How to implement (scientific) FAIR principles in my work?

Ammar Ammar
ORCID:0000-0002-8399-8990

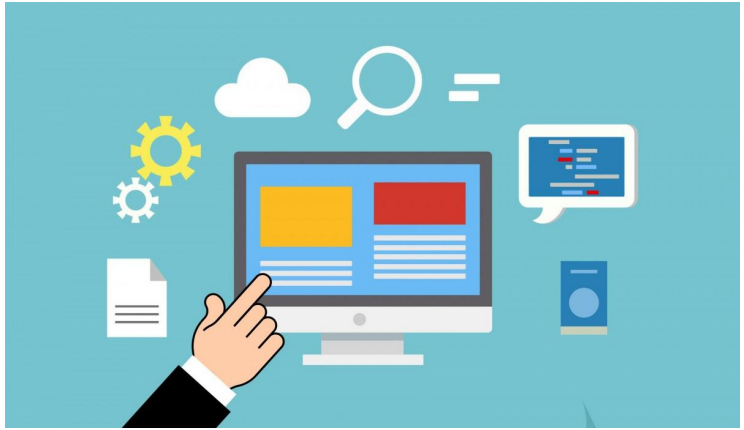
PhD candidate
BiGCAT, NUTRIM, FHML, Maastricht University

16-11-2020

Maastricht University

Why do we need FAIR?

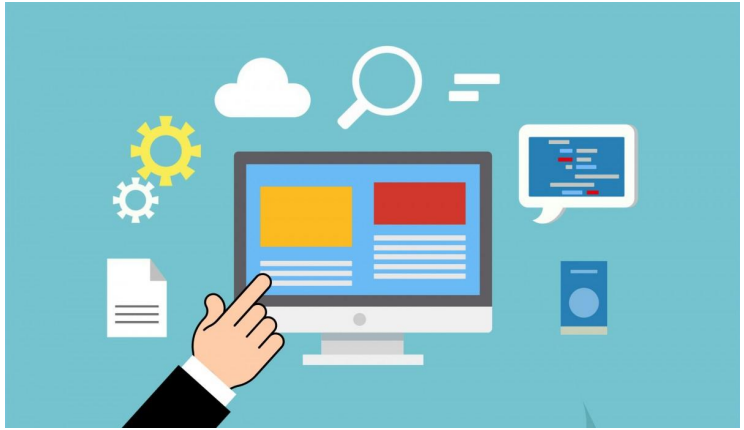
- Data sharing and reuse are beneficial for time efficiency and increased productivity in scientific research.



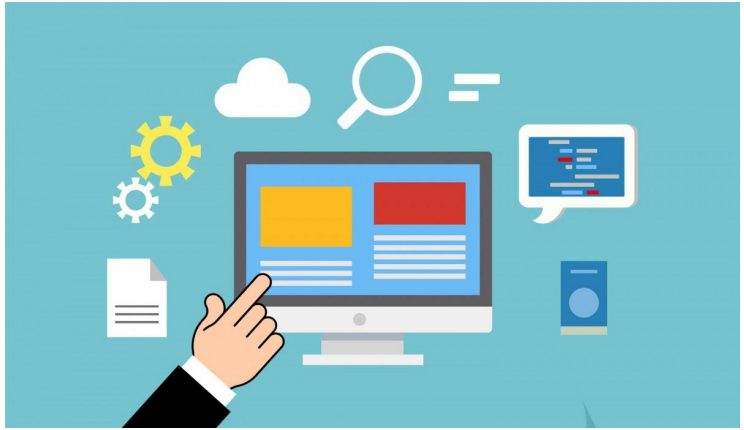
Why do we need FAIR?

- 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.



Why do we need FAIR?

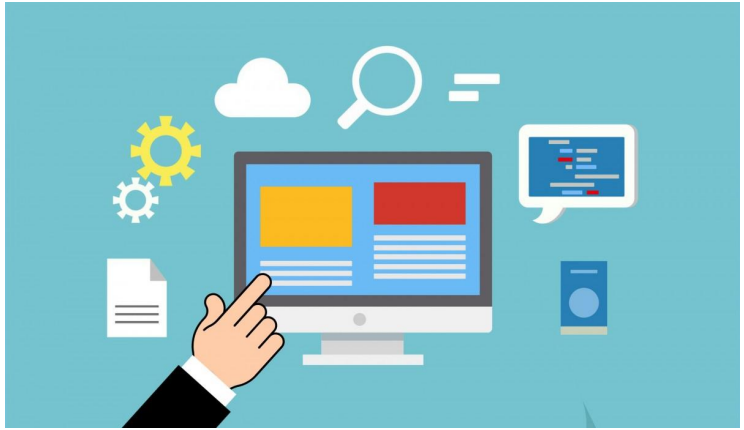


- Data sharing and reuse are beneficial for time efficiency and increased productivity in scientific research.

- Data reuse remains difficult → lack of infrastructures, standards, and policies.

- FAIR (findable, accessible, interoperable, reusable) aims to provide guidance to increase data discovery and reuse.

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) aims to provide guidance to increase data discovery and reuse.
- FAIRness of a dataset can be assessed using maturity indicators.

https://www.ands.org.au/working-with-data/fairdata/training

How to be FAIR in your work ?

1. Data repositories/registries

The research data repository provides additional information on its service.

The research data repository was reviewed by the re3data.org Team.

The research data repository is either certified or supports a repository standard.

The research data repository provides a policy.

The research data repository provides open/restricted/closed access to its data.

The research data repository uses a persistent identifier system to make its provided data persistent, unique and citable.

