FAIR maturity indicators in the life sciences

Ammar Ammar
ORCID: 0000-0002-8399-8990

PhD candidate
BiGCAT, NUTRIM, FHML, Maastricht University

14-10-2020

Maastricht University

Acknowledgment

Serena Bonaretti
Transparent MSK Research, Maastricht, The Netherlands (https://tmskr.github.io/)

Laurent Winckers, Jeaphianne van Rijn and Egon Willighagen
Department of Bioinformatics - BiGCaT, NUTRIM, Maastricht University, The Netherlands

Joris Quik, Martine Bakker
National Institute for Public Health and the Environment (RIVM), NL-3720 BA Bilthoven, The Netherlands

Dieter Maier
Biomax Informatics AG, Planegg, Germany

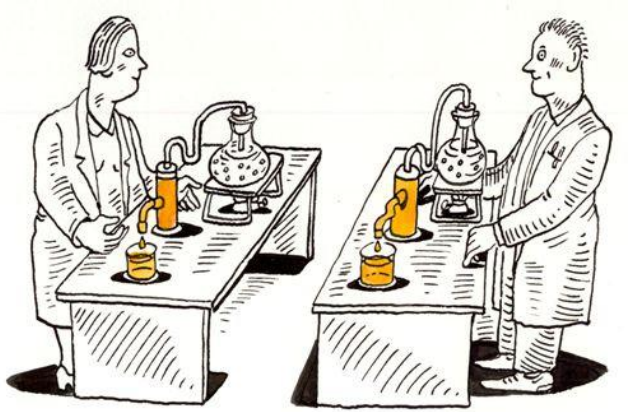
Iseult Lynch
School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, UK

Why do we need FAIR?

- Data sharing and reuse are beneficial for time efficiency and increased productivity in scientific research.
- Data reuse remains difficult → lack of infrastructures, standards, and policies.
- FAIR (findable, accessible, interoperable, reusable) aim to provide guidance to increase data discovery and reuse.
- Maturity indicators are a way to assess the FAIRness of a dataset.

Is research FAIR enough?

- 40% of qualitative datasets were never downloaded [1].
- About 25% of data is used less than 10 times [1].
- Reproducibility of landmark studies are strikingly low:
  - 39% in psychology [2]
  - 21% in pharmacology [3]
  - 11% in cancer [4]
- The availability of existing datasets associated with published articles decreases 17% per year [5], why?



[1] Libby Bishop, SAGE Open, 2017
[2] Monya Baker, Nature, 2015
[3] Florian Prinz et.al, Nature reviews, 2011
[4] C. Glenn Begley et.al, Nature, 2012
[5] Timothy H. Vines et. al, Current Biology, 2014

Why do we need maturity indicators?

- FAIR principles do not specify how to implement them.

Why do we need maturity indicators?

- FAIR principles do not specify how to implement them.

- Lack of practical specifications:
  - generated a large spectrum of interpretations and concerns.
  - raised the need to define measurements of data FAIRness .

Why do we need maturity indicators?

- FAIR principles do not specify how to implement them.
- Lack of practical specifications:
  - generated a large spectrum of interpretations and concerns.
  - raised the need to define measurements of data FAIRness .
- The majority of the proposed tools are online questionnaires
  - researchers and repository curators manually assess the FAIRness of their data.

Why do we need maturity indicators?

- FAIR principles do not specify how to implement them.
- Lack of practical specifications:
  - generated a large spectrum of interpretations and concerns.
  - raised the need to define measurements of data FAIRness .
- The majority of the proposed tools are online questionnaires
  - researchers and repository curators manually assess the FAIRness of their data.
- The FAIR metrics guidelines emphasize the importance of creating “objective, quantitative, and machine-interpretable” evaluators.

Problem statement

- Data reusability in the life sciences domain is hard to quantify.
- FAIR assessment is mostly done manually, which makes the process slow and less objective.
- We lack the means of comparing the FAIRness of life sciences data in a visual easy-to-read manner.

Research Aim

- Develop a computational approach to calculate 12 FAIR maturity indicators in the life sciences domain proposed by [6] and [7].
- Apply it on several datasets/databases with toxicology and/or nanotoxicology related data.
- Create a visualization tool to summarize and compare FAIR maturity indicators across various datasets.

[6] Mark D. Wilkinson et. al, Sci. Data, 2016
[7] Annika Jacobsen et. al, Data Intell., 2020

Box 2 | The FAIR Guiding Principles

To be 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

To be 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

To be 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

To be 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

Materials and methods

- Three uses case, six databases:



- For example:
  - What can eNanoMapper database tell us about nanoscale titanium dioxide (TiO_2) toxicity?

Materials and methods

- Three uses case, six databases:



- For example:
  - What can eNanoMapper database tell us about nanoscale titanium dioxide (TiO_2) toxicity?

- importance of *data* and *metadata* being “machine-interpretable” -> we collected information application programming interfaces (API).

Materials and methods

- Three uses case, six databases:



- For example:
  - What can eNanoMapper database tell us about nanoscale titanium dioxide (TiO_2) toxicity?

- importance of *data* and *metadata* being “machine-interpretable” -> we collected information application programming interfaces (API).

- We queried re3data.org to compute the maturity indicators for the principles F1, A2, and R1.2, related to providing persistent global identifier, metadata data policy and metadata provenance respectively.

