

FAIR data in research setting under the umbrella of UM open science

Ammar Ammar

 0000-0002-8399-8990

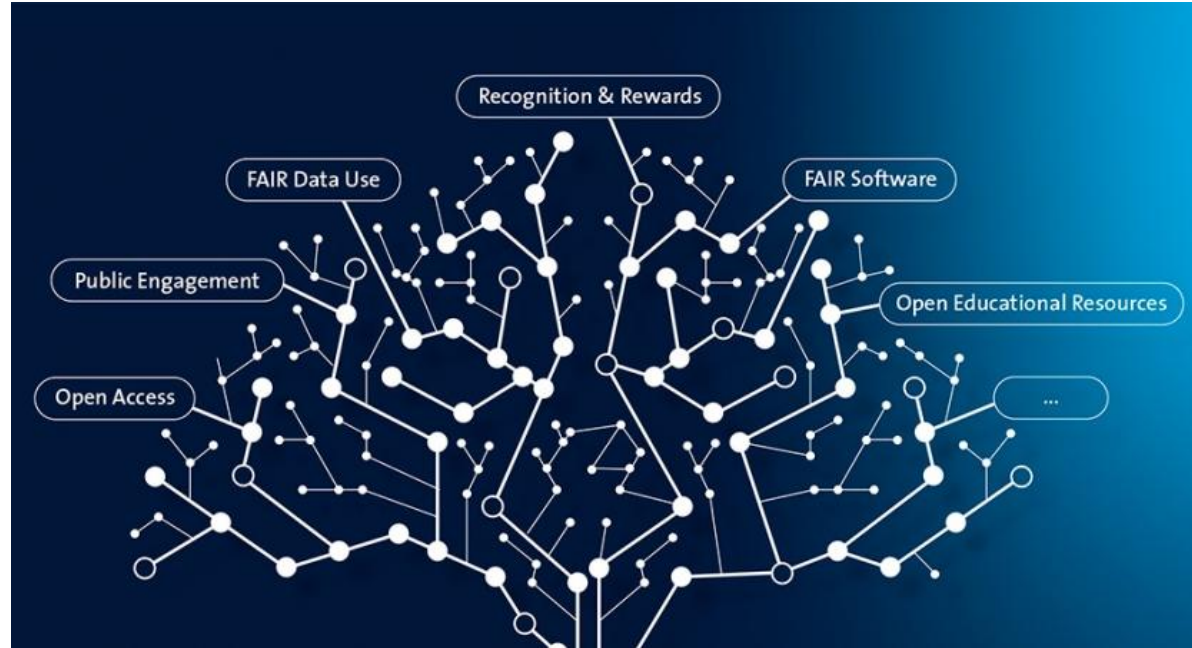
PhD candidate
BiGCAT, NUTRIM, FHML, Maastricht University

03-03-2023



UM open science

Open Science makes research more transparent, controllable, faster, more efficient, reproducible and more sustainable. The idea is that civil society organisations, patient organisations, companies and other organisations can all benefit from easy access to scientific research.

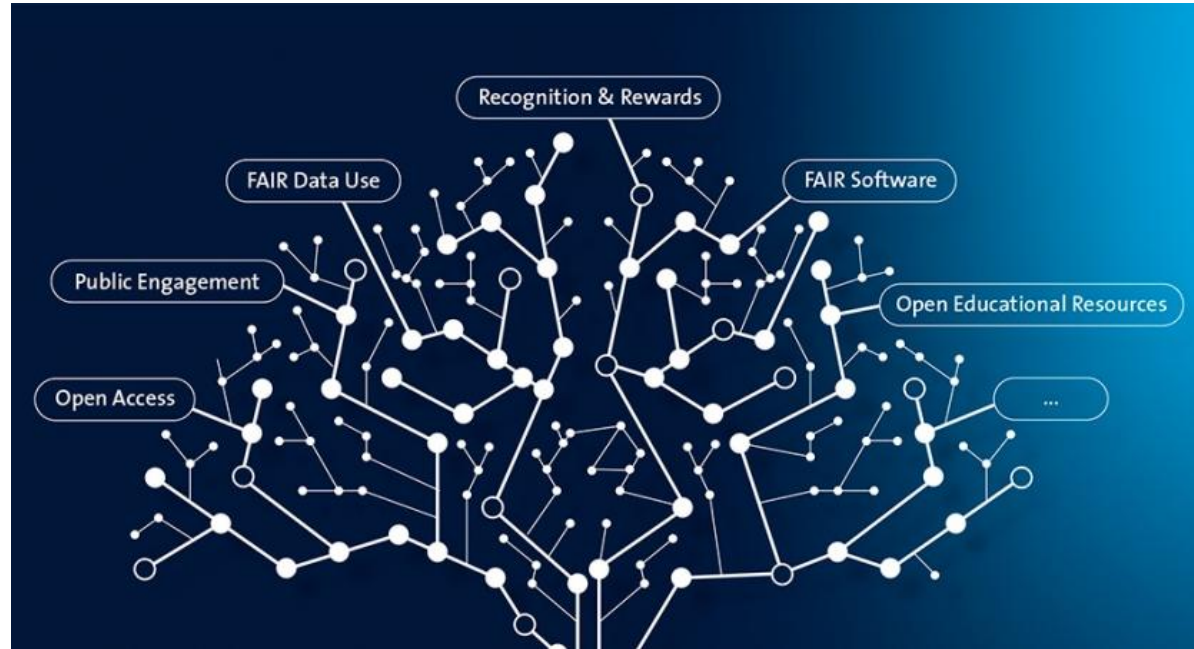


<https://www.maastrichtuniversity.nl/research/open-science>

UM open science

FAIR data use: if possible, research data must be Findable, Accessible, Interoperable and Reusable. UM wants to be fully FAIR by 2023.