The terms of use and licenses of the data are provided by the research data repository.

https://commons.wikimedia.org/wiki/File:Journal.pone.0078080.g001.png

1. Data repositories/registries

Data registry
Provides information on repositories for the permanent storage and access of data sets to researchers, funding bodies, publishers and scholarly institutions (e.g. re3data)

https://www.openaire.eu/opendatapilot-repository-guide

1. Data repositories/registries

Data registry

Provides information on repositories for the permanent storage and access of data sets to researchers, funding bodies, publishers and scholarly institutions (e.g. re3data)

Repository Badge for eNanoMapper (re3data)

re3data.org

REGISTRY OF RESEARCH DATA REPOSITORIES

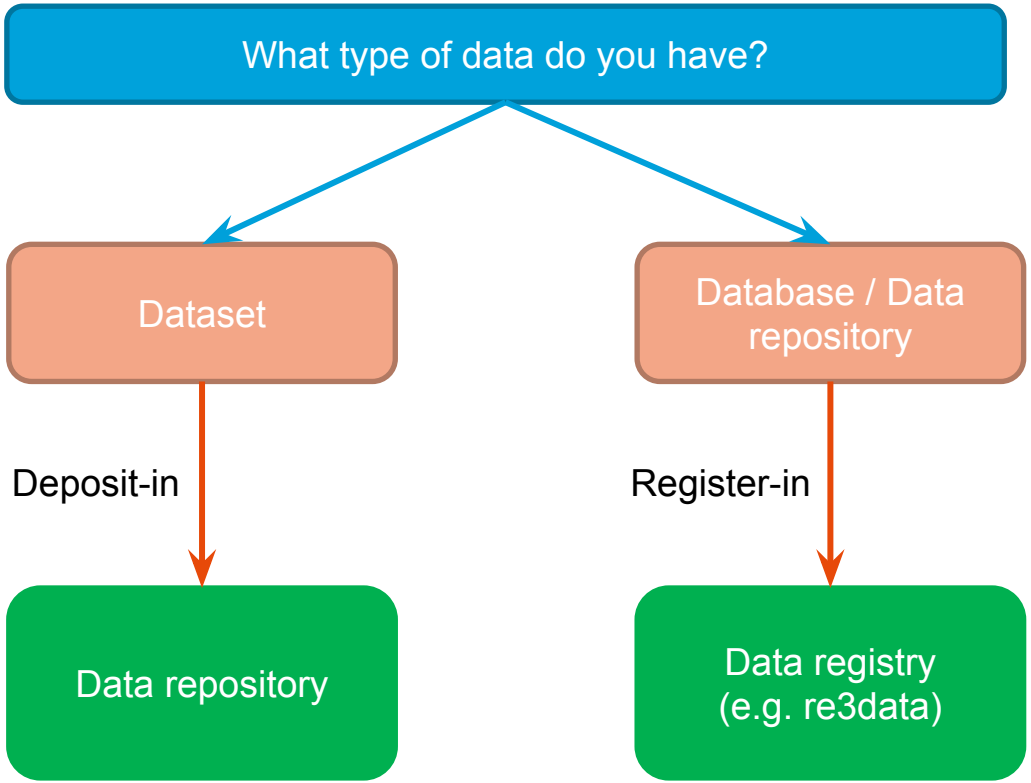
http://doi.org/10.17616/R31NJMKC

data.enanomapper

https://www.re3data.org/resources/badge/100013052

https://www.openaire.eu/opendatapilot-repository-guide

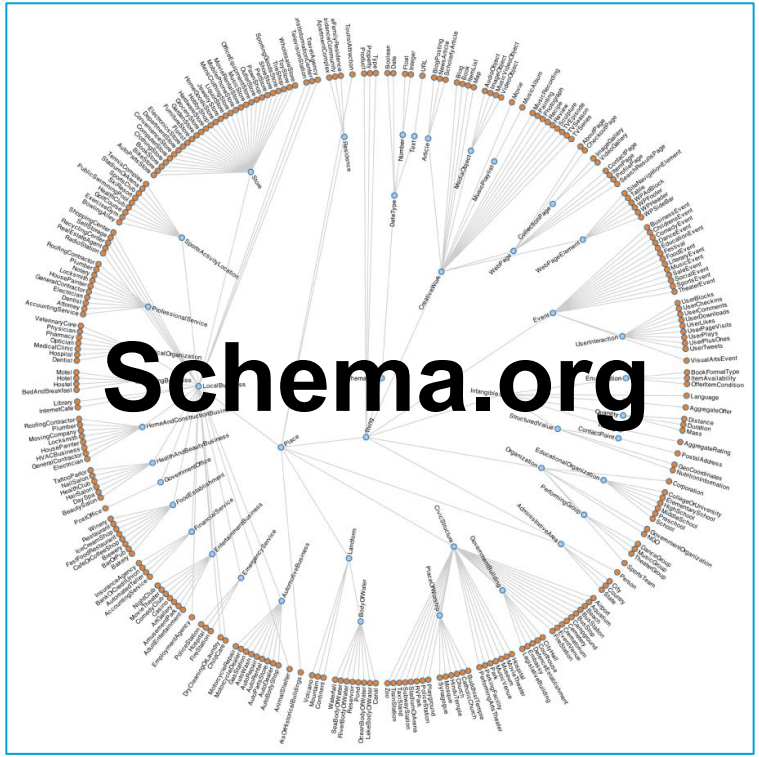
1. Data repositories/registries



2. Metadata/Controlled vocabularies

Bioschemas

- High quality metadata improves data discovery.

