EUTOPIA TRAIN
Transforming Research and Innovation

Selecting a trustworthy research data repository

Training guide data repositories

Julie Jordens – Data archivist
Centre for Academic and Secular Humanist Archives (CAVA), Vrije Universiteit Brussel
In coordination with the Research & Data Management Team (R&D), Vrije Universiteit Brussel

Based on the EUTOPIA TRAIN "Training guide data repositories": DOI: 10.5281/zenodo.7258306 (published november 2022)

This programme has received funding fromthe European's Union's Horizon 2020 Framework programme under grant agreement N° 101017419 EUTOPIA-TRAIN.

VRIJE UNIVERSITEIT BRUSSEL
CY CERGY PARIS UNIVERSITÉ
UNIVERSITY OF GOTHENBURG
University of Ljubljana
upf. Universitat Pompeu Fabra Barcelona
WARWICK

Data Repositories

What and Why?

Data repositories: What?

Database infrastructure for management, storage and dissemination of research data.

- Database is searchable
- Data discoverable through metadata.

- Publish (meta)data during/after research project
- Prolonged public access to (meta)data
- “as open as possible, as closed as necessary”

Types of data repositories

- General purpose
  - Data from all disciplines
- Discipline-specific
  - Linked and relevant to a specific scientific discipline
  - **Choose discipline-specific repository instead of general repository**
- National
  - Linked to specific country
- Institutional
  - Financed , maintained by institution. Captures intellectual output
- Project-specific
  - Focus on specific research project
- Journal-specific
  - But more frequently: list of recommended, trustworthy repositories

Data repositories: Why?

- Facilitates data sharing & reuse (rise citation rate)
- Verification of research results in publications easier
- Avoids unnecessary data collection/creation
- Possibility to integrate several datasets
- Valorisation research output
- Makes sure your data doesn't get lost
- Increasingly required (publishers/funders)

- In summary: **Good scientific practice!**

Source: Foster – Loss of data cartoon
License: Attribution 4.0 International (CC BY 4.0)

What's in it for you as a researcher?

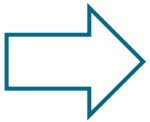
Build scientific **reputation**:

- Compliance data management requirements → more *credibility* as scientist
- Especially if publicly funded research (Give back to taxpayers, open data "public good")

More **citations**:

- Greater *visibility* & impact:

Helps with **funding**



Contribute to science on whole other level AND get rewarded for it!

TIP: Don’t deposit all data in a repository.

- Classified data that shouldn’t be available to general public
- Examples:
  - Personal data regarding human research subjects
    - Identification of participants?
    - Other sensitive information: e.g. religion, political opinion, sexual orientation, genetics, etc.
  - Dual-use: civilian & military applications
  - Data related to possible valorisation research
    - Economic activities based on research

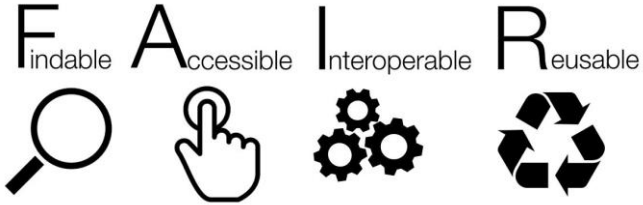
FAIR Data Repositories

The role of “FAIR” in choosing a repository

What is “FAIR”?

FAIR principles:

1. **Findable:** Easy to find by both humans and computer systems.
2. **Accessible:** Stored for long term such that they can be easily accessed and/or downloaded.
3. **Interoperable:** Ready to be combined with other datasets by humans as well as computer systems;
4. **Re-usable:** Ready to be used for future research and processed further using computational methods.



Findability

How can data in repository be findable?