Open Access: promoting free online access to scientific information, such as publications and data.



<https://www.maastrichtuniversity.nl/research/open-science>

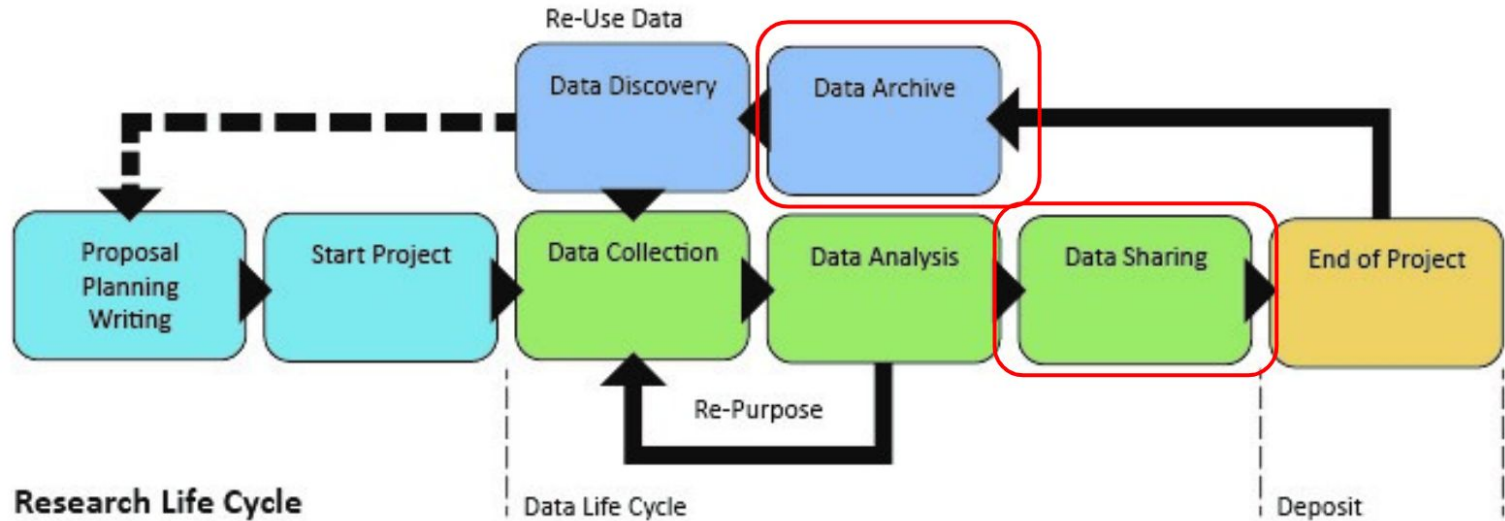
Benefits of FAIR data

For researchers:

- Increase the visibility of your research / maximise research impact
- Satisfy the data-management expectations of funding agencies, journals and peers
- Save time and avoid duplication of effort by reusing an existing dataset



<https://www.andis.org.au/working-with-data/fairdata/training>



Outline

1. Data preparation
 - 1.1. Adopt community standards and best practices
 - 1.2. Data (pseudo)anonymization
 - 1.3. Data formats, structuring and organization
2. Storage and backup
3. Data sharing
4. Data Publishing & Archiving
 - 4.1. Deposit in a repository / register in registry
 - 4.2. Use persistent identifiers
 - 4.3. Use licenses

1.1. Adopt community standards and best practices in data processing and reporting

[Front Neuroinform.](#) 2019; 13: 61.

Published online 2019 Sep 4. doi: [10.3389/fninf.2019.00061](https://doi.org/10.3389/fninf.2019.00061)

PMCID: PMC6738271

PMID: [31551745](https://pubmed.ncbi.nlm.nih.gov/31551745/)

Recommendations for Processing Head CT Data

[John Muschelli](#)*

[Int Urogynecol J.](#) 2021; 32(6): 1387–1390.

Published online 2020 Oct 28. doi: [10.1007/s00192-020-04575-z](https://doi.org/10.1007/s00192-020-04575-z)

PMID: [33112967](https://pubmed.ncbi.nlm.nih.gov/33112967/)

Minimum standards for reporting outcomes of surgery in urogynaecology

[Philip Tooze-Hobson](#),¹ [Fiona Bach](#),² [J. Oliver Daly](#),^{3,4} and [Niels Klarskov](#)^{5,6}

