DATA PUBLICATION 101

UNIVERSITEIT
GENT

OUTLINE

— What is data publishing?

— Why data sharing?

— Data sharing policy

— Data sharing using repositories

    — What is a data repository?

    — Examples

    — How repositories lead to FAIR and/or open data (metadata, persistent identifiers, standards)

    — How to select a data repository?

    — How to use a data repository

— Data re-use, data citation, licenses

— Biblio data register (refer to demos)

UNIVERSITEIT
GENT

PUBLISHING DATA

— Sharing data from

→ <mark>completed</mark> (parts of) research projects

→ <mark>outside</mark> your project or research team.

— Not

— exchanging data with collaborators while your

research is active.

— Share data with your promotor

RATIONALE FOR DEVELOPING FAIR PRINCIPLES

— Urgent need to improve the infrastructure supporting the reuse of research data

— Specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals

— The FAIR principles are a guideline to enhance the data reusability

https://book.fosteropenscience.eu/

UNIVERSITEIT GENT
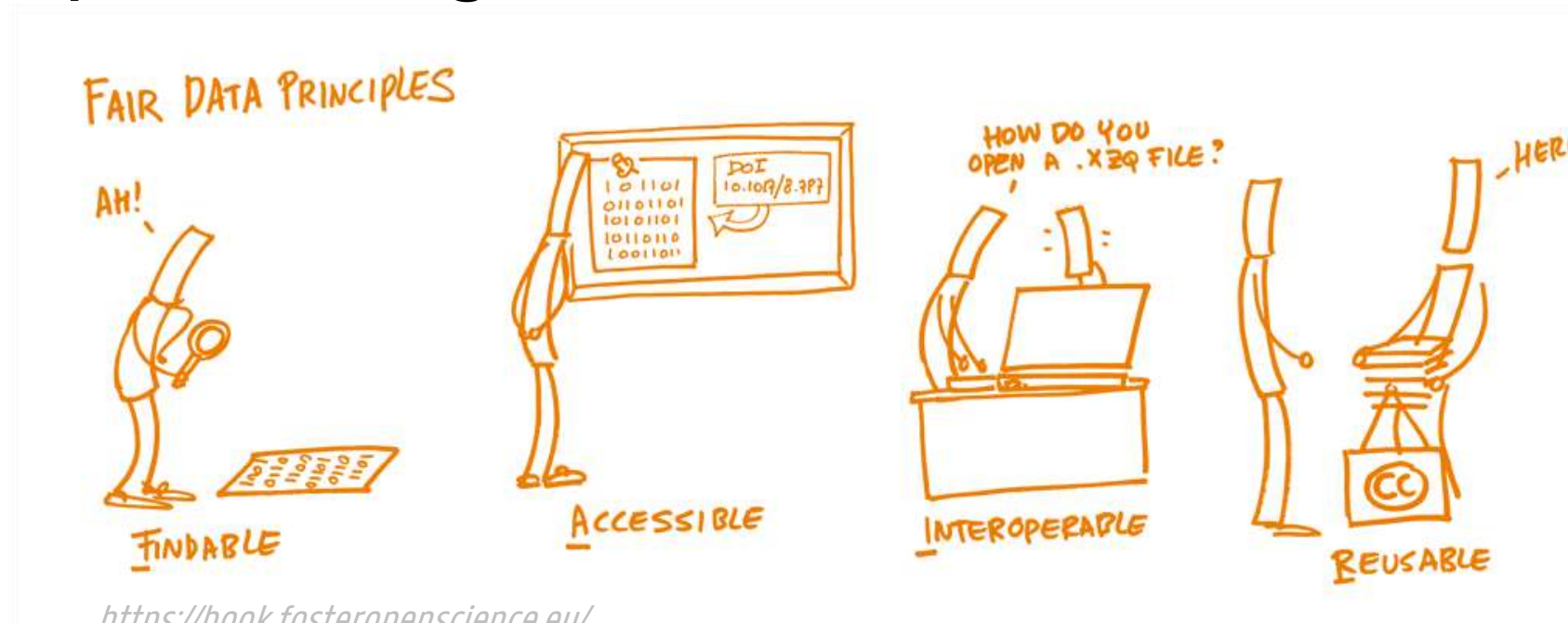
WHY SHARE DATA?

THE BENEFITS

FOR RESEARCHERS:

- More visibility and citations.
- Opportunities for collaboration.
- Career recognition.
- Decrease of non-compliance risks (legal, ethical, institutional and funders' policies).

FOR SCIENCE:

- Facilitates data finding and re-use.
- Enables new research and new insights on the data.
- Valuable data is protected.
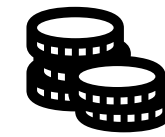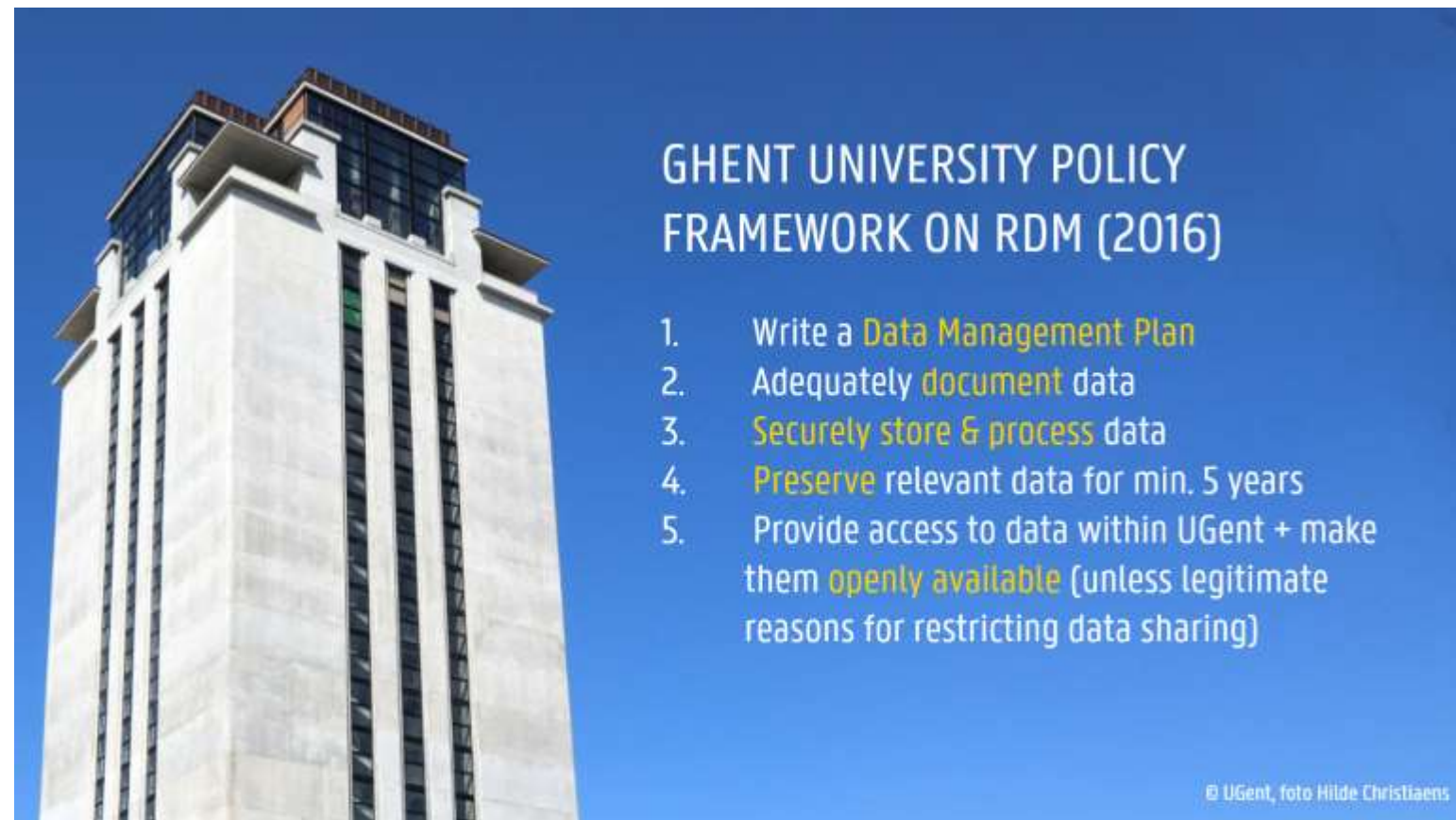- Supports research integrity and reproducibility.