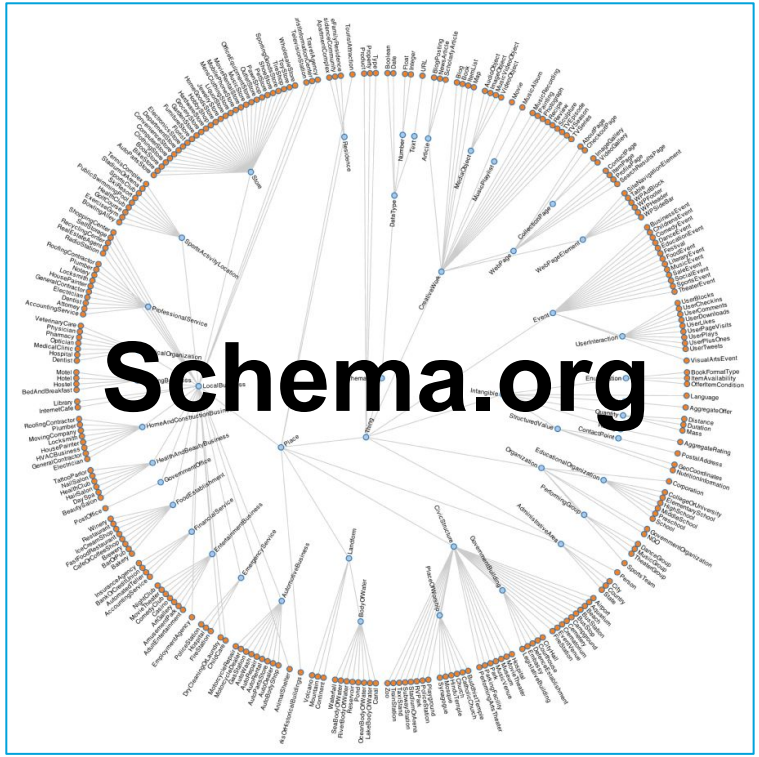


https://commons.wikimedia.org/wiki/File:Schema.org_(5925660995).png

2. Metadata/Controlled vocabularies

Bioschemas

- High quality metadata improves data discovery.
- Using controlled vocabularies -> increase chance to be discovered user searches.
- Using a metadata schema to mark up a dataset can make your data findable to the world.

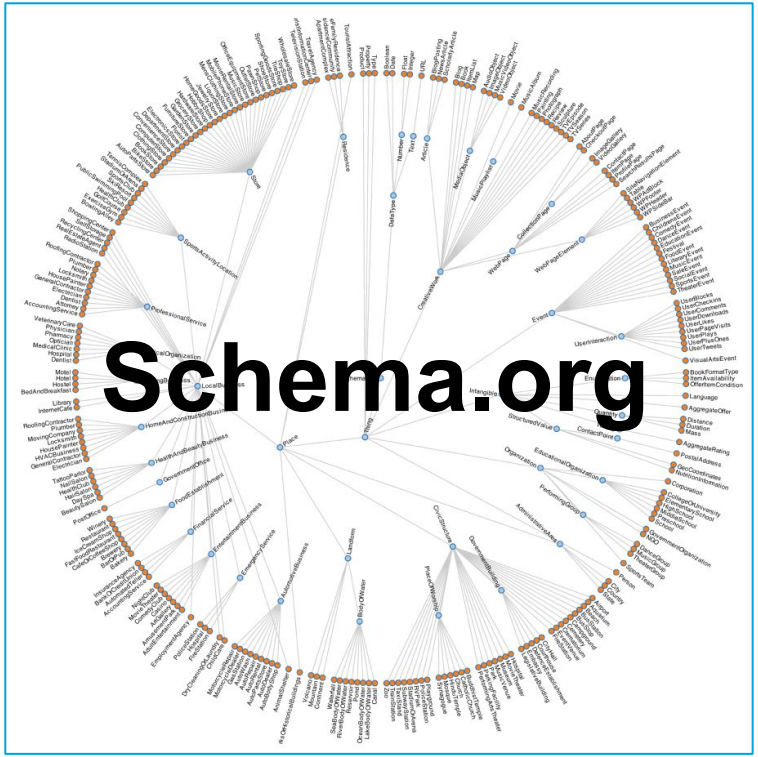


https://commons.wikimedia.org/wiki/File:Schema.org_(5925660995).png

2. Metadata/Controlled vocabularies

Bioschemas

- High quality metadata improves data discovery.
- Using controlled vocabularies -> increase chance to be discovered user searches.
- Using a metadata schema to mark up a dataset can make your data findable to the world.
- Adding markup from **Schema.org** and its extension for the life sciences **Bioschemas.org** to your personal/institute web site -> indexed by **Google Dataset Search**



https://commons.wikimedia.org/wiki/File:Schema.org_(5925660995).png

3. Permanent Identifiers

- Web links can break.
- Tracking down data based on a general description can be extremely challenging.
- **Solution!!** Permanent identifiers.

3. Permanent Identifiers

- Web links can break.
- Tracking down data based on a general description can be extremely challenging.
- **Solution!!** Permanent identifiers.

- Example of permanent identifiers: DOI and ORCID

Example:

https://doi.org/10.5468/ogs.2016.59.1.1

DOI Directory | Prefix | Suffix

3. Permanent Identifiers

- Web links can break.
- Tracking down data based on a general description can be extremely challenging.
- **Solution!!** Permanent identifiers.

- Example of permanent identifiers: DOI and ORCID

Benefits?

- Keeping track of data
- Data does not get lost or misidentified.
- Easier to cite and track the impact of datasets, much like cited journal articles.