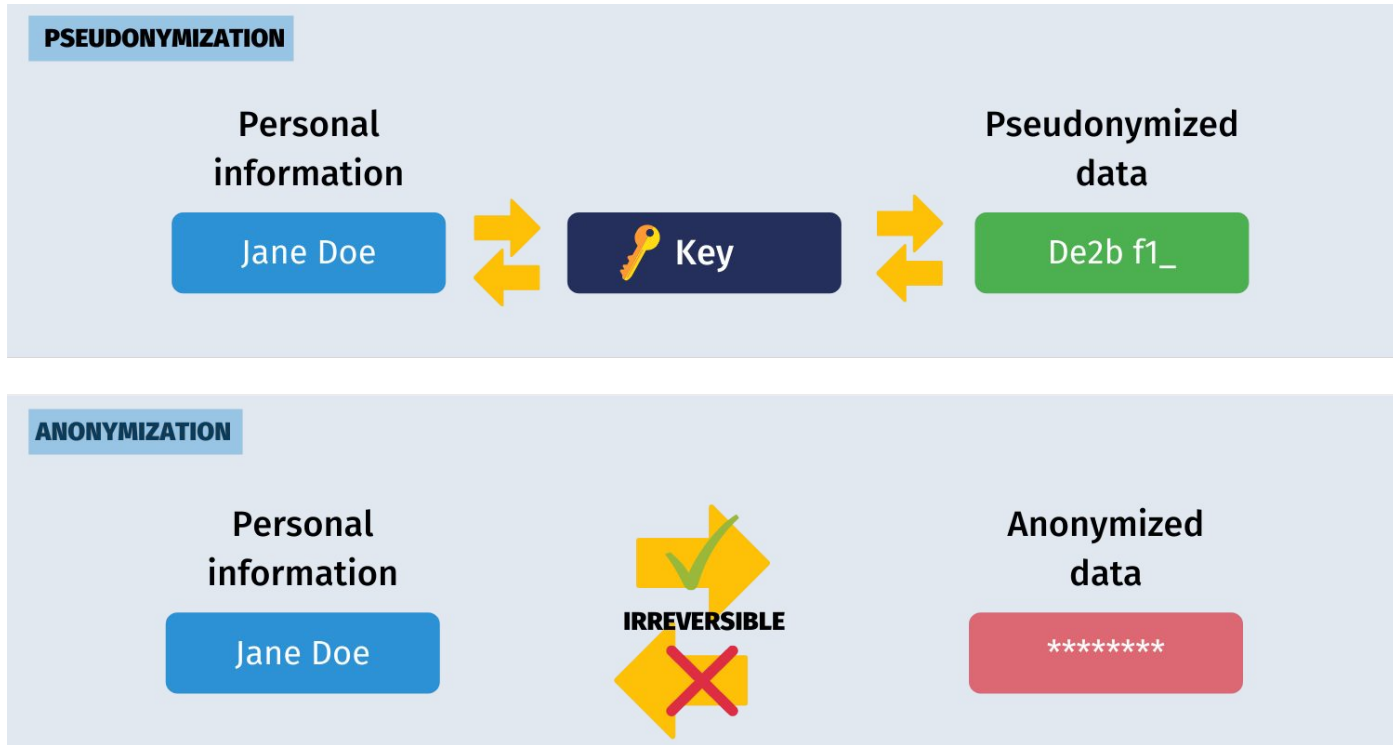
Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools

Oliver Diaz^{a,*}, Kaisar Kushibar^a, Richard Osuala^a, Akis Linardos^a, Lidia Garrucho^a, Laura Igual^a, Petia Radeva^a, Fred Prior^b, Polyxeni Gkontra^a, Karim Lekadir^a

^a Faculty of Mathematics and Computer Science, University of Barcelona, Barcelona, Spain

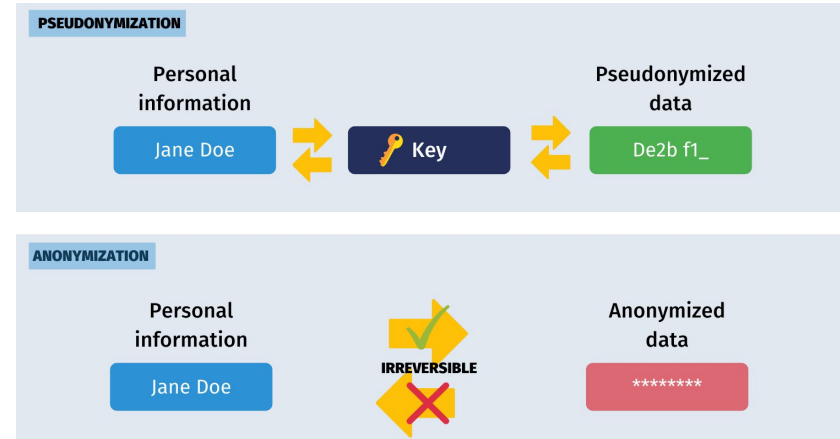
^b Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Arkansas, USA

1.2. Data (pseudo)anonymization



1.2. Data (pseudo)anonymization

- Clinical information (mostly tabular)
 - excluding columns (e.g. BSN)
 - numeric rounding (e.g. salary)
 - adding noise (e.g. birth date)
 - masking (e.g. email: a***b@ex.ample)
- Imaging data
 - scrub patient information from images
 - remove DICOM metadata



1.2. Data (pseudo)anonymization

Example



[Open Data Anonymizer](#)

	name	age	birthdate	salary	web	email	ssn
0	Bruce	33	1915-04-17	59234.32	http://www.alandrosenburgcpapc.co.uk	josefrazier@owen.com	343554334
1	Tony	48	1970-05-29	49324.53	http://www.capgeminiamerica.co.uk	eryan@lewis.com	656564664

