# Landscape of data repositories & platforms

## The quest for guiding researchers in choosing data dissemination platforms

## Open Data, but where?

Scholarly communication is slowly shifting from papers to richer content, as for instance research data and code. **Data dissemination** has a crucial role and yet it is in its infancy, with visible fragmentation of practices and tools, and too many unknowns for researchers to get onboard.

With our work, we want to mitigate this situation and **help researchers decide** which platforms would be best suited to disseminate their data and code.

The table we elaborated is based on precious feedback from EPFL researchers, and its purpose is both to capture a **variety of use-cases**, and to guide researchers in making their choice.

## At researchers' service

University services act in the continuous effort of supporting students and researchers, providing them with guidance and tools, and anticipating their needs. All while aligning with the services' mission.

We aim to help researchers in **streamlining datasets publication**: shorten the time to dissemination, create awareness, guide them in choosing the ones the best fit their needs.

The **work** towards the current version involved: 3 university services, 4 people, 2 collaborative platforms, 4 major versions of the table, ~60 total work hours, ~6 meetings, lot of coffee.

## Beyond data repos

We started from a list of data repositories used by a specific faculty, then expanded the scope to include other forms of data dissemination (code repositories, databanks, etc.), more faculties and eventually addressed **3 major blurred lines**:

- Between platforms
- Between content
- Between scientific domains of their users

While the fragmented landscape of tools for Open Data and Open-Source imposes to look beyond data repositories, the large variety of practices among researchers implies a combined approach **by research domain and by service**.

## Comparative table WIP

Based on EPFL researchers' input and on our experience, we divide the data dissemination solutions in **8 categories**:

- Data repositories
- Data archives
- Data banks
- Cloud storage
- Code repositories
- ELN / LIMS
- Data analysis platforms
- IT infrastructure

The table is composed by:

**10 tabs**: the 1st tab contains the 8 definitions of data dissemination solutions, plus 8 tabs corresponding to each data dissemination type, and finally a tab for uncategorized platforms.

**22 columns** or more for each category tab: ranging from general information (ex. research domain, managing institution, country), to information about pricing, handling of private data, findability, sharing, etc.

Because it is still **work in progress** and because of two main challenges, voids are a natural component of the table.

As **next steps**, we will explore these ideas for improvement:

1. Streamline the information and fill the voids!
2. Publish it as a filtrable table and collaborative spreadsheet
3. Create a short guide to accompany the researchers
4. Make it evolve as a Decision tree or Chatbot or other form

| TYPE | SHORT DEFINITION (main use) | SHORT DEFINITION (secondary use) | Data (D) / Code (C) / Both (B) | SOME FEATURES | DATA "TEMPERATURE | RETRIEVAL FREQUENCY | PROVIDE (DOI, e |
|---|---|---|---|---|---|---|---|
| 02 - Data repository | Store<br>Retrieve<br>Preserve<br>Discover | Publish, store | B | Online infrastructure where researchers can submit their data. Allows to manage, share, and access datasets for the long-term. Can be generic (discipline-agnostic) or specialized (dicipline-specific): specialized data repositories often integrate tools and services useful for a discipline. The main purpose is to preserve data to make it reusable for future research. Data repositories may have specific requirements concerning: research domain; data re-use and access; file format; data structure; types of metadata. They can also have restrictions on who can deposit data, based on: funding; academic qualification; quality of data. Institutional data repositories may collect a university or consortium of | Warm / Cold | Medium | Mostly yes |
| 03 - Data archive | Store<br>Retrieve | - | D | Storage infrastructure enabling the long-term retention and reusability of data. Provides secure locations for | Cold | Low | Mostly no |