FOR SOCIETY:

- Efficient use of public resources.
- Better quality research can lead to better decision-making.
- Opportunities for citizen science.
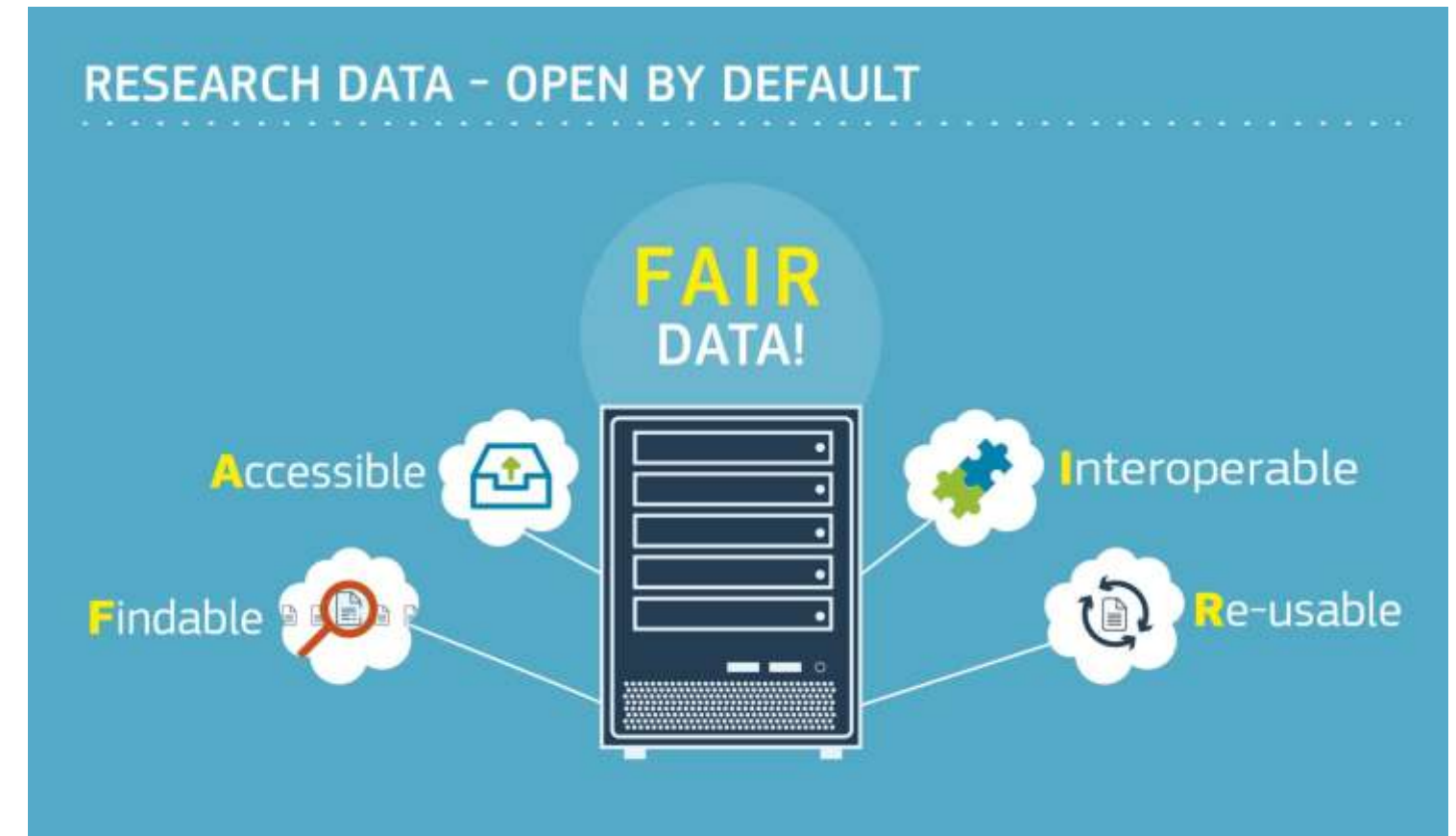- Increased transparency and trust in science.

UNIVERSITEIT GENT

POLICIES AND REQUIREMENTS ON DATA SHARING

🏛️ Institutional policies

GHENT UNIVERSITY POLICY
FRAMEWORK ON RDM (2016)

1. Write a Data Management Plan
2. Adequately document data
3. Securely store & process data
4. Preserve relevant data for min. 5 years
5. Provide access to data within UGent + make them openly available (unless legitimate reasons for restricting data sharing)

© UGent, foto Hilde Christiaens

🪙 HEU Open Science requirements

RESEARCH DATA – OPEN BY DEFAULT

FAIR DATA!

Accessible

Findable

Interoperable

Re-usable

📖 Journals & publishers

"*PLOS journals require authors to make all data necessary to replicate their study's findings publicly available without restriction at the time of publication. When specific legal or ethical restrictions prohibit public sharing of a data set, authors must indicate how others may obtain access to the data.*" PLOS ONE Data availability
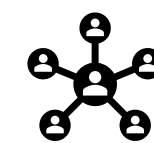
ALLEA European Code of Conduct for Research Integrity

- Access as open as possible, as closed as necessary
- Data as legitimate and citable products of research

— RDM = key part of FWO policy for all support channels

→ "Sound data management is required for all research supported by the FWO."

— Specific expectations:

1. preserve data for a minimum of 5 years after end of research

2. submit a Data Management Plan

3. encourages researchers to ensure that their data management complies as much as possible with the 'FAIR' principles

4. advises researchers to apply the principle 'as open as possible, as closed as necessary'

— Reporting

→ the FWO obliges to register all relevant output via the institutional 'CRIS' systems

https://www.fwo.be/en/the-fwo/research-policy/data-management-plan/
https://www.fwo.be/en/the-fwo/research-policy/reporting/

3. Valuation of diverse research outputs

- **Research applications**
    - Adapted application form: Space for narrative, CV with a broad set of activities, 5 main publications and/or achievements, other scientific output and impact
    - Adapted scoring grid for evaluation
- **Scientific reporting**
    - Specific question: 'Are one or more referable datasets linked to this publication or to the project?'
    - Questions about metadata and accessibility