Example:

https://doi.org/10.5468/ogs.2016.59.1.1

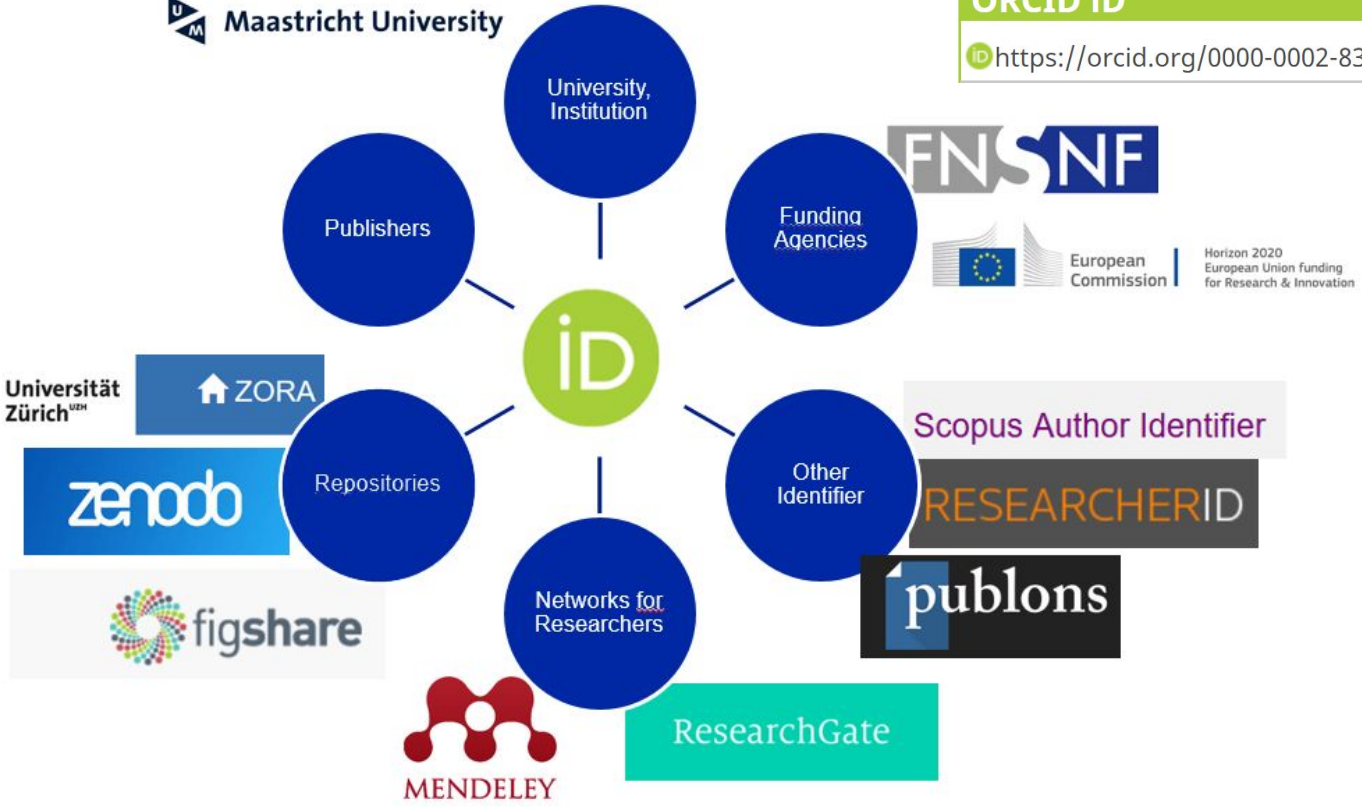
DOI Directory | Prefix | Suffix

3. Permanent Identifiers

Maastricht University

ORCID iD

https://orcid.org/0000-0002-8399-8990



https://www.uzh.ch/blog/hbz/2019/05/29/manage-your-research-output-with-orcid/?lang=en

4. Data formats, structuring and organization

Well-structured and well-organized data:

-> can be **reused** much more easily

-> can be **interoperable**

https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989

4. Data formats, structuring and organization

Well-structured and well-organized data:

-> can be **reused** much more easily

-> can be **interoperable**

- Many life sciences researchers capture their data in spreadsheets.

https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989

4. Data formats, structuring and organization

Well-structured and well-organized data:

-> can be **reused** much more easily

-> can be **interoperable**

- Many life sciences researchers capture their data in spreadsheets.

- Notes:

E	F	G
NA	1.5	2.4
1.3		4.1

https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989

4. Data formats, structuring and organization

Well-structured and well-organized data:

-> can be **reused** much more easily

-> can be **interoperable**

- Many life sciences researchers capture their data in spreadsheets.

- Notes:

E	F	G
NA	1.5	2.4
1.3		4.1

P	Q	R
Core_size/Surface_charge	Core_size	Surface_charge
313.8, 74.2	313.8	74.2

https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989

4. Data formats, structuring and organization

Well-structured and well-organized data:

-> can be **reused** much more easily

-> can be **interoperable**

- Many life sciences researchers capture their data in spreadsheets.

- Notes:

E	F	G
NA	1.5	2.4
1.3		4.1

P	Q	R
Core_size/Surface_charge	Core_size	Surface_charge
313.8, 74.2	313.8	74.2

F	G
2020-11-16	11/16/2020
2020-11-15	11/15/20
2020-11-14	14-Nov-20
yyyy-mm-dd	

https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989

4. Data formats, structuring and organization

Well-structured and well-organized data:

-> can be **reused** much more easily

-> can be **interoperable**

- Many life sciences researchers capture their data in spreadsheets.