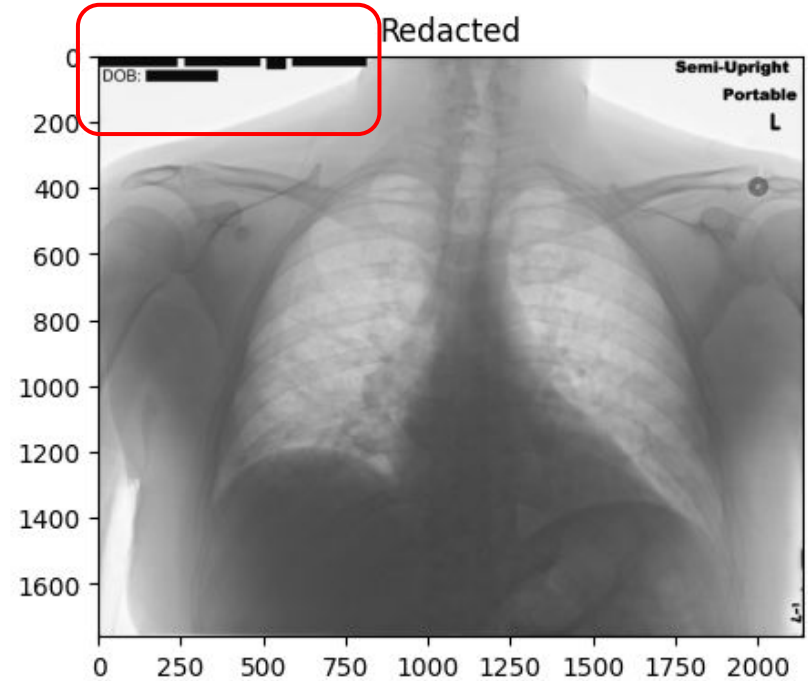
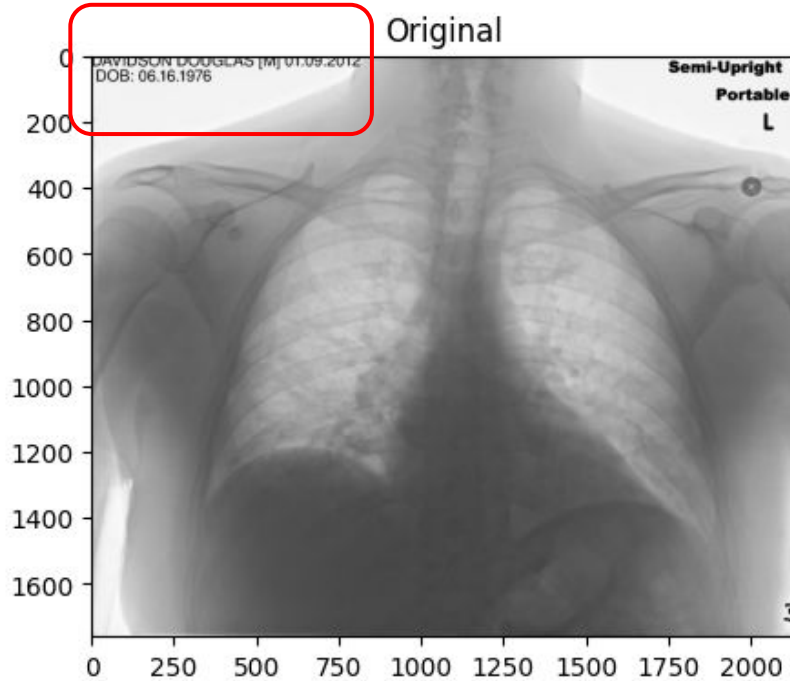
```
# Calling the generic function
>>> anonym = dfAnonymizer(df)
>>> anonym.anonymize(inplace = False) # changes will be returned, not applied
```

	name	age	birthdate	age	web	email	ssn
0	Stephanie Patel	30	1915-05-10	60000.0	5968b7880f	pjordan@example.com	391-77-9210
1	Daniel Matthews	50	1971-01-21	50000.0	2ae31d40d4	tparks@example.org	872-80-9114

1.2. Data (pseudo)anonymization



Microsoft Presidio

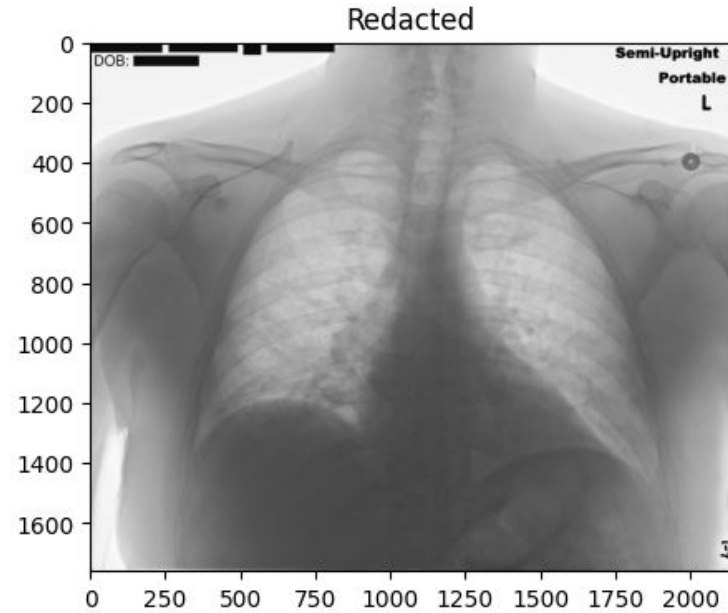
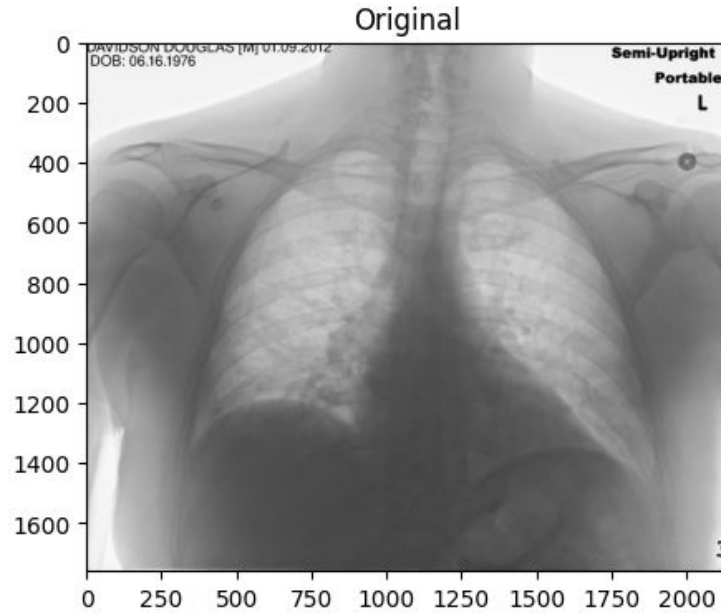