UNIVERSITEIT GENT

WAYS OF SHARING DATA

UNIVERSITEIT
GENT

DATA UPON REQUEST

Often unavailable

Recent research has found that as little as

6,8% of data was readily made available upon

request

— Only 14% responded
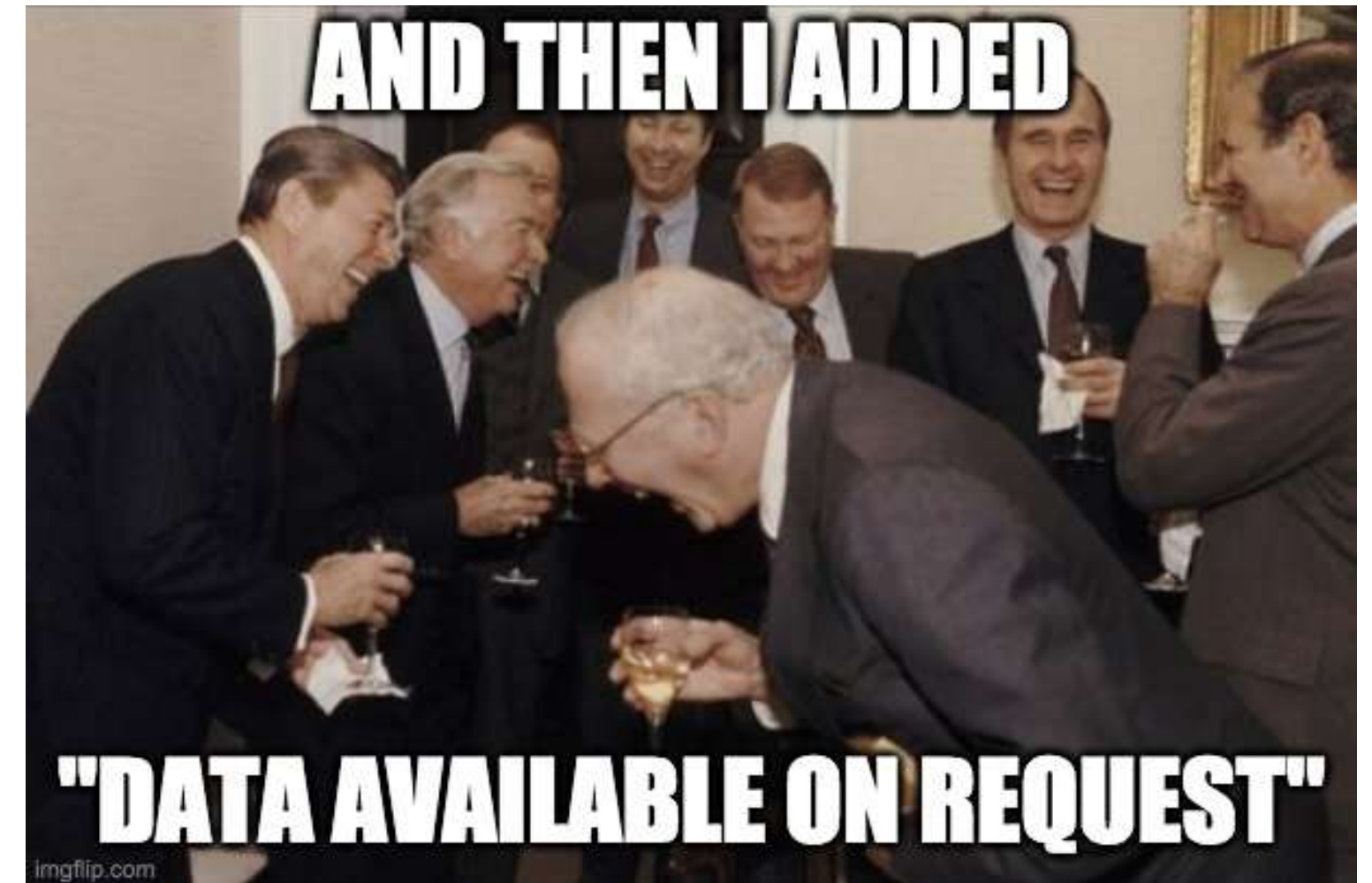
https://figshare.com/articles/presentation/Handson_fMRI_code_and_data_sharing/12911165/1

UNIVERSITEIT
GENT

Gabelica et al. 2022, https://doi.org/10.1016/j.jclinepi.2022.05.019

WAYS OF SHARING DATA
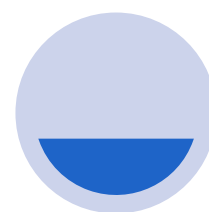
 Upon request
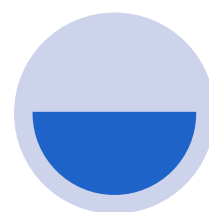
 Personal or project website
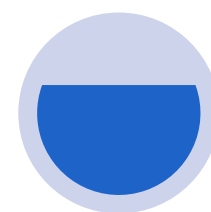
 Article suppl. material
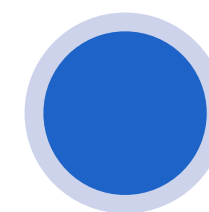
 Data repository or archive

F**indable** A**ccessible** I**nteroperable** R**eusable**

UNIVERSITEIT GENT

DATA LINKED TO ARTICLE PUBLICATION

Data availability

Raw data of scRNA-seq (and output files from CellRanger) and spatial transcriptomics (Visium and stereo-seq) are deposited at the GEO under accession number GSE207592. Processed count matrices and cell annotations (scRNA-seq and Visium) are available online (https://marinelab.sites.vib.be/en). Molecular cartography data are available at Zenodo (https://zenodo.org/record/6856193#.Ytj-fnZBz-g). All other data supporting the findings of this study are available on reasonable request. Source data are provided with this paper.

Karras, P., Bordeu, I., Pozniak, J. *et al.* A cellular hierarchy in melanoma uncouples growth and metastasis. *Nature* **610**, 190–198 (2022). https://doi.org/10.1038/s41586-022-05242-7

FAIR DATA IN REPOSITORIES
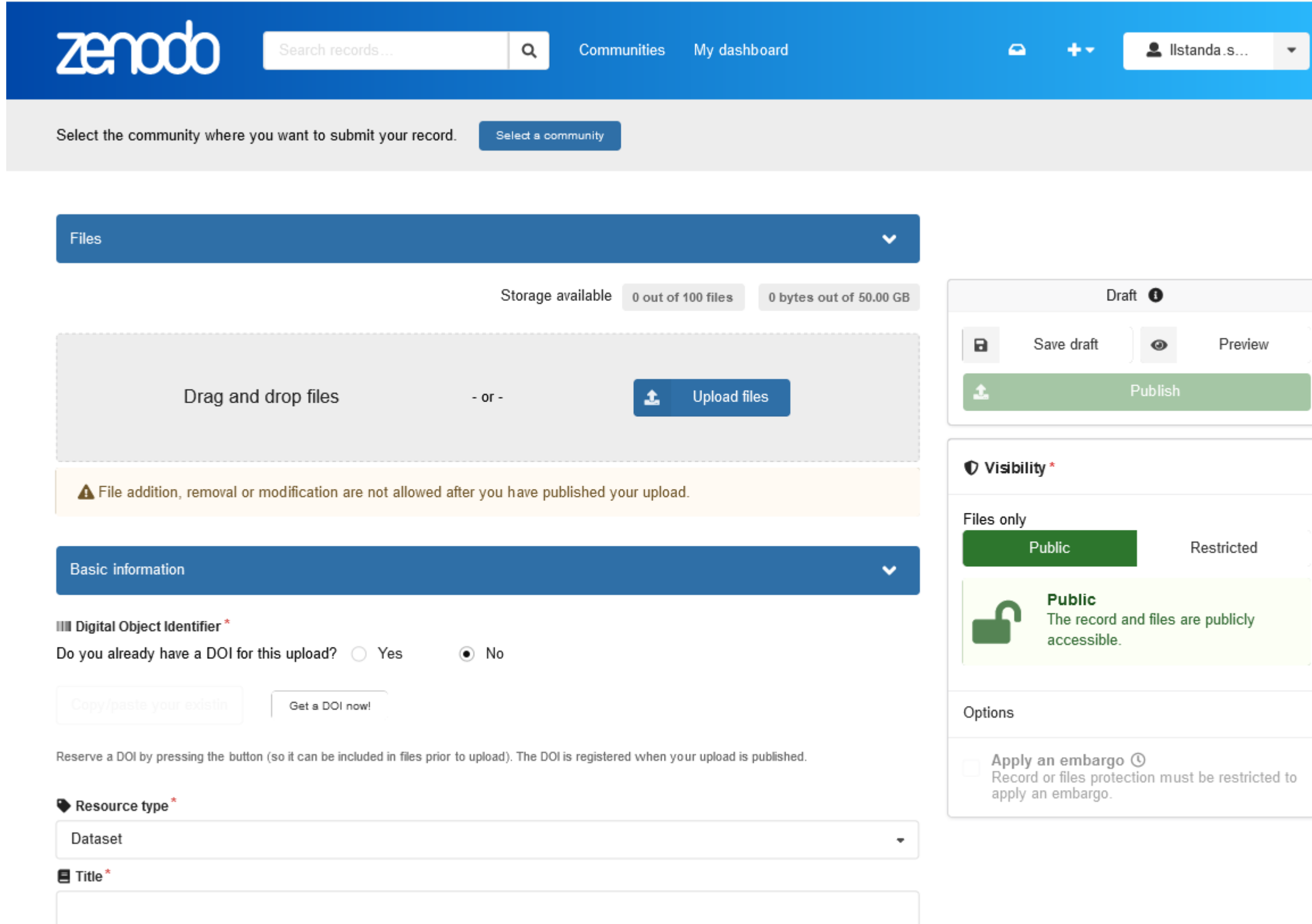
WHAT IS A DATA REPOSITORY?