- Notes:

E	F	G
NA	1.5	2.4
1.3		4.1

P	Q	R
Core_size/Surface_charge	Core_size	Surface_charge
313.8, 74.2	313.8	74.2

F	G
2020-11-16	11/16/2020
2020-11-15	11/15/20
2020-11-14	14-Nov-20
yyyy-mm-dd	

+120	120
+80	80
-40	40
-60	60

https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989

4. Data formats, structuring and organization

- **Data model + data dictionary**
- **Data dictionary documents the model:**

DATA

last_name	nin	dept_id
Martinez	HH 45 09 73 D	1
Goldstein	SA 75 35 42 B	2
Cornelsen	NE 22 63 82	2
Petculescu	XY 29 87 61 A	1
Stadick	MA 12 89 36 A	15
Scardelis	AT 20 73 18	2
Hunter	HW 12 94 21 C	6
Evans	LX 13 26 39 B	6
Berndt	YA 49 88 11 A	3
Eaton	BE 08 74 68 A	1

DATA DICTIONARY (METADATA)

Column	Data Type	Description
emlployee_id	int	Primary key of a table
first_name	nvarchar(50)	Employee first name
last_name	nvarchar(50)	Employee last name
nin	nvarchar(15)	National Identification Number
position	nvarchar(50)	Current postion title, e.g. Secretary
dept_id	int	Employee deparmtnet. Ref: Departmetns
gender	char(1)	M = Male, F = Female, Null = unknown
employment_start_date	date	Start date of employment in organization.
employment_end_date	date	Employment end date.

Dataedo

https://dataedo.com/kb/data-glossary/what-is-data-dictionary

4. Data formats, structuring and organization

- **Data model + data dictionary**
- **Data dictionary documents the model:**

- A list of all the column names used in the data spreadsheet

DATA

last_name	nin	dept_id
Martinez	HH 45 09 73 D	1
Goldstein	SA 75 35 42 B	2
Cornelsen	NE 22 63 82	2
Petculescu	XY 29 87 61 A	1
Stadick	MA 12 89 36 A	15
Scardelis	AT 20 73 18	2
Hunter	HW 12 94 21 C	6
Evans	LX 13 26 39 B	6
Berndt	YA 49 88 11 A	3
Eaton	BE 08 74 68 A	1

DATA DICTIONARY (METADATA)

Column	Data Type	Description
emlployee_id	int	Primary key of a table
first_name	nvarchar(50)	Employee first name
last_name	nvarchar(50)	Employee last name
nin	nvarchar(15)	National Identification Number
position	nvarchar(50)	Current postion title, e.g. Secretary
dept_id	int	Employee deparmtnet. Ref: Departmetns
gender	char(1)	M = Male, F = Female, Null = unknown
employment_start_date	date	Start date of employment in organization.
employment_end_date	date	Employment end date.

Dataedo

https://dataedo.com/kb/data-glossary/what-is-data-dictionary

4. Data formats, structuring and organization

- **Data model + data dictionary**
- **Data dictionary documents the model:**

- A list of all the column names used in the data spreadsheet
- A description of the purpose and the contents of the columns.

DATA

last_name	nin	dept_id
Martinez	HH 45 09 73 D	1
Goldstein	SA 75 35 42 B	2
Cornelsen	NE 22 63 82	2
Petculescu	XY 29 87 61 A	1
Stadick	MA 12 89 36 A	15
Scardelis	AT 20 73 18	2
Hunter	HW 12 94 21 C	6
Evans	LX 13 26 39 B	6
Berndt	YA 49 88 11 A	3
Eaton	BE 08 74 68 A	1

DATA DICTIONARY (METADATA)

Column	Data Type	Description
emlployee_id	int	Primary key of a table
first_name	nvarchar(50)	Employee first name
last_name	nvarchar(50)	Employee last name
nin	nvarchar(15)	National Identification Number
position	nvarchar(50)	Current postion title, e.g. Secretary
dept_id	int	Employee deparmtnet. Ref: Departmetns
gender	char(1)	M = Male, F = Female, Null = unknown
employment_start_date	date	Start date of employment in organization.
employment_end_date	date	Employment end date.

Dataedo

https://dataedo.com/kb/data-glossary/what-is-data-dictionary

4. Data formats, structuring and organization

- **Data model + data dictionary**
- **Data dictionary documents the model:**

- A list of all the column names used in the data spreadsheet
- A description of the purpose and the contents of the columns.
- Give an indication of the units of measurement.

DATA

last_name	nin	dept_id
Martinez	HH 45 09 73 D	1
Goldstein	SA 75 35 42 B	2
Cornelsen	NE 22 63 82	2
Petculescu	XY 29 87 61 A	1
Stadick	MA 12 89 36 A	15
Scardelis	AT 20 73 18	2
Hunter	HW 12 94 21 C	6
Evans	LX 13 26 39 B	6
Berndt	YA 49 88 11 A	3
Eaton	BE 08 74 68 A	1

DATA DICTIONARY (METADATA)

Column	Data Type	Description
emlployee_id	int	Primary key of a table
first_name	nvarchar(50)	Employee first name
last_name	nvarchar(50)	Employee last name
nin	nvarchar(15)	National Identification Number
position	nvarchar(50)	Current postion title, e.g. Secretary
dept_id	int	Employee deparmtnet. Ref: Departmetns
gender	char(1)	M = Male, F = Female, Null = unknown
employment_start_date	date	Start date of employment in organization.
employment_end_date	date	Employment end date.

Dataedo

https://dataedo.com/kb/data-glossary/what-is-data-dictionary

4. Data formats, structuring and organization

- **Data model + data dictionary**
- **Data dictionary documents the model:**

- A list of all the column names used in the data spreadsheet
- A description of the purpose and the contents of the columns.
- Give an indication of the units of measurement.
- Describe the measures that have been taken to ensure the correctness and the consistency of the data.