1.2. Data (pseudo)anonymization



Microsoft Presidio

Note: Performance is best when the burnt-in text is also present within the DICOM metadata. It is recommended to remove the metadata after image redaction.



1.2. Data (pseudo)anonymization

DicomAnonymizer

Group	Action	Action definition
D_TAGS	replace	Replace with a non-zero length value that may be a dummy value and consistent with the VR**
Z_TAGS	empty	Replace with a zero length value, or a non-zero length value that may be a dummy value and consistent with the VR**
X_TAGS	delete	Completely remove the tag
U_TAGS	replace_UID	Replace all UID's random ones. Same UID will have the same replaced value
Z_D_TAGS	empty_or_replace	Replace with a non-zero length value that may be a dummy value and consistent with the VR**
X_Z_TAGS	delete_or_empty	Replace with a zero length value, or a non-zero length value that may be a dummy value and consistent with the VR**
X_D_TAGS	delete_or_replace	Replace with a non-zero length value that may be a dummy value and consistent with the VR**
X_Z_D_TAGS	delete_or_empty_or_replace	Replace with a non-zero length value that may be a dummy value and consistent with the VR**
X_Z_U_STAR_TAGS	delete_or_empty_or_replace_UID	If it's a UID, then all numbers are randomly replaced. Else, replace with a zero length value, or a non-zero length value that may be a dummy value and consistent with the VR**
ALL_TAGS		Contains all previous defined tags

1.2. Data (pseudo)anonymization

10 tools analyzed

Only one tool was able to de-identify all required elements with the default setting (RSNA MIRC Clinical Trials Processor)

Springer Open Choice

[Eur Radiol.](#) 2015; 25(12): 3685–3695.

Published online 2015 Jun 3. doi: [10.1007/s00330-015-3794-0](https://doi.org/10.1007/s00330-015-3794-0)

PMCID: PMC4636522

PMID: [26037716](https://pubmed.ncbi.nlm.nih.gov/26037716/)

Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy

[K. Y. E. Aryanto](#),[✉] [M. Oudkerk](#), and [P. M. A. van Ooijen](#)

► [Author information](#) ► [Article notes](#) ► [Copyright and License information](#) [Disclaimer](#)

Abstract

[Go to:](#) ►

Purpose

To compare non-commercial DICOM toolkits for their de-identification ability in removing a patient's personal health information (PHI) from a DICOM header.

1.3. Data formats, structuring and organization

Well-structured and well-organized data:

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

- Many 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

- **XLS is a proprietary file format**
- **XLSX is an open file format**

A list of preferred formats can be found on [DANS](#) and [4TU.ResearchData](#)



1.3. 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
Comelsen	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
Bemdt	YA 49 88 11 A	3
Eaton	BE 08 74 68 A	1

DATA DICTIONARY (METADATA)		
Column	Data Type	Description
employee_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 position title, e.g. Secretary
dept_id	int	Employee department. Ref: Departments
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.



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

1.3. Data formats, structuring and organization

- Define a folder structure in advance
- Define logical categories
- Use a naming convention and document this in a README file
- Keep file names clear and short
- Avoid the use of spaces, dots and special characters in file names
- File names must be consistent, meaningful and easy to find
- Store raw data separately, leave it untouched and use a working copy
- Separate data in progress from completed data
- Avoid ambiguous filenames such as Final_1, Final_2
- Use for example YYYYMMDD for the notation of dates in file names and use this notation consistently
- Use major and minor versions like:
 - Major versions: v01, v02, v03
 - Minor versions: v01_01, v02_02, v03_03

Outline

1. Data preparation
 - 1.1. Adopt community standards and best practices
 - 1.2. Data (pseudo)anonymization
 - 1.3. Data formats, structuring and organization
2. **Storage and backup**
3. Data sharing
4. Data Publishing & Archiving
 - 4.1. Deposit in a repository / register in registry
 - 4.2. Use persistent identifiers
 - 4.3. Use licenses