A database infrastructure that collects, manages and stores data for preservation and sharing
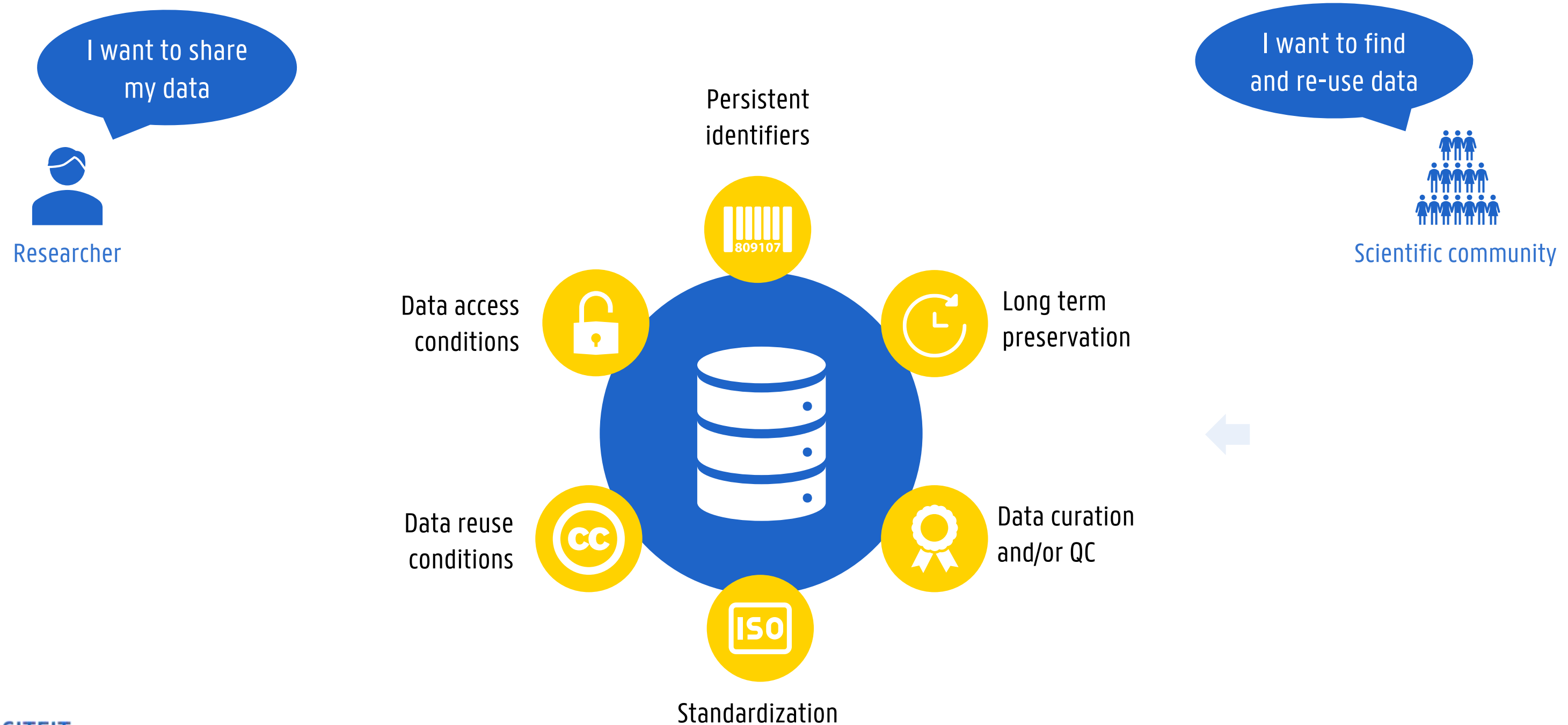
EXPLORING DATA REPOSITORIES

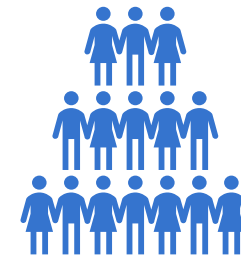— Some repositories have a test environment: sandbox.zenodo.org/ (vs. zenodo.org/)

WHAT IS A DATA REPOSITORY?

A database infrastructure that collects, manages and stores data for preservation and sharing

DATA
FINDABILITY

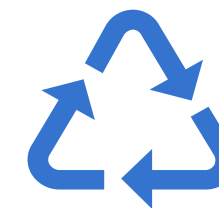*FAIR* 👍

RESEARCH COMMUNITY
SUPPORT

PERSISTENT & UNIQUE
IDENTIFIER

FAIR 👍

LONG TERM
DATA PRESERVATION

FAIR 👍

DATA ACCESS CONDITIONS

FAIR 👍

DATA RE-USE CONDITIONS

FAIR 👍

**UNIVERSITEIT
GENT**

DATA REPOSITORIES & FAIR DATA

→ Key infrastructures that provide functionalities to make your data FAIR.

→ Shared responsibilities between repositories and researchers in making data FAIR

FAIR principle	In practice	Repository	Researcher
FINDABLE	Assign a persistent identifier (PID)	●	
	Data are described with <u>rich</u> metadata	●	●
	Metadata include the PID, are indexed & searchable online	●	
ACCESSIBLE	(Meta)data retrievable via PID using a standardised communications protocol (e.g. html)	●	
	Authentication and authorisation procedure	●	
INTEROPERABLE	Metadata standards	●	
	Controlled vocabularies	●	●
	References to other (meta)data	●	●
REUSABLE	(Machine readable) license / terms of use	●	●
	Detailed provenance documentation		●
	Domain standards	●	●

Shared responsibilities between repositories and researchers in making data FAIR

DATA REPOSITORIES & FAIR DATA

→ Key infrastructures that provide functionalities to make your data FAIR.

→ Shared responsibilities between repositories and researchers in making data FAIR

FAIR principle	In practice	Repository	Researcher
FINDABLE	Assign a persistent identifier (PID)	●	
	Data are described with rich metadata	●	●
	Metadata include the PID, are indexed & searchable online	●	
ACCESSIBLE	(Meta)data retrievable via PID using a standardised communications protocol (e.g. html)	●	
	Authentication and authorisation procedure	●	
INTEROPERABLE	Metadata standards	●	
	Controlled vocabularies	●	●
	References to other (meta)data	●	●
REUSABLE	(Machine readable) license / terms of use	●	●
	Detailed provenance documentation		●
	Domain standards	●	●

Persistent IDentifier

Shared responsibilities between repositories and researchers in making data FAIR

PERSISTENT IDENTIFIERS

	Identifier	Globally unique identifier (guid, uuid)	Persistent identifier (PID)
What	a name that identifies a unique object gene identifier ENSG00000139618 book identifier ISBN 817525766-0	ensures that there are no two identical identifiers that point to different digital objects 123e4567-e89b-12d3-a456-426614174000	ensures that the digital object will remain findable over time and reduces the risk of broken links
How	can be (arbitrarily) assigned by anyone	by using a mechanism (e.g. using a registry service) for assigning identifiers	is determined by the commitment of the organisation that assigns and manages the identifier

Examples of globally unique, persistent identifiers

Identifier:

Organisation:

Scope: Digital objects   Researchers   Research organisations

ORCID FOR IDENTIFICATION OF RESEARCHERS

Connection to contributions and affiliations

Unambiguous identity

Commonly used and required

ORCID: Open Researcher and Contributor ID
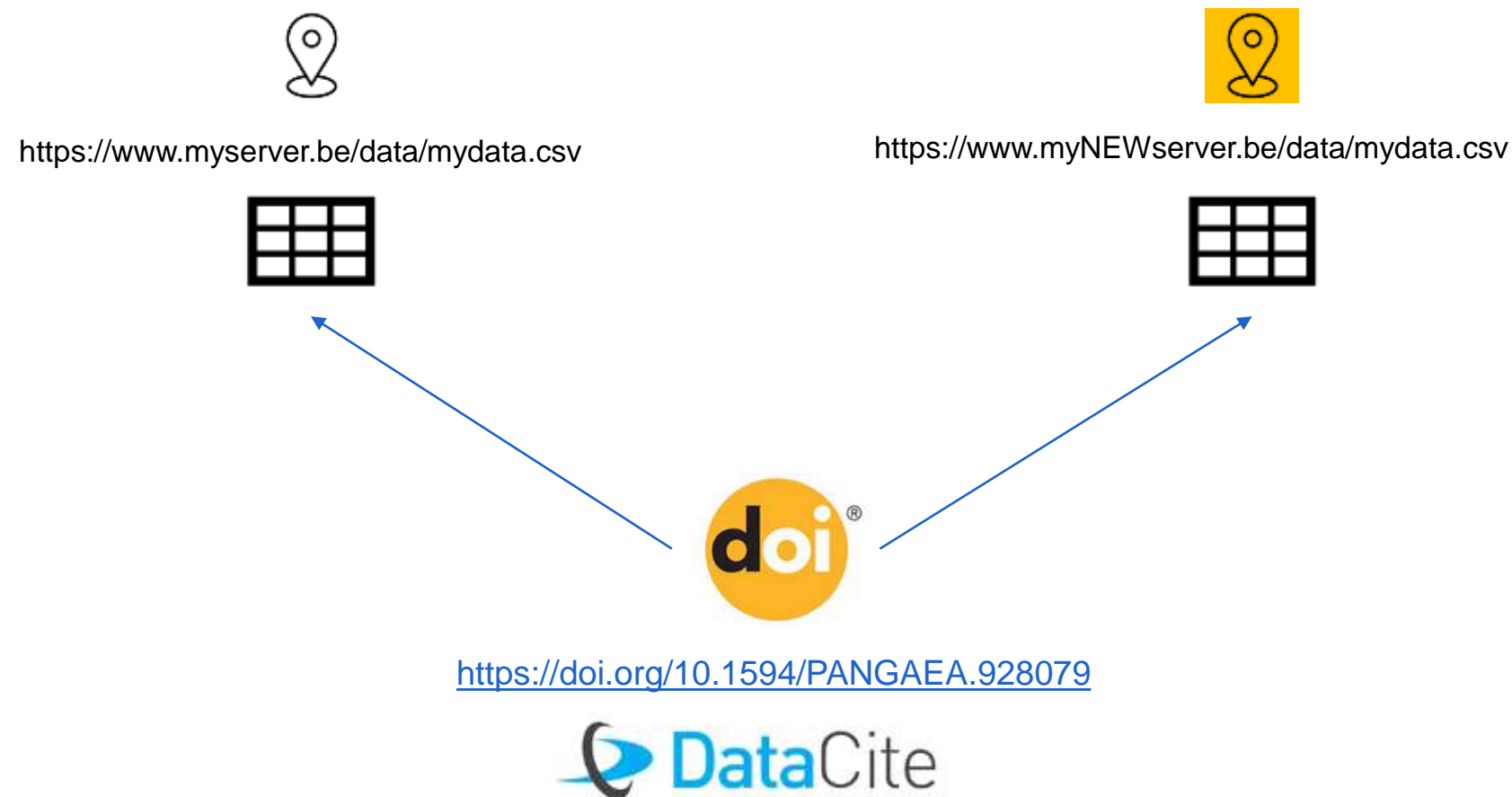
PID: PERSISTENT IDENTIFIER

Globally unique: an unambiguous reference for humans or machines, an identifier associated with only one resource at any time

Persistent: stable and resolvable identifier for the long term

e.g. DOI: Digital Object Identifier

https://www.myserver.be/data/mydata.csv

https://www.myNEWserver.be/data/mydata.csv

https://doi.org/10.1594/PANGAEA.928079

UNIVERSITEIT GENT

DATA REPOSITORIES & FAIR DATA

→ Key infrastructures that provide functionalities to make your data FAIR.

→ Shared responsibilities between repositories and researchers in making data FAIR

FAIR principle	In practice	Repository	Researcher
FINDABLE	Assign a persistent identifier (PID)	●	
	Data are described with <u>rich</u> metadata	●	●
	Metadata include the PID, are indexed & searchable online	●	
ACCESSIBLE	(Meta)data retrievable via PID using a standardised communications protocol (e.g. html)	●	
	Authentication and authorisation procedure	●	
INTEROPERABLE	Metadata standards	●	
	Controlled vocabularies	●	●
	References to other (meta)data	●	●
REUSABLE	(Machine readable) license / terms of use	●	●
	Detailed provenance documentation		●
	Domain standards	●	●

Metadata

Shared responsibilities between repositories and researchers in making data FAIR

UNIVERSITEIT
GENT

DOCUMENTATION

Any contextual information that allows data to be understandable and reusable.

METADATA

Information about data, recorded in a highly structured way, using a set of defined information fields or elements to be machine readable.

The dataset...

Title	
Creator	
Abstract	
Keywords	
Temporal scope	
File type	
License	
Related publication	

UNIVERSITEIT GENT

METADATA IS DATA...

— About data

— A (very) structured type of documentation: records information in sets of pre-defined fields or elements.

— Meant to be read & exchanged by machines

— To facilitate (online) data discovery

— To quickly judge usefulness of data

UNIVERSITEIT
GENT

METADATA IN DATA REPOSITORIES

— Repositories: infrastructure to allow us to share & find data

— Provide **services to create & manage metadata**

— Typically, by filling a form with

— Pre-defined fields

— Lists of controlled terms to choose from (e.g. standard licenses)

— Metadata adheres to a specific metadata standard (interoperability)

UNIVERSITEIT GENT

New upload

Instructions: (i) Upload minimum one file or fill-in required fields (marked with a red star). (ii) Press "Save" to save your upload for editing later. (iii) When ready, press "Publish" to finalize and make your upload public.

Files ❯   Choose files Start upload

Communities ❓   recommended ❯

Upload type   required ⌄

Publication ⦿   Poster ◯   Presentation ◯   Dataset ◯   Image ◯   Video/Audio ◯   Software ◯   Lesson ◯   Physical object ◯   Other ◯

Publication type   [Journal article   ⌄]
Publication type.

Basic information   required ⌄

Digital Object Identifier   [e.g. 10.1234/foo.bar]
Optional. Did your publisher already assign a DOI to your upload? If not, leave the field empty and we will register a new DOI for you. A DOI allows others to easily and unambiguously cite your upload. Please note that it is NOT possible to edit a Zenodo DOI once it has been registered by us, while it is always possible to edit a custom DOI.

[Reserve DOI]

Publication date *   [2021-03-03]
Required. Format: YYYY-MM-DD. In case your upload was already published elsewhere, please use the date of first publication.

Title *   [   ]
Required.

Authors *   [Family name, given names] [Affiliation] [ⓘD ORCID (e.g.: 0000-0002-1825-0097] ⇕ ✕
Optional.

➕ Add another author

Description *   [B I S x₂ x² | ∞ | ☰ ☰ | ≣ ≣ " ⊞ | ↶ ↷ Iₓ | Σ Ω | Source]

DATA REPOSITORIES & FAIR DATA

→ Key infrastructures that provide functionalities to make your data FAIR.