- Globally unique and persistent identifier (PID)
  - E.g., DOI
- Indexed in different catalogues and services.
  - Repository linked to different platforms
    - E.g., DataCite, Open Aire, Google Dataset Search …

Accessibility

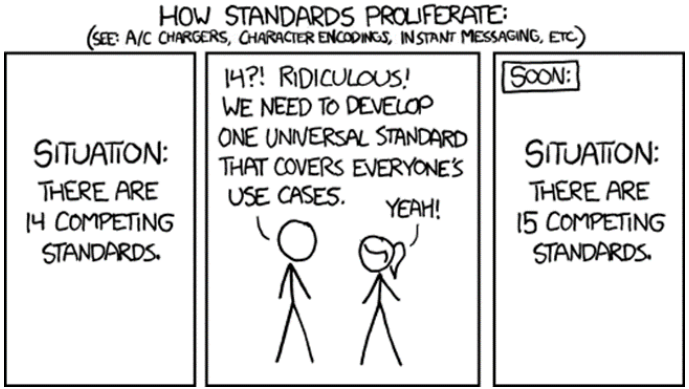
What level of accessibility should repository provide?

- **Open access by default**: data freely shared with anyone
- But: access restrictions!
  - **Restricted access:** Data shared if certain conditions are met. Data are very confidential
    - cf. person-related data
  - **Closed access**: Under no circumstances shared with other researchers
  - **Embargo**: Freely shared, after embargo-period
    - During period data are inaccessible.

Access conditions for data in restricted-access repository → **data use agreement** (*reusability!*).

Interoperability

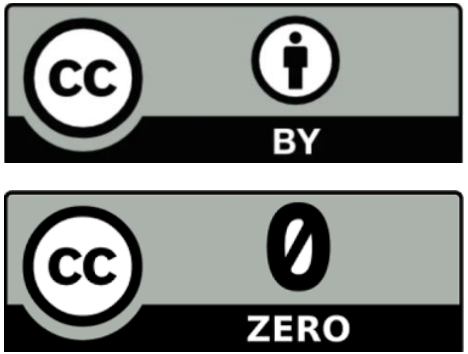
- **Make use of persistent identifiers** (+ findability)
  - Data found through different information platforms
    - Permanent and persistent “link”
- **Comply with metadata standards**
  - Metadata can be exchanged between information systems
  - Common components of metadata that give description and context
    - Date, names, places …
  - More info in EUTOPIA training concerning metadata
- **Standardised formats**
- **Clear exit strategy**



Source: XKCD – Standards
License: Creative Commons Attribution-NonCommercial 2.5 License.

Reusability

- **Terms of reuse** determined, at least:
  - Attribution
  - Copyright requirement
  - Control on commercial exploitation
- If severe restrictions: **data use agreement**
- **Attach reuse license** to data
  - Conditions of reuse instantly clear
  - Example: Creative Commons (and types)
  - Most used:
    - CC BY: Content reused when source acknowledged
    - CC0: Content “public domain”, no requirements/obligations



Characteristics of Trustworthy Data Repositories

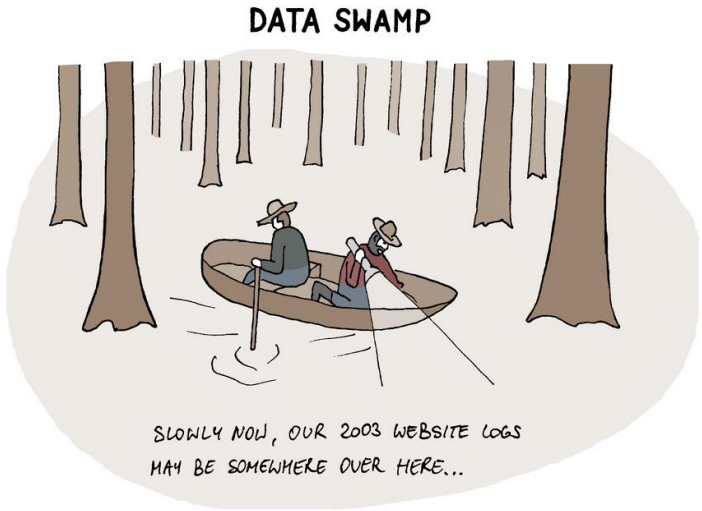
How to know which repository to trust?

