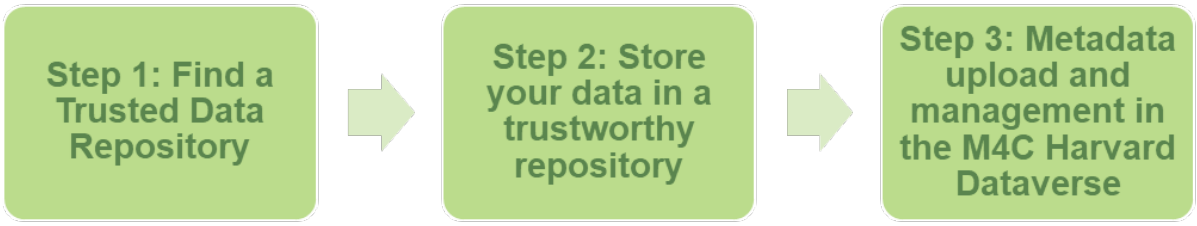
# Quick Guidelines for Data Access and Management (Draft)




## Step 1: Find a Trusted Data Repository

**Locate a data repository** that fits your data's needs and aligns with your discipline. In the context of the M4C project, we recommend to deposit your data in one of the repositories listed in Table 1. If you opt to upload your data in a different repository, we recommend to contact M4C WP5 members (m4c.wp5@mirri.org) to appraise the chosen repository. For example, to assess if the repository is trustworthy and aligned with our data management plan, and assures long-term preservation of the data.

For your information, in addition to Table 1 repositories, resources like Re3data offer a directory of repositories for the permanent storage and access of datasets. You can search for a repository that fits your needs by clicking on search and filtering the results on the Re3data Web interface, for example, filtering only repositories that are open to upload data. We recommend to give preference to an **open data repository** that uses a persistent identifier (PID) system to make the provided data persistent, unique and citable. A list of data repositories open to upload data is available here. Note that blue icons such as doi (see Figure 1) shows repositories that were catalogued as having some sort of persistent identifiers (PIDs). For further details about icons used by re3data, check its FAQ page, see section "What do the icons mean?".


Figure 1. A Re3data repository entry.

**Tip 1**: If a data repository is listed by Re3data as not providing a PID, double check it, because a PID might actually be provided by the repository. For example, Massive (Mass Spectrometry Interactive Virtual Environment) repository is listed by Re3data as not providing a PID, but it does provide a domain specific one, more precisely, the Universal Spectrum Identifier (USI), for instance, MSV000079514.
**Tip 2**: If a domain-specific repository is unavailable, consider using a general-purpose repository like Zenodo (M4C Community), which offers a reliable alternative for storing and sharing research data.

| Name | Data Type | How to submit your data |
|---|---|---|
| GBIF | Checklist data | https://ipt.gbif.org/manual/en/ipt/late |

| | | |
|---|---|---|
| | | st/checklist-data |
| GBIF | Occurrence data | https://ipt.gbif.org/manual/en/ipt/latest/occurrence-data |
| GBIF | Sampling event data | https://ipt.gbif.org/manual/en/ipt/latest/sampling-event-data |
| metabolights | Metabolomics experiments and derived data | https://www.ebi.ac.uk/metabolights/ |
| Mass Spectrometry Interactive Virtual Environment (MassIVE) | Mass Spectrometry data | https://ccms-ucsd.github.io/MassIVEDocumentation/#submit_data/ |
| Zenodo M4C community | Any data | https://help.zenodo.org/docs/deposit/ |
| eDAL! | Research data | https://edal.ipk-gatersleben.de/ |
| Jülich DATA | Research data | https://data.fz-juelich.de/guide/juelich/ |
| EBI ENA Browser | Sequencing data (raw sequencing data, sequence assembly information, and functional annotation) | https://www.ebi.ac.uk/ena/browser/submit |
| EBI MGnify | Microbiome & metagenome data | MGnify \| EMBL-EBI Training |
| EBI BioSamples | Descriptions and metadata about biological samples used in research and development by academia and industry | BioSamples < EMBL-EBI |
| EBI BioStudies | Submissions of all biological data that do not fit in the other, specialised EBI resources, as well as data packages that link together datasets in other resources (e.g., multi-omics) | Submit < BioStudies < EMBL-EBI |
| GitHub M4C organization | Any software code. It is also useful for storing small datasets that require a more advanced “data version control” for tracking changes. | https://docs.github.com/en/get-started/start-your-journey/about-github-and-git |