→ Shared responsibilities between repositories and researchers in making data FAIR

FAIR principle	In practice	Repository	Researcher
FINDABLE	Assign a persistent identifier (PID)	●	
	Data are described with <u>rich</u> metadata	●	●
	Metadata include the PID, are indexed & searchable online	●	
ACCESSIBLE	(Meta)data retrievable via PID using a standardised communications protocol (e.g. html)	●	
	Authentication and authorisation procedure	●	
INTEROPERABLE	Metadata standards	●	
	Controlled vocabularies	●	●
	References to other (meta)data	●	●
REUSABLE	(Machine readable) license / terms of use	●	●
	Detailed provenance documentation		●
	Domain standards	●	●

Vocabularies

Shared responsibilities between repositories and researchers in making data FAIR

UNIVERSITEIT GENT

VOCABULARIES AND STANDARDS

STANDARDIZED METADATA

Metadata fields or elements	Metadata values Vocabularies, ontologies, ISO standards
DataCite — Title	▌▌▌▌▌▌ ▌▌▌▌
DataCite — Creator	ORCID
DataCite — Description	▌▌▌▌▌▌▌ ▌▌▌▌ ▌▌▌▌▌
DataCite — Subject	▌▌▌▌▌▌ ▌▌▌▌
DataCite — Date	ISO-…
DataCite — Format	▌▌▌
DataCite — Rights	▌▌▌▌
DataCite — Related identifier	▌▌▌▌▌▌▌ ▌▌▌▌ ▌▌▌▌

UNIVERSITEIT GENT

CONTROLLED VOCABULARIES TO STANDARDIZE TERMINOLOGY

Controlled vocabulary
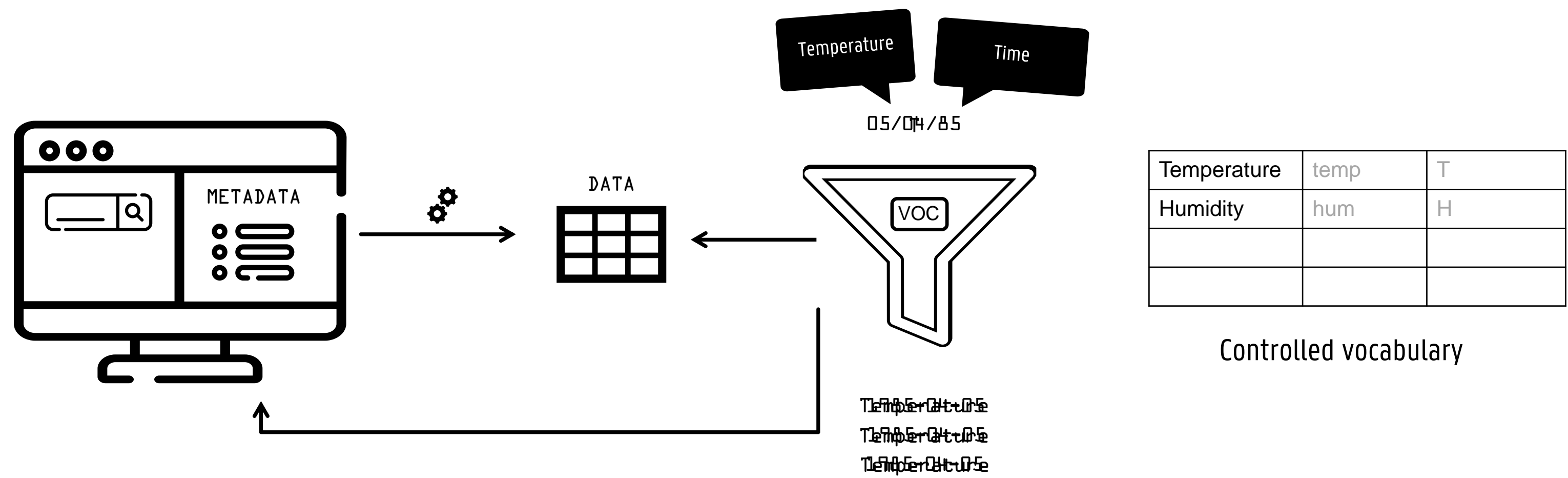
STANDARD VOCABULARIES & ONTOLOGIES

- Enable machine readability (use might be required in some domain repositories)
- Day-to-day example:
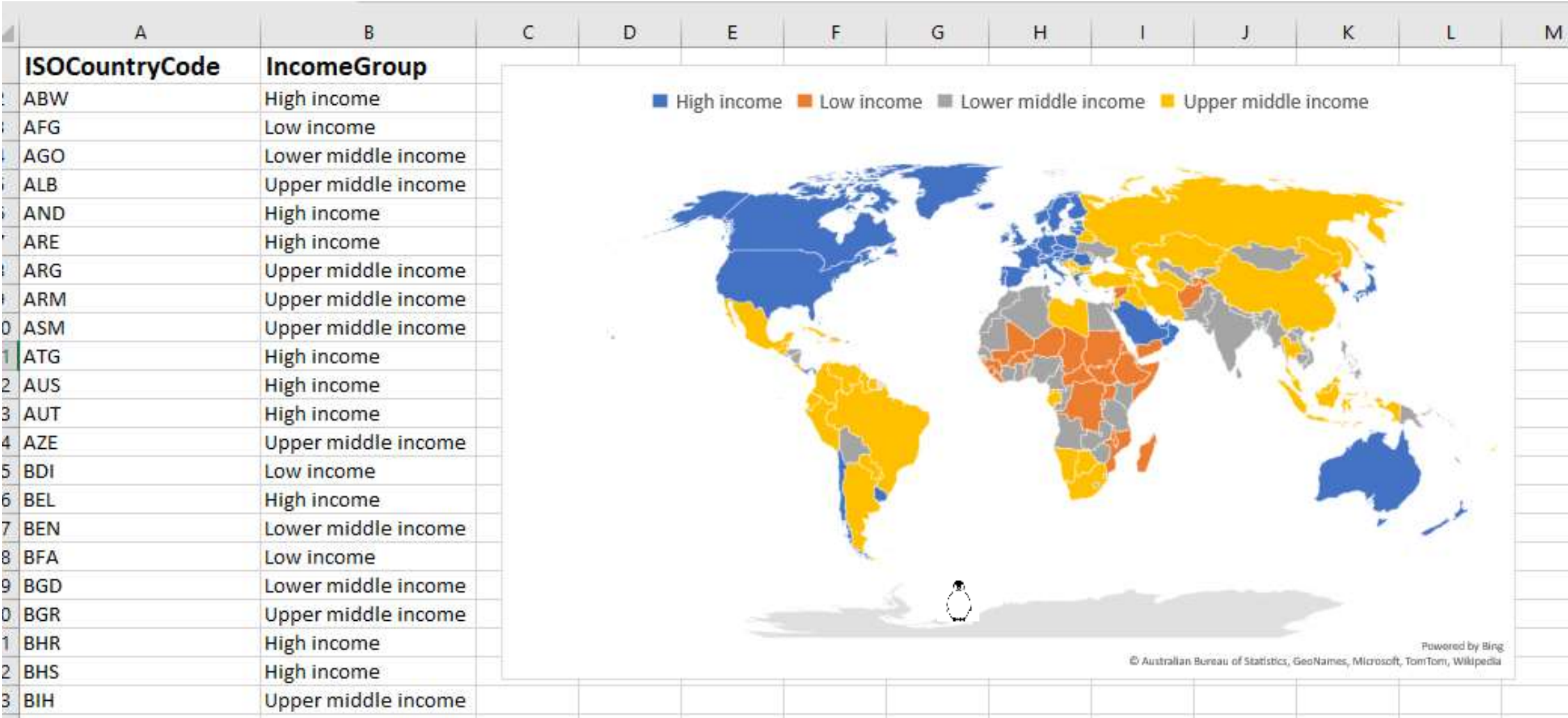
MS Excel recognizes ISO Country codes (or other place names in geonames.org) to automatically create maps

FINDING REPOSITORIES

WHICH TYPES OF REPOSITORIES EXIST?

Types of repositories

GENERIC REPOSITORY

DOMAIN-SPECIFIC REPOSITORY

INSTITUTIONAL REPOSITORY

HOW TO SELECT A DATA REPOSITORY?

Recommended repositories

| RESEARCH DOMAIN'S CONSENSUS | RESEARCHERS' BEST PRACTICES | JOURNAL RECOMMENDATIONS | DATA REPOSITORY REGISTRIES |

e.g. data repositories
recommended by Nature

re3data.org demo

UNIVERSITEIT
GENT

HOW TO SELECT A DATA REPOSITORY?

— Does your publisher or funder recommend a repository?

