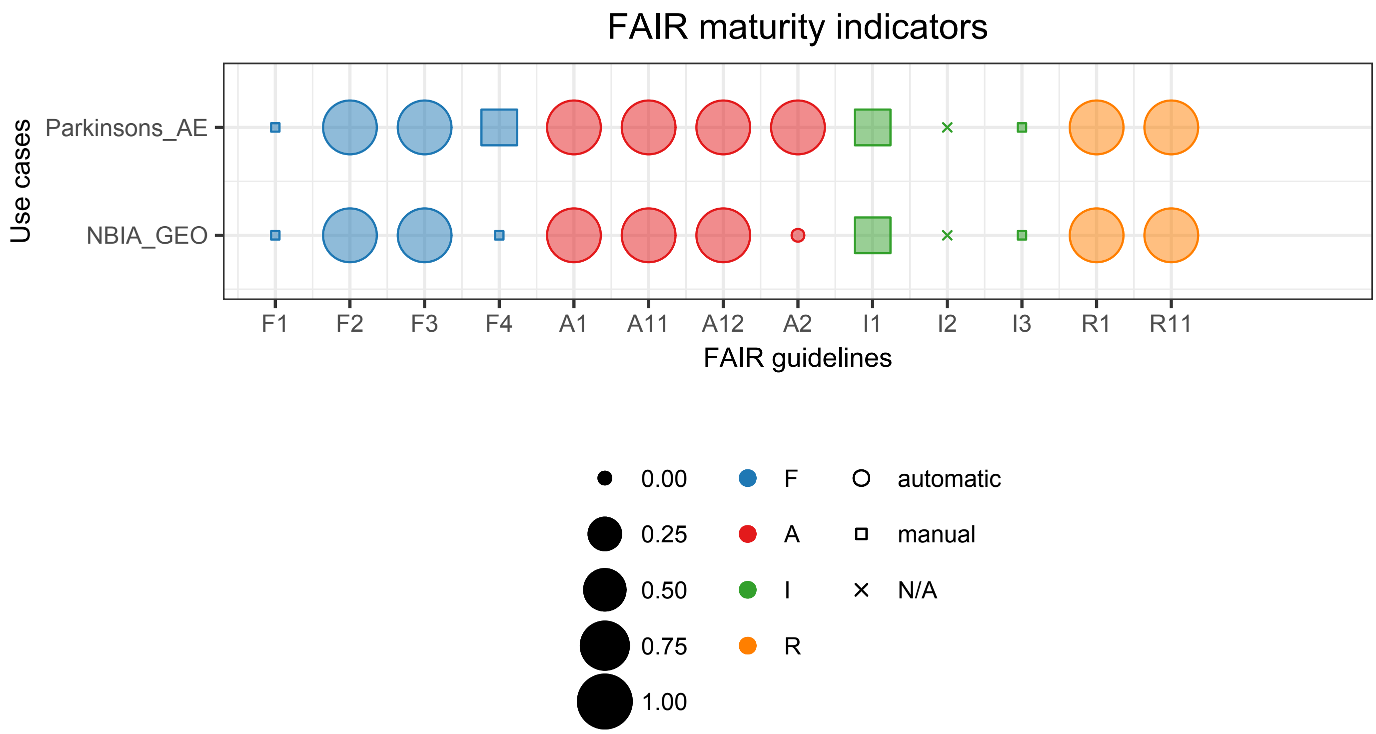

Figure 1: Comparative summary of FAIR maturity indicators for the two use cases evaluated in this work. Shape size corresponds to the numerical value of mutual indicators, colors represent FAIR categories, and shapes illustrate the way we retrieved information (N/A = not available). The graph can be 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 (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 field (1)	58 metadata field (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, Snsoleaga 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

This is still quite a lot work in progress, ready in 1-2 days

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 Before calculating the metrics we decided to simulate dataset retrieval via API. We assumed that a researcher does not know a priori the

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

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”

Similary, 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 simplyfy our computations, we put all tags lower case and vectorized for quick 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.

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 the we exploited shapes, colors, and sizes to put all possible information. On the other side, the fact that each row represent a dataset allows for comparison among datasets The platform FAIRshake provide visualizations that can dinamically 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 “insignas”, where they use color gradents

- 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

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 · PMID: 29281658 · PMCID: 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 · PMID: 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 · PMID: 21765886 · PMCID: 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

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 · PMID: 26978244 · PMCID: 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/gfkrrv
DOI: 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/gfkrrt
DOI: 10.1038/sdata.2018.118 · PMID: 29944145 · PMCID: 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

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

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

12. **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

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 · PMID: 12519949 · PMCID: 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 · PMID: 11752295 · PMCID: 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

18. **Open collaborative writing with Manubot**
Daniel S. Himmelstein, Vincent Rubinetti, 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 · PMID: 31233491