A trustworthy data repository should be…

Not a data swamp!

BUT,

- Compliant with **FAIR** data principles
- Compliant with **metadata** standards
- Linked to data **creator** (e.g., via ORCID ID)
- Discoverable via (meta)data **catalogues**
- Includes **access/reuse information** (license)
- Ideally, also **certified**… (see next section)



Source: Data Swamp – Dataedo Data Cartoon
License: Creative Commons Attribution-NoDerivs 3.0 License

Repository Selection Criteria

Trustworthy repositories should meet the following minimum criteria:

1. Provision of Persistent and Unique Identifiers (PIDs)

- Allow data discovery and identification
- Enable searching, citing, and retrieval of data
- Provide support for data versioning

2. Metadata

- Finding of data
- Referencing to related relevant information, such as other data and publications
- Provide information that is publicly available and maintained, even for non-published, protected, retracted, or deleted data
- Use metadata standards that are broadly accepted (by scientific community)
- Ensure metadata are machine-retrievable

3. Data access and usage licenses

- Access to data under well-specified conditions
- Ensure data authenticity and integrity
- Enable retrieval of data
- Information on licensing and permissions (ideally machine-readable)
- Ensure confidentiality and respect rights of data subjects and creators

4. Preservation

- Ensure persistence of metadata and data
- Transparent about mission, scope, preservation policies, and plans (including governance, financial sustainability, retention period, and continuity plan)

Source: Science Europe – Practical Guide tot the International Alignment of Research Data Management

Looking for a repository to deposit your data?

- Check with institution for best options
  - When in doubt, go ***discipline-specific*** or ***institutional***– but definitely, go reputable
- Do not share all data
  - e.g., personal data? Consider only your institutional repository, if available.
- Ideally, go *certified*!

What is Certification?

Reputation of repository can be assessed in three levels:

1. **Listed** in registries (E.g., Re3data)/ broadly recognised in research domain
2. **Endorsed** by relevant funder, journal, or learned/professional society
3. **Certified** to appropriate international standard

Certification one of highest reputations repository can get.

NOTE! Small number repositories certified; many good data repositories not (yet)

- E.g., Zenodo not certified by CoreTrustSeal → not domain-specific

In principle, prefer certified repositories to non-certified repositories

- But: non-certified repository recommended more than certified one? → exercise your discretion.

How to find a trustworthy repository?: Search engines

- Narrow search with parameters
- Most widely recognized inventory of research data repositories:

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Appraisal and selection of research data: what to keep?

- **Your GOAL:** Document and preserve everything needed to verify and/or replicate study + to reuse data for other research.
- Necessary to keep ALL research data?
  - No, keeping large amounts = **costly**
- How to appraise your data?
  - Evaluation criteria
    - Legal or policy compliance
    - Reuse purpose
    - Long term value
    - Weigh up the costs
    - … (see training guide)

Final remarks: Do yourself a favour and plan ahead.

- Make decision to publish or share data in beginning of research project (You can still re-evaluate)

→ including depositing data in repository

  - Begin with end in mind: Write intentions in Data Management Plan
  - Appraisal data (what to keep) also in DMP

Have fun contributing to (open) science!

EUTOPIA TRAIN

Transforming Research and Innovation

Questions / Feedback

welcome at

DMP@vub.be

CAVA@vub.be

This programme has received funding fromthe European's Union's Horizon 2020 Framework programme under grant agreement N° 101017419 EUTOPIA-TRAIN.

VUB VRIJE UNIVERSITEIT BRUSSEL

CY CERGY PARIS UNIVERSITÉ

UNIVERSITY OF GOTHENBURG

University of Ljubljana

upf. Universitat Pompeu Fabra Barcelona

WARWICK