For example:

Data Repository Guidance

Scientific Data mandates the release of datasets accompanying our Data Descriptors,

View data repositories
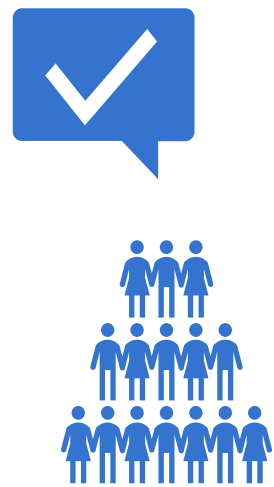- **Biological sciences:** Nucleic acid sequence; Protein sequence; Molecular & supramolecular structure; Neuroscience; Omics; Taxonomy & species diversity; Mathematical & modelling resources; Cytometry and Immunology; Imaging; Organism-focused resources
- **Health sciences**
- **Chemistry and Chemical biology**
- **Earth, Environmental and Space sciences:** Broad scope Earth & environmental sciences; Astronomy & planetary sciences; Biogeochemistry and Geochemistry; Climate sciences; Ecology; Geomagnetism & Palaeomagnetism; Ocean sciences; Solid Earth sciences
- **Physics**
- **Materials science**
- **Social sciences**
- **Generalist repositories**

UNIVERSITEIT
GENT

Source: a list of data repositories

HOW TO SELECT A DATA REPOSITORY?

— Does your publisher or funder recommend a repository?

European Commission | **Horizon Europe**

For example:

Deposit data in a trusted repository and provide **open access** to it ('as open as possible, as closed as necessary')

ELIXIR Deposition Database list

Deposition Database	Data type	International collaboration framework [1]
ArrayExpress	Functional genomics data. Stores data from high-throughput functional genomics experiments.	
BioModels	Computational models of biological processes.	
BioSamples	BioSamples stores and supplies descriptions and metadata about biological samples used in research and development by academia and industry.	NCBI BioSamples database
BioStudies	Descriptions of biological studies, links to data from these studies in other databases, as well as data that do not fit in the structured	

https://elixir-europe.org/platforms/data/elixir-deposition-databases

GENERIC VS. DOMAIN-SPECIFIC REPOSITORIES

COVID-19 datasets in Zenodo

COVID-19 data portal

UNIVERSITEIT GENT

EXAMPLE OF A DOMAIN-SPECIFIC REPOSITORY
COVID-19 DATA PORTAL

GENERIC VS. DOMAIN REPOSITORIES

— Domain-specific repositories

— Visibility: where your peers will likely look for data

— Use of discipline standards & vocabularies

— Search filters, visualization tools tailored to domain

— Interoperability: easier to integrate multiple datasets

— Dataset curation by repository advisors often part of submission process

— Generic repositories:

— Provide rich metadata to make it more findable & reusable

— Generic AND domain:

— DOI, link to publication(s), dataset(s) or other outputs

— Support minimum degree of FAIRness

DOMAIN SPECIFIC: LIFE SCIENCES

"*Caretta caretta*" in GBIF

Search filters tailored to domain & data type -> facilitated by domain metadata standard

Data visualization -> facilitated by data standardization

"*Caretta caretta*" in Zenodo

REPOSITORIES FOR SOFTWARE AND DATA ANALYSIS

Run and publish notebooks

Jupyter notebooks

https://opendreamkit.org/2017/11/02/use-case-publishing-reproducible-notebooks/

https://github.com/CPShub/sim-data

CODE OCEAN

https://codeocean.com
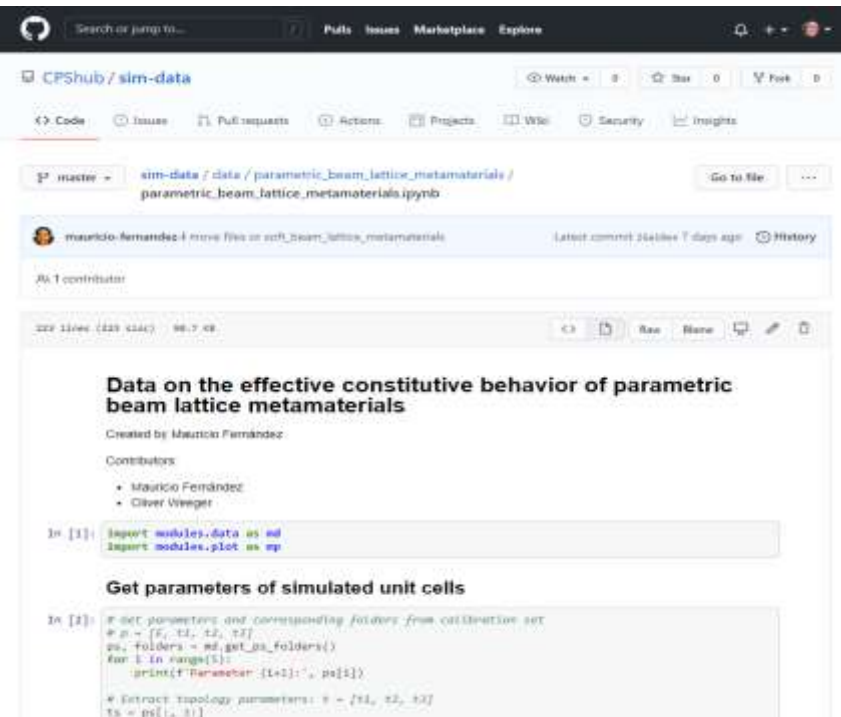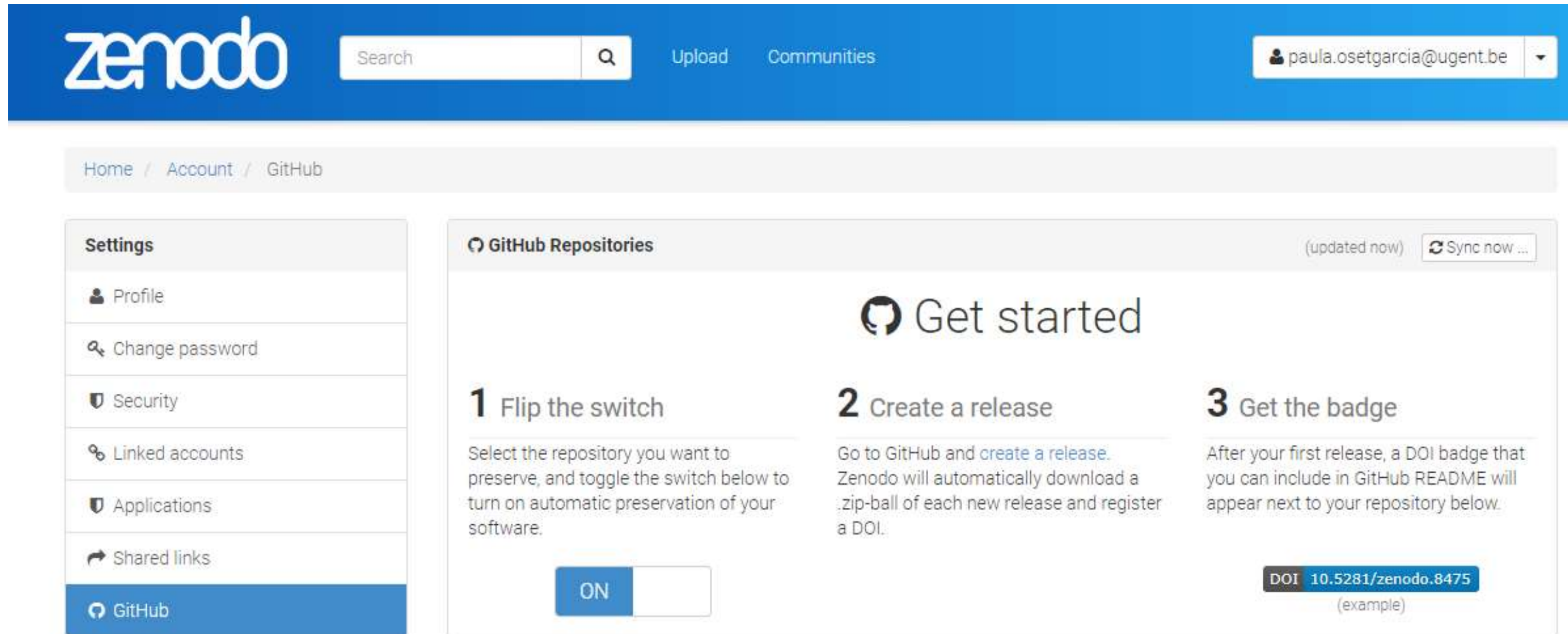
Publish git repositories

https://guides.github.com/activities/citable-code/

UNIVERSITEIT GENT
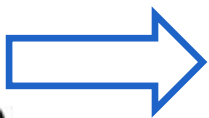
SHARING CODE AND MAKING CODE CITABLE

— GitHub – Zenodo integration: https://guides.github.com/activities/citable-code/

THE JOURNEY TO FAIR DATA

Plan for data management, keep your DMP updated

Check regulations. Explore potential data repositories and their requirements & guidelines.
- Inc. use of standards, vocabularies, etc.

