UNIVERSITÀ DEGLI STUDI DI MILANO

Checklist ‘Prepare your dataset’

DATA SELECTION

- ☐ Select research data that are at the basis of a publication or PhD or research data that you want to publish.
- ☐ Ensure the data has not been published yet or received a DOI (or other persistent identifier) in other (trusted) repositories.
- ☐ You can store 30 GB per year. If your dataset exceeds this limit, contact Unimi dataverse support.

LEGAL AND Intellectual Property QUESTIONS

Research at the University of Milan is carried out in accordance with and taking into account existing contractual agreements, legislation, regulations, guidelines or valorisation potential.

- ☐ Check whether you have the right to share the data, taking into account:
  - Existing contractual agreements
  - Rights of third parties
  - Confidentiality obligations
  - Informed consent forms
  - Ethical approvals
  - Funding requirements

- ☐ Check compliance with GDPR in case of personal data.
  - Contact “Ufficio di Segreteria Tecnica del Responsabile Protezione Dati” in order to register the processing of personal data
  - Anonymize/pseudonymize the data if possible. For tools on anonymization, refer to the OpenAIRE data anonymization tool (https://www.openaire.eu/item/amnesia).
  - Consult with the DPO support office in case of questions.

- ☐ Notify Unimi dataverse support (mailto:dataverse@unimi.it) before submitting or disclosing potentially patentable or commercially useful data, in accordance with the Unimi Regulations on Intellectual Property Rights to Research Results.

DATA CLEANING

- ☐ Check errors in the data.
- ☐ Label your data consistently (e.g. data headers, file naming, etc.).
- ☐ Preferably choose open formats or generally accepted standard formats (see UK data service).
- ☐ The size of each individual file must not exceed 5 GB. The ZIP file upload limit is 10 GB.

DOCUMENTATION

- ☐ Provide all the necessary information to understand your data.
- ☐ Complete as accurately as possible the mandatory metadata to describe your work.
- ☐ Create a readme.txt file. The readme.txt file is mandatory when the data format is not FAIR. Describe your collection methodology in the documentation. For more information, consult https://libguides.graduateinstitute.ch/rdm/readme.
- ☐ Explain acronyms
- ☐ If you have many files, please provide a file list (with file names, descriptions of the content and of any connections between the files).

TOOLS

There are tools that can help managing and checking research data. Some of them are free and publicly available:

- ☐ Data quality control: QAMyData (https://ukdataservice.ac.uk/about/research-and-development/) – developed by the UK Data Service (https://www.data-archive.ac.uk), this open source tool can be used to automatically assess and report on elements of quality, such as missingness, labelling, duplication, formats, outliers and direct identifiers. (Free)
- ☐ Data quality control: OpenRefine (https://openrefine.org/) – this data manipulation tool is used for data cleansing and quality control purposes. It is particularly useful for correcting erroneous text values in tabular data. (Free)
- ☐ Numeric data anonymisation: sdcMicro (https://cran.r-project.org/web/packages/sdcMicro/) – a practical R package for checking disclosure risk through examining combinations of key variables. (Free)

FURTHER HELP

- ➢ Didn't find what you are looking for? Check the guide.
- ➢ Still have some questions? Contact Unimi dataverse support.