2. Storage and Backup

Overview of the storage solutions at UM

	Suitable for sensitive data	Sharing / Collaborating possible with:			off line available	backup	synchronous collaboration	Size in GB	Costs	Remarks	
		UM-employees	UM-students	Non-UM							
Storage solutions											
Local storage (PC / laptop)											
	!	✗	✗	✗	✓	✗	✗			Files are not backed up. It is strongly discouraged to use local drives for data storage.	
Managed File Services (MFS)	I-drive (Personal)	✓	✗	✗	✗	✓	✗	10 (students 2)	Free	In case you leave our institution your account including the data on I-drive will be deleted. Make sure your research data are stored and available at your faculty. Use e.g. P-drive or DataverseNL. Storing sensitive data on J:\, L:\, N:\ or P:\ is in itself perfectly save. However, you must be aware of persons or groups that have access to a specific drive.	
	J-drive (Organisation)	✓	✓	✗	✗	✓	✗	unlimited	Paid*		
	L-drive (Projects)	✓	✓	✗	✗	✗	✓	✗	unlimited		Paid*
	N-drive (Education)	✓	✓	✓	✗	✗	✓	✗	unlimited		Paid*
	P-drive (Research)	✓	✓	✓	✗	✗	✓	✗	unlimited		Paid*
	R-drive (Research Data Management)	✓	✓	✓	✗	✗	✓	✗	100 * number of researchers / faculty free		
MS Teams	✗	✓	✗	✗	✗	✓	✓	15 GB / Team	Free	MS Teams offers (video-)conferencing. It also offers some storage. But its characteristics are very different from what, for example, MFS offers. We strongly advise against using MS Teams for storing data. Extension of the default storage size (15 GB / Team) can be requested via Servicedesk-ICT5	
SURFdrive	!	✓	✓	✓	!	✓	✓	500	Free	Please note that even though the SURFdrive platform is secure, activating synchronization to certain devices may introduce a risk to the security of your data. In case you leave our institution your account including the data on Surfdrivewill be deleted. Make sure your research data are stored and available at your faculty. Use e.g. P-drive or DataverseNL. Sharing and full collaboration is possible with all NL higher education institutes using SURFdrive and other institutes that have linked to the SURFdrive infrastructure. Collaboration with others is possible with restricted functionality.	
Atlassian (Jira / Confluence)	under construction										
Sending (big) files											
SURFfilesender	!	✓	✓	✓			✗	500	Free	NB: encryption limited to 2GB	

✓	Yes
!	Only in combination with additional measures (e.g. encryption; Data Management Plan)
✗	No

Outline

1. Data preparation
 - 1.1. Adopt community standards and best practices
 - 1.2. Data (pseudo)anonymization
 - 1.3. Data formats, structuring and organization
2. Storage and backup
3. **Data sharing**
4. Data Publishing & Archiving
 - 4.1. Deposit in a repository / register in registry
 - 4.2. Use persistent identifiers
 - 4.3. Use licenses

3. Data Sharing

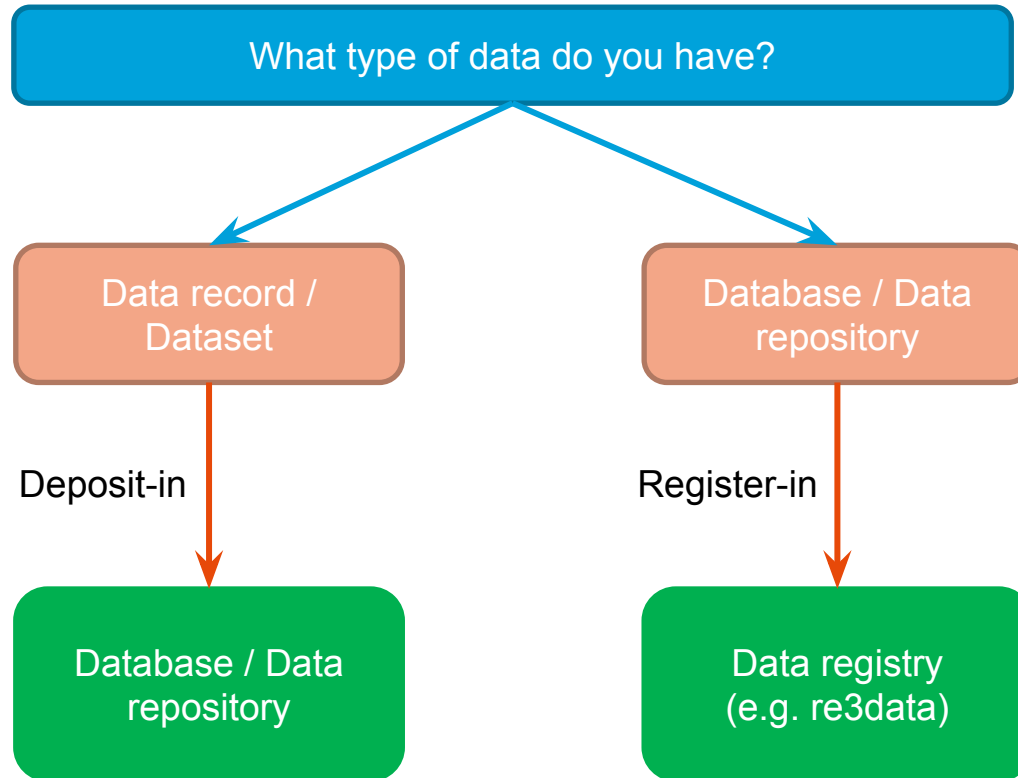
- UM offers a range of solutions for secure collaboration, such as SURFdrive, SURFfilesender and Virtual Research Environments (VREs). To start with SURFdrive or SURFfilesender, check the SURFdrive manual (PDF) and the SURFfilesender manual (PDF).
- Do not use ('free') online storage alternatives such as Dropbox, Google Drive, Box, Hotmail, OneDrive, WeTransfer, Evernote and many others. It is unclear how safe your data is when you store it there. There are even services that require you to transfer intellectual rights to the provider. UM has legal duty to protect (especially) sensitive data and intellectual property should never be transferred to a third party.