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

FAIRsharing

D|C|C

Document as you go

- Notebook
- Codebook, data dictionary
- Workflows
- Protocols
- Readme files
- Data overview with metadata
- File organization and naming

Deposit data & documentation in repository; describe it with rich metadata, inc. access & reuse conditions

Adopt domain standards during data collection & processing
Implement regulatory requirements

UNIVERSITEIT GENT

PREPARING DATA
DATA DEPOSITION CHECKLIST (GENERIC)

— Gather all relevant (meta)data needed for reproducibility:

— experimental method details, raw data files, organized data tables, scripts and code, data visualizations, and statistical output

— Migrate data to recommended, reusable formats

— Organize files in a logical schema

— Provide a separate readme file

Jørgen Stamp - Digitalbevaring.dk

More info at: https://datadryad.org/stash/best_practices

UNIVERSITEIT GENT

DATA LINKED TO ARTICLE PUBLICATION

https://doi.org/10.1098/rspb.2022.2456

— **Neutral processes underlying the macro eco-evolutionary dynamics of mixed-ploidy systems**

— Felipe Kauai, Frederik Mortier, Silvija Milosavljevic, Yves Van de Peer and Dries Bonte

— All relevant data are within the manuscript and its electronic supplementary material. The code developed for running numerical simulations is available on a GitHub repository at https://github.com/KauaiFe/Polyploidy-V1.0. Readers will also find the code documentation within the electronic supplementary material

— Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.6456230

DATA LINKED TO ARTICLE PUBLICATION

https://doi.org/10.1111/mec.16865

— **Runs of homozygosity reveal past bottlenecks and contemporary inbreeding across diverging populations of an island-colonizing bird**

— Claudia A. Martin, Eleanor C. Sheppard, Juan Carlos Illera, Alexander Suh, Krystyna Nadachowska-Brzyska, Lewis G. Spurgin, David S. Richardson

— The genomic data supporting this study and code used to perform the data analyses within this article are openly available in the Dryad Digital Repository: https://doi.org/10.5061/dryad.ksn02v75k (Martin et al., 2022).

Supporting Information

Filename	Description
mec16865-sup-0001-AppendixS1.pdf PDF document, 1.1 MB	**Appendix S1**

UNIVERSITEIT GENT

DATA LINKED TO ARTICLE PUBLICATION

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9627755/

— **At age 9, the methylome of assisted reproductive technology children that underwent embryo culture in different media is not significantly different on a genome-wide scale**

— Rebekka M Koeck, Florence Busato, Jorg Tost, Heleen Zandstra, Sylvie Remy, Sabine Langie, Marij Gielen, Ron van Golde, John C M Dumoulin, Han Brunner, Masoud Zamani Esteki, and Aafke P A van Montfoort

— The raw and processed array data from this study have been uploaded to the Gene Expression Omnibus (GEO) database and are available under the accession number GSE196432 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE196432). The raw array data from the FLEHS study are available within the GEO database under the accession number GSE110128.

ALWAYS OPEN SHARING?

UNIVERSITEIT
GENT

WHAT ABOUT SHARING OF SENSITIVE OR CONFIDENTIAL DATA?

"As open as possible, as closed as necessary"

OPEN   RESTRICTED   CLOSED

Can be freely used,   – Personal data   Under embargo
modified & shared by   – Confidential data   Unable to share
anyone for any purpose   – Sensitive data
                             – 3rd party data
                             – Copyright/database right
                             – Valorization potential
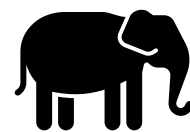
UNIVERSITEIT
GENT
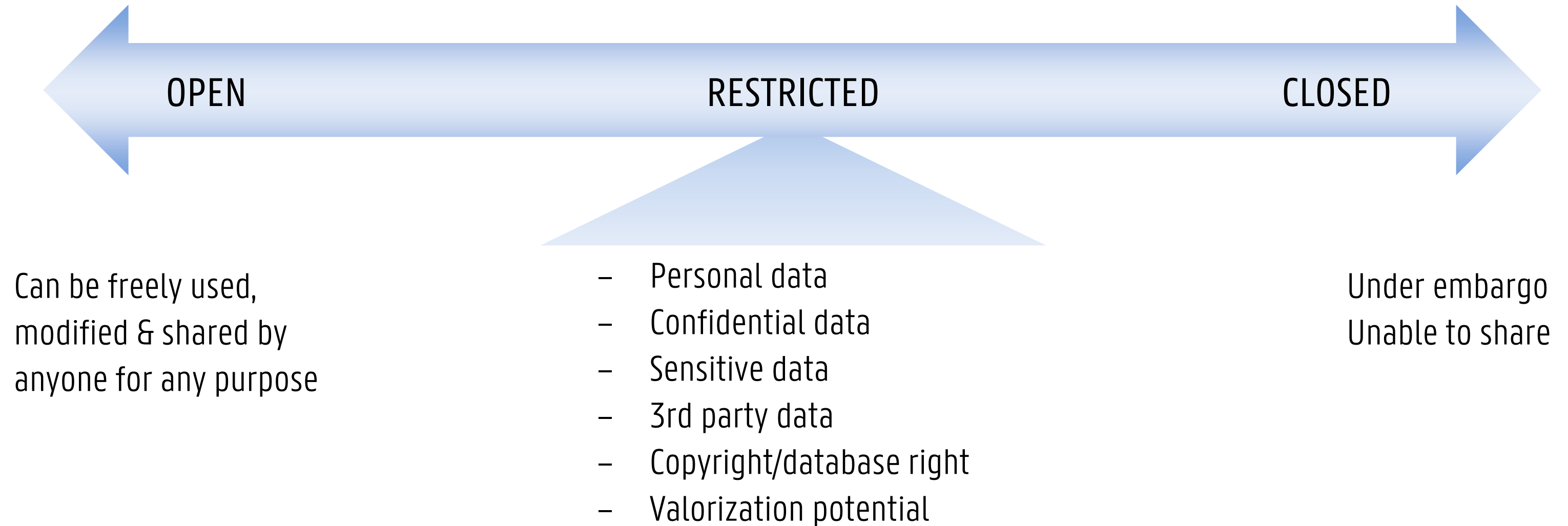
RESTRICTIONS ON DATA SHARING

Personal data   Confidential data   Sensitive data   Third-party data   Copyright   Valorization

More info in next sessions!
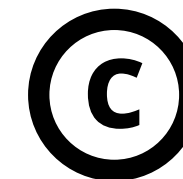
TAKE-HOME MESSAGES

— Publishing data can provide personal (career) advantages, and it benefits science and society at large

— Sharing data is a growing requirement from research institutes, funders and journals

— The recommended way to share data is using data repositories

  → Data repositories provide essential functionalities to make your data FAIR

— Check repository requirements to prepare for deposit during research

— Putting your data in a repository ≠ making data open

UNIVERSITEIT
GENT

Ghent University Data Stewards

RESEARCH DEPARTMENT - UNIVERSITY LIBRARY

✉ rdm.support@ugent.be

🐦 @UGentRDM

💻 ugent.be/en/research/datamanagement