Materials and methods

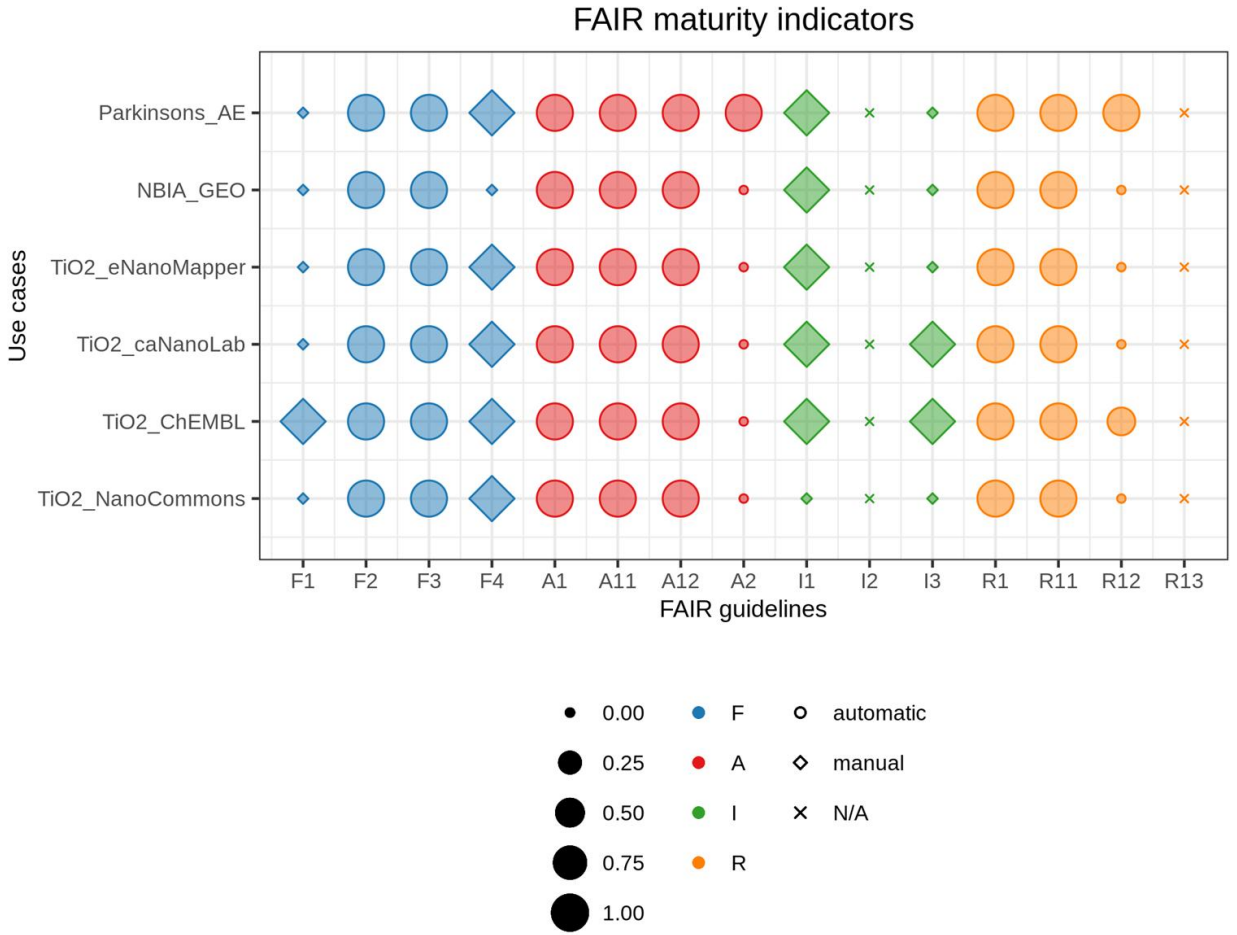
- Searchable resource: We queried Google Dataset Search, an emerging search engine specific for datasets, to quantify the principle F4, which relates to indexing of the metadata in a searchable resource.

- The output of queries consisted of information structured in XML or JSON, which were parsed using Python to extract information.

- Each maturity indicator was encoded as a binary value:
  - “1” if the criterion was satisfied and
  - “0” in the opposite case.

- With the exception of indicators F2 and R1.2.

Results


FAIR maturity indicators
Use cases
Parkinsons_AE
NBIA_GEO
TiO2_eNanoMapper
TiO2_caNanoLab
TiO2_ChEMBL
TiO2_NanoCommons
F1
F2
F3
F4
A1
A11
A12
A2
I1
I2
I3
R1
R11
R12
R13
FAIR guidelines
0.00
0.25
0.50
0.75
1.00
F
A
I
R
automatic
manual
N/A

Conclusion

- In this research, we developed a semi-automated workflow to assess FAIRness and applied it on 6 life sciences resources using maturity indicators.
- We implemented our workflow in a Jupyter notebook to make our analysis open and reproducible.
- We created a FAIR balloon plot to summarize and compare FAIRness compliance.
- Such a workflow could help the developers of the databases to improve their FAIRness.
- Changes to APIs or metadata attributes could affect reproducibility of the results.
- For new datasets, FAIR maturity indicators could be evaluated by changing the search procedure and the values assigned manually.

Acknowledgment

Serena Bonaretti
Transparent MSK Research, Maastricht, The Netherlands (https://tmskr.github.io/)

Laurent Winckers, Jeaphianne van Rijn and Egon Willighagen
Department of Bioinformatics - BiGCaT, NUTRIM, Maastricht University, The Netherlands

Joris Quik, Martine Bakker
National Institute for Public Health and the Environment (RIVM), NL-3720 BA Bilthoven, The Netherlands

Dieter Maier
Biomax Informatics AG, Planegg, Germany

Iseult Lynch
School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, UK