<https://library.maastrichtuniversity.nl/research/rdm/guide/collecting-processing-analysing-data/>

Outline

1. Data preparation
 - 1.1. Adopt community standards and best practices
 - 1.2. Data (pseudo)anonymization
 - 1.3. Data formats, structuring and organization
2. Storage and backup
3. Data sharing
4. Data Publishing & Archiving ([Link](#))
 - 4.1. Deposit in a repository / register in registry
 - 4.2. Use persistent identifiers
 - 4.3. Use licenses

4.1. Data archiving



4.1. Data archiving

Database / data repository Examples

- Proteins: UniProt
- Chemical compounds: ChEMBL & PubChem
- Biological Pathways: Wikipathways & KEGG
- Genes: Ensembl
- Omics (Microarray): GEO
- Imaging Data: [QIDW](#), [TCIA](#), [OpenNeuro](#), [others](#)

Other repositories:

- For datasets/document: Zenodo, Figshare
- For code: GitHub, SourceForge


More examples:

<https://www.nature.com/sdata/policies/repositories>



4.1. Data archiving

- Maastricht University Library supports DataverseNL as the midterm storage facility for our institution.
- DataverseNL offers storage up to the prescribed ten years after the last publication based on the data or the completion of the study. This is in accordance with UM's Code of Conduct for RDM. Depending on the discipline the retention period may even be fifteen years and longer.



The image shows a screenshot of the DataverseNL website. At the top left is the DataverseNL logo, which consists of three blue circles of varying sizes connected by lines, with the text 'Data' above 'verseNL' in a blue sans-serif font. Below the logo is a search bar with the placeholder text 'Search for datasets in DataverseNL' and a blue 'Search' button. Below the search bar are two blue buttons: 'About DataverseNL' and 'Browse Dataverses'. The background of the screenshot is a light blue network diagram with nodes and connecting lines.

Online storage, sharing and publishing of research data

4.2. Use persistent identifiers

Example:

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

DOI Directory Prefix Suffix

- Web links can break.
- Tracking down data based on a general description can be extremely challenging.
- **Solution!!** persistent identifiers.
- Example of persistent identifiers: DOI and ORCID



4.2. Use persistent identifiers

For Example: Register data in a 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 & fairsharing)

Repository Badge for eNanoMapper (re3data)