| | GENERAL | | | | | | | License & Pricing | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Name | URL | Domain / Technique | Managing institution | Country | Type | Type 2 | Profit/ Non-4-profit | License | Pricing | |
| 04 - Data bank | | | | | | | | | | |
| Zenodo | https://zenodo.org/ | Generic | CERN & OpenAIRE | EU/CH | 02 - Data repository | | N4P | Many (> 400, from SPDX.org) | free | |
| Figshare | https://figshare.com/ | Generic | Digital Science | UK/self-hosting | 02 - Data repository | | P | CC | free | |
| Dryad | https://datadryad.org/stash | Generic | Springer Nature | USA | 02 - Data repository | | N4P | CC | multiple | |
| Dataverse | https://dataverse.harvard.edu/ | Generic | Harvard University | USA/international | 02 - Data repository | | N4P | CC0 | free | |
| FORSbase | l.ch/ | Social sciences | UNIL | CH | repository | | N4P | ? | free | |
| IEEE Dataport | dataport.org/ | IEEE scope | IEEE | USA | repository | | N4P | CC | multiple | |
| MaterialsCloud | https://www.materialscloud.org/ | Materials Science, Chemistry, Physics | MARVEL NCCR | CH | 02 - Data repository | | N4P | CC-BY-SA | only, see https://archive.materialsclou | |
| DaSCH | https://dasch.swiss/ | humanities | Center for the Humanities, unibas | CH | 02 - Data repository | | N4P | CC | free ??? | |
| Mendeley Data | https://data.mendeley.com/ | Generic | Elsevier | NL | 02 - Data repository | | P | CC | free | |
| EnviDat | t.ch/ | Environment | WSL (ETH-Domain) | CH | repository | | N4P | Many | free | |
| European Nucleotide Archive (EBI ENA) | .uk/ena/browser/home | nucleotide sequencing information | EMBL-EBI | UK/EU/International | 02 - Data repository | | N4P | CC | free | |
| NCBI GEO | m.nih.gov/gds/ | expression | NCBI | USA | repository | | N4P | Many | free | |
| Neuromorpho | neuromorpho.org | curated neurons | George Mason University et al. | USA | 02 - Data repository | | N4P | Other | free | |
| NCBI dbGap | http://www.ncbi.nlm.nih.gov/gap | curated genomes and phenomes | NLM | USA | 02 - Data repository | | N4P | PublicDomain, Other | free | |
| Openneuro.org | openneuro.org | Neuroimaging Data | Standford | USA | 02 - Data repository | | N4P | CC0, ODC, CC | free | |
| PRIDE EMBL | .uk/pride/ | Proteomics data | EMBL-EBI | EU | repository | | N4P | CC | free | |
| | | biological molecules, | | | | | | | | |

01_DEFINITIONS | 02_DATA_REPO | 03_DATA_ARCHIVES | 04_DATA_BANKS | 05_CLOUD_STORAGE | 06_CODE_REPO | 07_ELN-LIMS | 08_ANALYSIS_PLATF

This poster and a static version of the **Data Dissemination table** in their current form are **freely available** at

**go.epfl.ch/FORCE21epfl**

## Contributors

WHO PUT THE HANDS ON THIS PROJECT

**Miriam Braskova**
Scientific advisor
@ EPFL School of Engineering

**Eliane Blumer**
Coordinator or the EPFL RDM team
@ EPFL Library

**Francesco Varrato**
Data Librarian and RDM trainer
@ EPFL Library

**Charlotte Weil**
Data Scientist
@ EPFL School of Architecture, Civil and Environmental Engineering

## Main challenges

### Collecting information

INFORMATION IS POWER, TO EMPOWER

As our own knowledge of platforms for data and code dissemination is limited, we **surveyed** EPFL researchers, asking for the platforms they use. Once we listed the used platforms, we did the online **due diligence** to get information for each platform.

Ideally, the table should be self-explanatory, however the **fragmented landscape** for data dissemination exacerbates this challenge.

For ex., while *re3data.org* is a great resource for data repositories, it leaves out **basic information** as pricing, max allowed size of datasets, physical servers' location, etc. We also contacted directly some platform providers to get the information.

### Categorization

YOUR MINDMAP IS NOT MY MINDMAP

Once we collected the list of platforms and their specifics, we needed to **align our perspectives** (basic vocabulary + collaborative online tools + etc.), integrating our different backgrounds, different tasks, different service goals and expectations.

To make sure to be on the same page we tackled this complexity by structuring the information with categorization of the Data Dissemination platforms that would **make sense** both for us (the services' representatives) and of course for the researchers.

Such a categorization is **not unique**. To anyone wanting to tackle a similar project, we suggest that you engage and leverage on different stakeholders, in particular: RDM experts, researchers, faculty representatives and central services.

Bibliothèque

FORCE11 The Future of Research Communications and e-Scholarship

**FORCE11 online conference: *FORCE2021*, 7-9 December 2021**
**"Joining Forces to Advance the Future of Research Communications"**