DATA

last_name	nin	dept_id
Martinez	HH 45 09 73 D	1
Goldstein	SA 75 35 42 B	2
Cornelsen	NE 22 63 82	2
Petculescu	XY 29 87 61 A	1
Stadick	MA 12 89 36 A	15
Scardelis	AT 20 73 18	2
Hunter	HW 12 94 21 C	6
Evans	LX 13 26 39 B	6
Berndt	YA 49 88 11 A	3
Eaton	BE 08 74 68 A	1

DATA DICTIONARY (METADATA)

Column	Data Type	Description
emlployee_id	int	Primary key of a table
first_name	nvarchar(50)	Employee first name
last_name	nvarchar(50)	Employee last name
nin	nvarchar(15)	National Identification Number
position	nvarchar(50)	Current postion title, e.g. Secretary
dept_id	int	Employee deparmtnet. Ref: Departmetns
gender	char(1)	M = Male, F = Female, Null = unknown
employment_start_date	date	Start date of employment in organization.
employment_end_date	date	Employment end date.

Dataedo

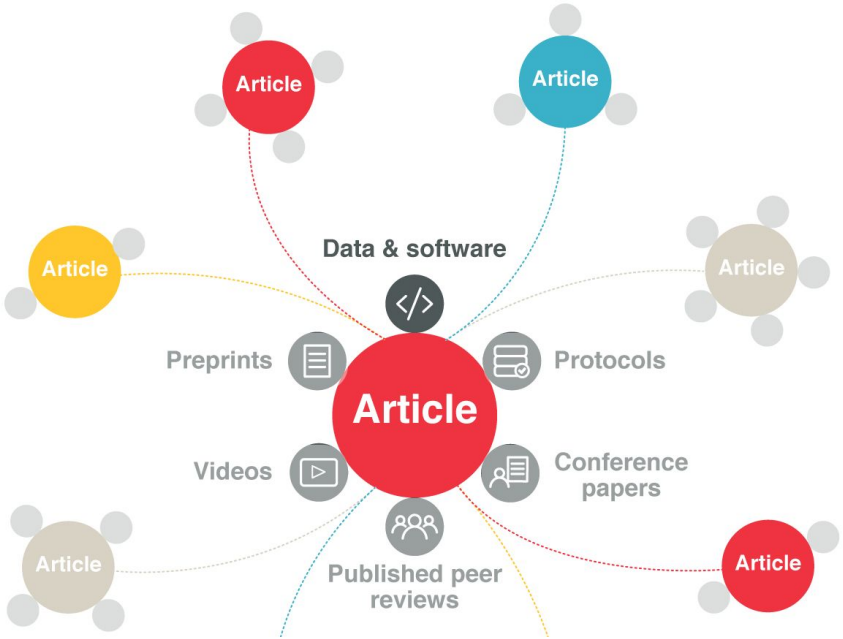
https://dataedo.com/kb/data-glossary/what-is-data-dictionary

5. Licensing/Data citation

Give your data a license

A license describes the conditions under which your data or software is (re)usable

Check https://creativecommons.org/licenses/

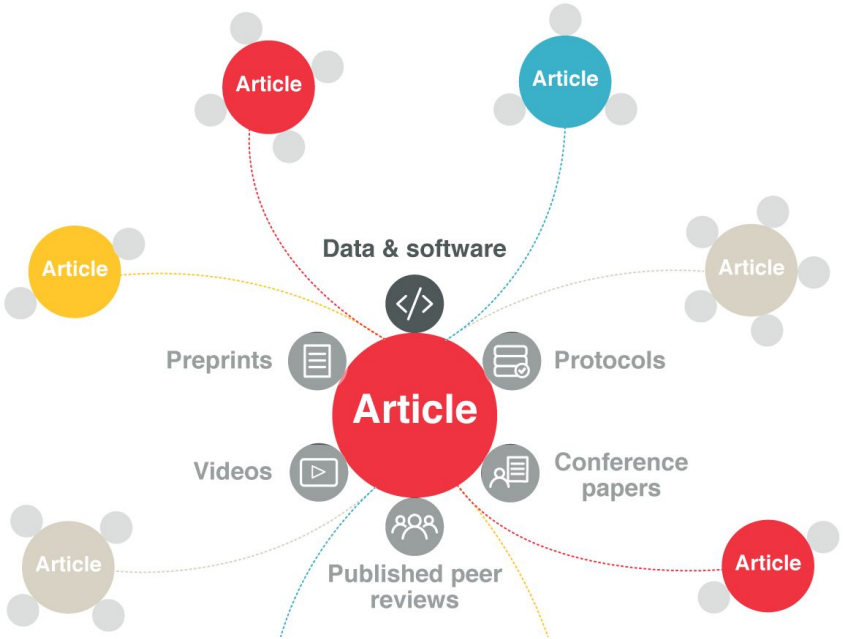


https://www.crossref.org/blog/data-citation-lets-do-this/

5. Licensing/Data citation

Give your data a license
A license describes the conditions under which your data or software is (re)usable
Check https://creativecommons.org/licenses/

State how to cite your data
- A data citation should include: author/creator, date of publication, title of dataset, publisher/organization, and unique identifier.