*Table 1. List of Recommended Data Repositories.*

## Step 2: Store your data in a trustworthy repository

**Deposit your datasets**, including the necessary metadata and tools/instruments for access and reuse, in the selected data repository.

Key steps:

- Attach an open license, such as a Creative Commons license, to datasets that can be shared publicly. We highly recommend to apply the least restrictive license as possible for your data to foster open science and open data, for example, CC0 and CC BY 4.0.
- Ensure that metadata and documentation are complete to maximize usability and reproducibility.
- Get the Web address to access and download the deposited data, for instance, give

preference to its Digital Object Identifier (DOI). This Web address (e.g., DOI) will be informed in the next step, when providing information about your data in the M4C Metadata repository (i.e., a Harvard Dataverse instance).

- Choose open and interoperable file formats whenever possible. This ensures greater data reusability and prevents reliance on proprietary software licenses for future access and reuse. For long-term storage, the preferred option is typically the original 'raw' file format or a lossless format derived directly from it. While different formats may be used during data processing and analysis, it is recommended to convert files to an accepted open format, if any available, before archiving or publishing. Ideally, selected formats should be non-proprietary, unencrypted, uncompressed, and widely adopted within the research community. If an open format cannot be used, metadata should include details about the software required for access, including its name, version, license, and developer. A list of recommended data types and formats in available in **Table 2**

| Data Type | Data Format |
|---|---|
| Documents | PDF (.pdf), OpenDocument Format (.odt, .fodt), Text file (.txt), Microsoft Word (.doc, .docx) |
| Presentations | OpenDocument Format (.odp, .fops), Microsoft PowerPoint (.pptx) |
| Tabular | CSV (.csv), OpenDocument Format for spreadsheets (.ods, .fods), Microsoft Excel (.xls, .xlsx), Relational data (e.g., SQL) |
| Images | PNG (.png), JPEG (.jpg), GIF (.gif), TIFF (.tiff), Hyperspectral images (BIP, BSQ, BIL), X-ray Computed Tomography (.tif, .vol, .raw) |
| Videos | MPEG-4 (.mp4) |
| Audio | MP3 (.mp3), FLAC (.flac), WAV (.wav) |
| Spatial | Shapefiles (.shp, etc.), GeoTIFF (.tif, etc.) |
| Sequences | FASTQ, FASTA, FAS |
| Code | Plain text (with an extension representing the source language) |
| Other | Compressed files (.zip, .gz), Hierarchical data (XML-based formats like RSML), Resource Description Framework (RDF) for metadata |

*Table 2. Data types and formats (from table 6 of DMP).*

**[DRAFT] Step 3: Metadata upload and management in the M4C Harvard Dataverse**

- Uploading Metadata:
  - Create a new dataset entry in the M4C Harvard Dataverse, so you can inform where that dataset is deposited (i.e., hosted in a recommended data repository, see Steps 1 and 2).
  - Add comprehensive metadata (generic and domain-specific metadata) to describe your data.

- Managing Datasets:
  - Edit metadata, add or remove files, and organize datasets into collections as needed.
  - Data Access and Sharing: Define data visibility settings (e.g., public or restricted access). Share links to datasets with collaborators or other stakeholders.

- Dataset Updates:
    - Update existing datasets and manage new dataset versions to reflect corrections or additional information.

## 3. CONTACT INFORMATION AND TECHNICAL SUPPORT

Our team will provide guidance on repository usage, metadata creation, or resolving technical issues.

For general technical support or inquiries about data management, contact the project's data management team at m4c.wp5@mirri.org.

For support on the M4C Dataverse instance, please contact Marco Beccuti.