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



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

4.2. Use persistent identifiers

re3data

<https://www.re3data.org/>



Fairsharing

<https://fairsharing.org/>

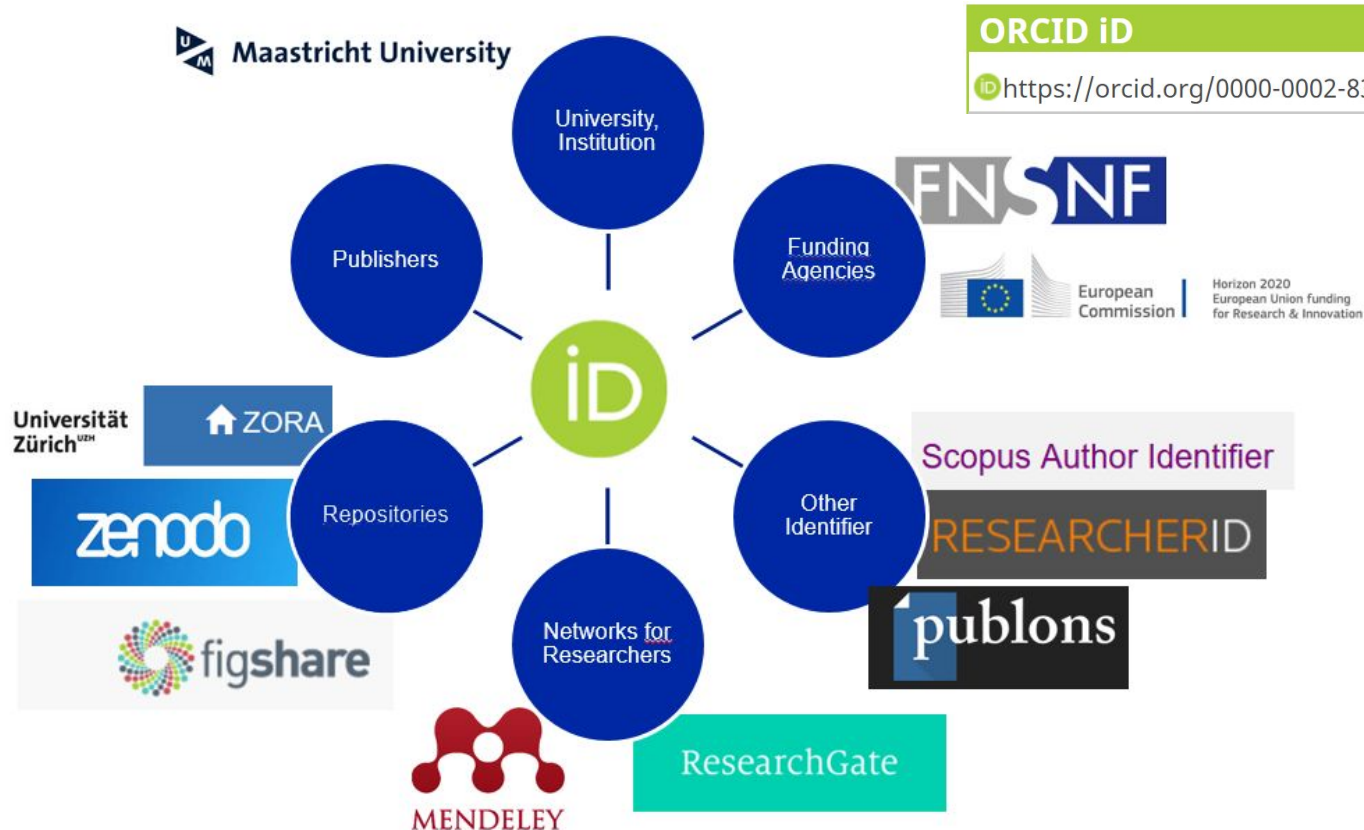


Zenodo

<https://zenodo.org/>



4.2. Use persistent identifiers



ORCID iD
id <https://orcid.org/0000-0002-8399-8990>



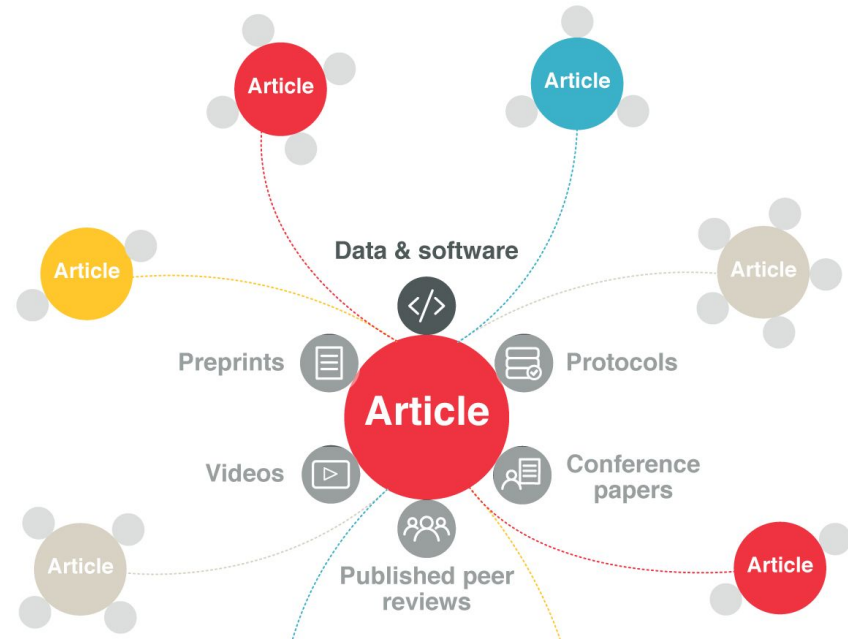
4.3. Use licenses

Give your data a license

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

Choosing an open license

- **General**
<https://choosealicense.com/>
- **Creative Commons licenses**
<https://creativecommons.org/choose>
- **GNU licenses**
<https://www.gnu.org/licenses/license-recommendations.en.html>

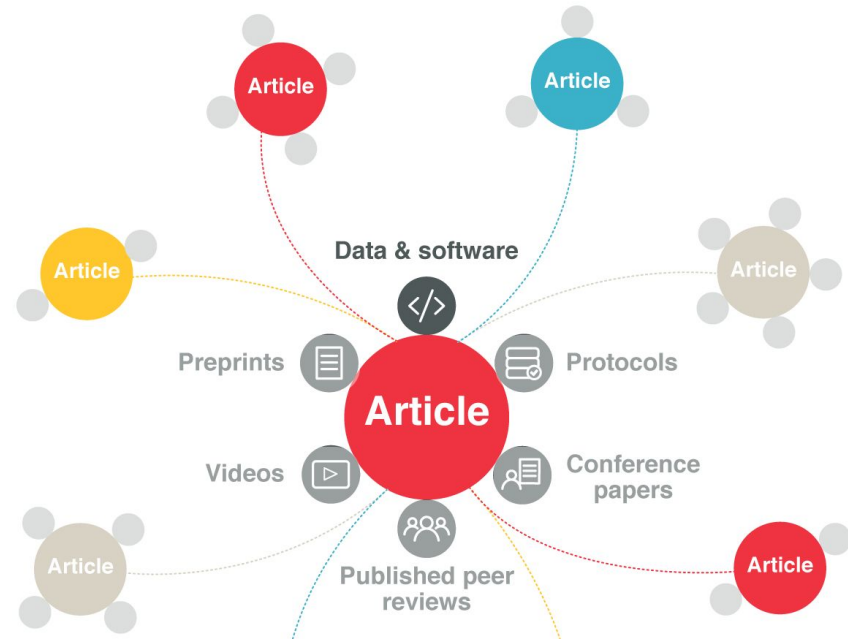


<https://www.crossref.org/blog/data-citation-lets-do-this/>

4.3. Use 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/>

More resources

More on FAIR

- <https://www.go-fair.org/fair-principles/>
- <https://www.openaire.eu/how-to-make-your-data-fair>
- <https://zenodo.org/record/6381648#.ZAC8-dLMJkg>
- <https://zenodo.org/record/7334235#.ZAC9N9LMJkg>

Open Access guide

<https://library.maastrichtuniversity.nl/research/sharing-output/open-access-guide/>

Select the right journal for your paper

<https://library.maastrichtuniversity.nl/research/sharing-output/open-access-guide/select-the-right-journal/>

Five recommendations for FAIR software

<https://fair-software.nl/>

UM RDM consultancy

<https://www.maastrichtuniversity.nl/about-um/other-offices/memic/products-and-services/datamanagement-consultancy>

UM RDM practice

<https://www.maastrichtuniversity.nl/about-um/other-offices/memic/products-and-services/custom-software-and-apps>

Thank you