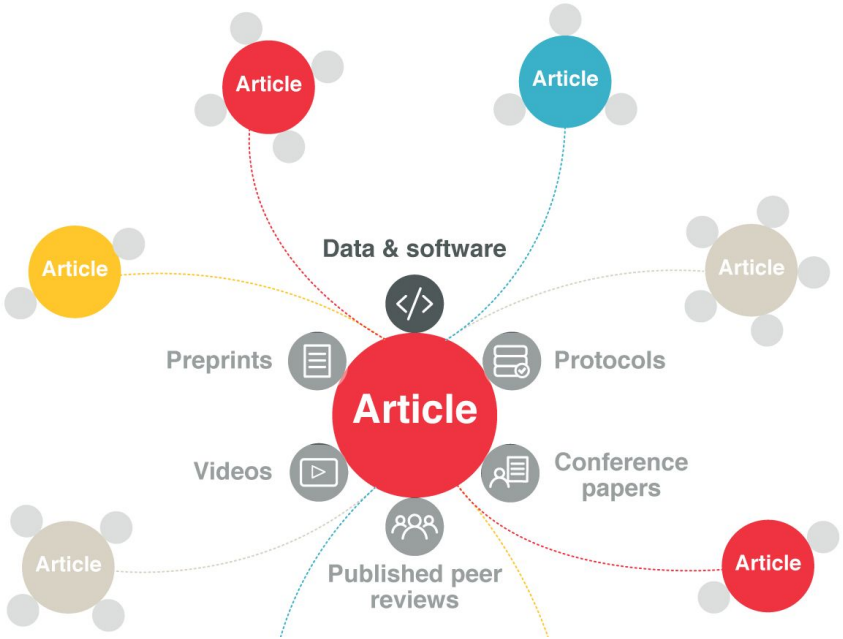
https://www.crossref.org/blog/data-citation-lets-do-this/

5. Licensing/Data citation

Give your data a license
A license describes the conditions under which your data or software is (re)usable
Check https://creativecommons.org/licenses/

State how to cite your data
- A data citation should include: author/creator, date of publication, title of dataset, publisher/organization, and unique identifier.

- Long-term data stewardship is an important factor for keeping data open and accessible for the long term.



https://www.crossref.org/blog/data-citation-lets-do-this/

6. Maturity indicators (assess the FAIRness of data)

- FAIR principles do not specify any technical requirement.

6. Maturity indicators (assess the FAIRness of data)

- FAIR principles do not specify any technical requirement.
- Data reusability in the life sciences domain is hard to quantify.

6. Maturity indicators (assess the FAIRness of data)

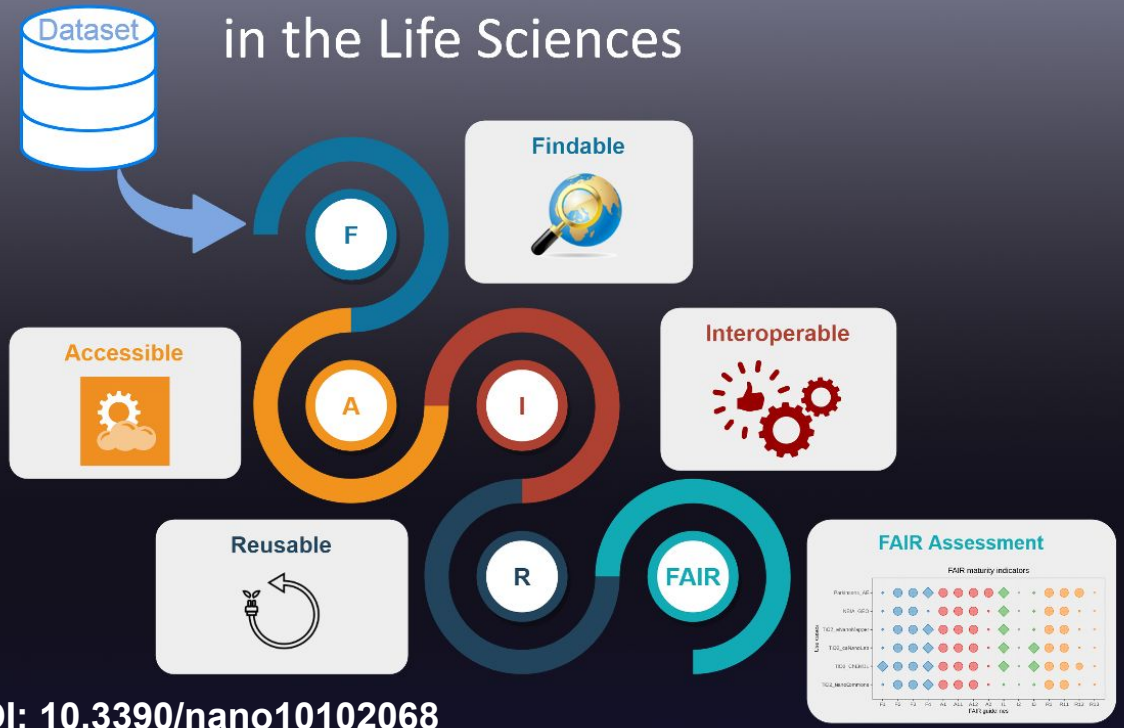
- FAIR principles do not specify any technical requirement.
- Data reusability in the life sciences domain is hard to quantify.
- FAIR assessment is mostly done manually, which makes the process slow and less objective.

6. Maturity indicators (assess the FAIRness of data)

- FAIR principles do not specify any technical requirement.
- 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.

nanomaterials

A Semi-Automated Workflow for FAIR Maturity Indicators in the Life Sciences



DOI: 10.3390/nano10102068

python™

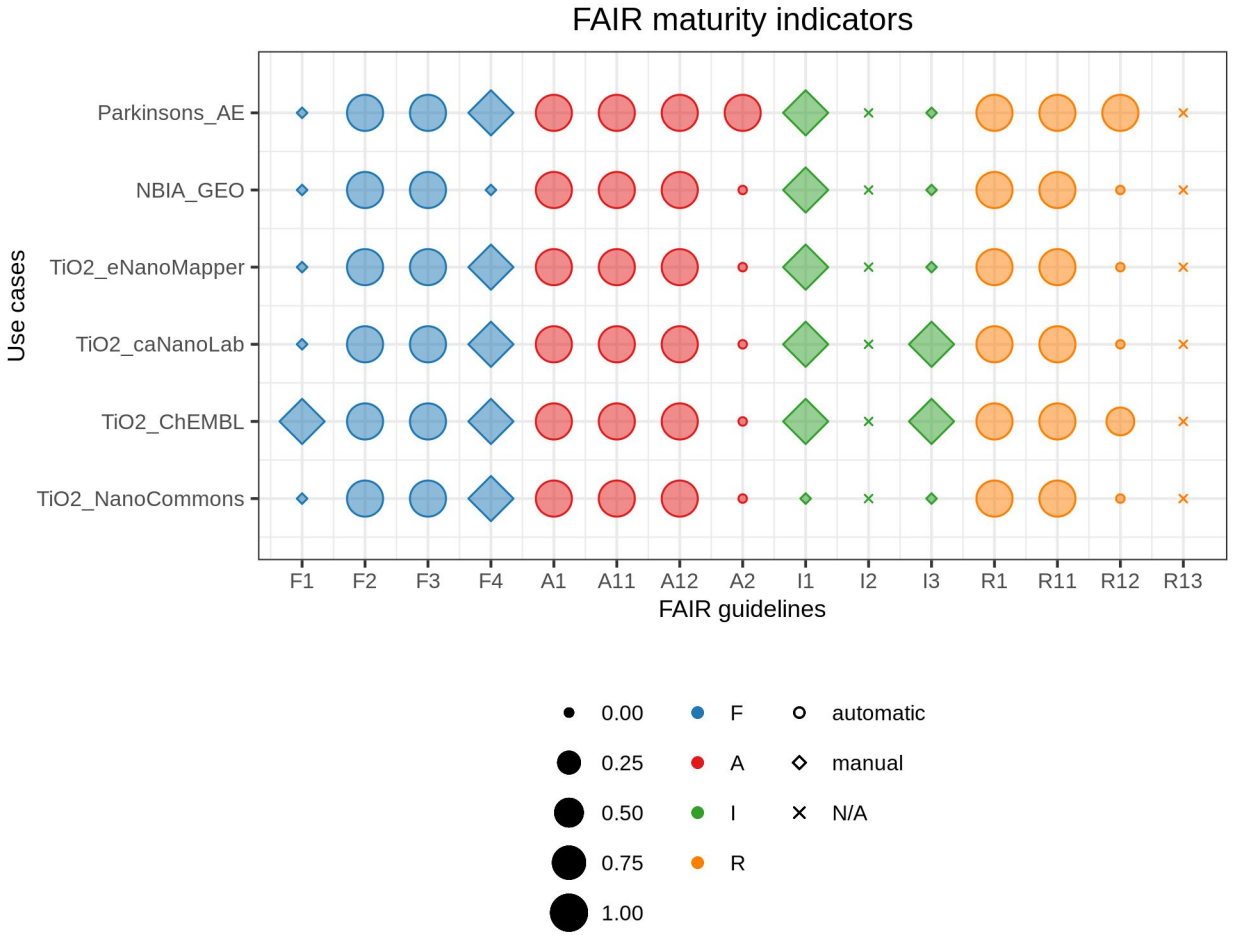
jupyter

binder

GitHub

https://nanocommons.github.io/wgf-fairdata-analysis/

Study Results



Work in progress

- Develop new maturity indicators more specific for nano-QSAR applications, especially related to the R (reusable) and I (interoperable) principles.

Work in progress

- Develop new maturity indicators more specific for nano-QSAR applications, especially related to the R (reusable) and I (interoperable) principles.

- Develop a maturity indicator about standardized formats used (IOM, JRC, ISA-TAB) and minimal reporting standards that should be meet in assessed datasets.

Work in progress

- Develop new maturity indicators more specific for nano-QSAR applications, especially related to the R (reusable) and I (interoperable) principles.

- Develop a maturity indicator about standardized formats used (IOM, JRC, ISA-TAB) and minimal reporting standards that should be meet in assessed datasets.

- Observations from 15 nano-QSAR article review:
  - At least 3 features/descriptors were used in any QSAR study
  - Presence of units is important to build the QSAR model.
  - Frequent physio-chemical features used: surface area, porosity, shape, size distribution, zeta potential.

Conclusion

- Implementing FAIR principles in our daily work is crucial to enable data discovery and reusability.

Conclusion

- Implementing FAIR principles in our daily work is crucial to enable data discovery and reusability.

- Making our data/software/workflows FAIR is as important as our publications.

Conclusion

- Implementing FAIR principles in our daily work is crucial to enable data discovery and reusability.
- Making our data/software/workflows FAIR is as important as our publications.
- There is many options (tools, standards, etc) so pick up what benefits you the most.

Conclusion

- Implementing FAIR principles in our daily work is crucial to enable data discovery and reusability.
- Making our data/software/workflows FAIR is as important as our publications.
- There is many options (tools, standards, etc) so pick up what benefits you the most.
- FAIRness can be measured.

Conclusion

- Implementing FAIR principles in our daily work is crucial to enable data discovery and reusability.
- Making our data/software/workflows FAIR is as important as our publications.
- There is many options (tools, standards, etc) so pick up what benefits you the most.
- FAIRness can be measured.
- We developed a semi-automated workflow to assess FAIRness and applied it on six life sciences resources using maturity indicators. Such a workflow could help the developers of the databases to improve their FAIRness.

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

RISK GONE

NanoSolveIT

Gov4Nano
meeting the needs of nanotechnology

NanoSafety Cluster

Thank you