Research Data Flow Diagram

User guide

About the Research Data Flow Diagram

The Research Data Flow Diagram (RDFD) is proposed by the Faculty of Medicine of the University of Geneva. The RDFD has been elaborated by the members of the working group “Data Storage” of the “Data Project”. The “Data Project” has been established to identify challenges and propose measures to facilitate research data management in the Faculty of Medicine.

The RDFD has been built upon the data life cycle analysis of the RDMkit (https://rdmkit.elixir-europe.org/data_life_cycle). RDMkit is the ELIXIR Research Data Management toolkit for Life Sciences (https://rdmkit.elixir-europe.org).

The RDFD is designed using Draw.io software (https://www.drawio.com/about). Draw.io is a registered trademark of JGraph Ltd and draw.io AG. JGraph Ltd is a company registered in England, draw.io AG is a company registered in Switzerland.

What is a Research Data Flow Diagram?

The RDFD is intended for research groups and research platforms. The aim of the RDFD is three folds:

- A reflection instrument to evaluate the data lifecycle in your lab
- A decision-making support document to implement better research data management strategies
- A tool to identify improvement needs in terms of practices and infrastructure

The RDFD is a simple yet effective representation of the life cycle of your research data.

Why should I design my own Research Data Flow Diagram?

- **Overview of current lab practices**

Your group is generating and processing huge amounts of data from a large variety of data types. Tools, technologies and research programs evolve. You need an overview of your current research data management practices to ensure their alignment with your objectives, and with institutional regulations and legal obligations.

- **Homogeneity of practices**

You want all members of your research group or platform to have the same level of information about data management. You want the whole group to follow the same best practices, and you want to ease the onboarding of new group members. You need a common framework to easily retrieve data.

- **Identification of improvement directions**

While reflecting on your RDFD, you might realize that you have too many copies of the same datasets, or you might evidence slow data transfers. The RDFD is your entry tool to improve your research data management practices, and to discuss with supporting service providers

- **DMP preparation and communication with journal editors**

Your RDFD is a valuable asset for preparing the Data Management Plan (DMP) required by the Swiss National Science Foundation (SNSF). All information is gathered in a simple diagram. Your RDFD may also be useful when communicating with journal editors.

What should I consider to design my Research Data Flow Diagram?

The design of your RDFD is a reflection on the practices in your lab or platform. The first thing to consider is devoting enough time to this reflection.

You could also consider including several people in the reflection phase: what you think is happening with your research data might not be what is actually happening…

How do I prepare my Research Data Flow Diagram?

The very important question that you must answer first is: "**do I process personal and/or sensitive data?**" This question is far less than obvious. For more information, you can check the personal or sensitive data definition and the ethical and legal obligations at UNIGE, the comments of the CUREG2.0, and the position papers of swissethics. The sensitive and non-sensitive data should be treated separately.

You can start to elaborate a draft of RDFD on a simple piece of paper or on a whiteboard.

It is advised to consider separately the data storage location (where the data are physically) and the nature of the data (*i.e.* raw data, analyzed data), and to start the analysis by the nature of the data. Your data flow will most probably follow part of the 7-step data lifecycle reference.

Resource: Data Life Cycle from the RDMkit (https://rdmkit.elixir-europe.org/data_life_cycle). RDMkit: The ELIXIR Research Data Management toolkit for Life Sciences (https://rdmkit.elixir-europe.org)

How to use the Research Data Flow Diagram template and tool?

The diagram is sketched using Draw.io. This free and open-source solution has been selected, as it is very intuitive to use and easy to handle. For more information, see: https://www.drawio.com/doc/

- **Tool**

Draw.io exits in two versions: a downloadable program that can obtained here, or a web application that can be accessed at https://app.diagrams.net/?src=about#

Once you have installed the program or opened the web application, you should import the icons that you will use. On the left panel, click on the pencil of the section "scratchpad", and import the Drawio_workflow_icons file.

Finally, you should download your preferred starting template and open it in Draw.io:

- Without sensitive data – empty
- Without sensitive data – prefilled
- With sensitive data – empty
- With sensitive data – prefilled

You are ready to go!

Draw.io is very intuitive. For example:

- double click to create a new object (*i.e.* a rectangle). You can even write inside this new object (*i.e.* Raw Data).
- the cursor on the contour of an object makes a green dot appear, drag this green dot to another object and it will link the two objects with an arrow.
- Drag an icon from the scratchpad (*i.e.* large data volume and metadata) to your sketch to document your RDFD.

Now that everything is set up and you master Draw.io, you can design your RDFD.

- **Design**

Horizontally, the RDFD follows the data flow over time, from left to right.

Vertically, data placed at the top are stored for a short period of time while data placed at the bottom are stored for a long period of time.

In the case of a RDFD **with sensitive data**, the non-sensitive (top) and sensitive (bottom) data are separated. It is possible that sensitive data becomes non-sensitive during the data flow (deidentification/anonymization) and thus connect the non-sensitive and sensitive parts of the RDFD (*i.e.* example below).

It is advised to focus first on the Nature of the Data at each step of the data lifecycle, and on the corresponding Storage Duration. You can annotate and comment on the diagram to ease understanding. You can also use red arrows to highlight problematic data transfers.

- **Definitions**

There are some terms that need to be defined:

- Back-up 

: temporarily saved data.
- Preserve 

: data preservation consists of a series of activities necessary to ensure safety, integrity and accessibility of data for as long as necessary. Should not be used in place of Back-up.
- Share 

: sharing data with collaboration partners or with the global research community and society at large
- Data temperature 
: frequency of access to the data (cold = low frequency, up to hot = high frequency)
- Challenges 
: everything you deem interesting to mention

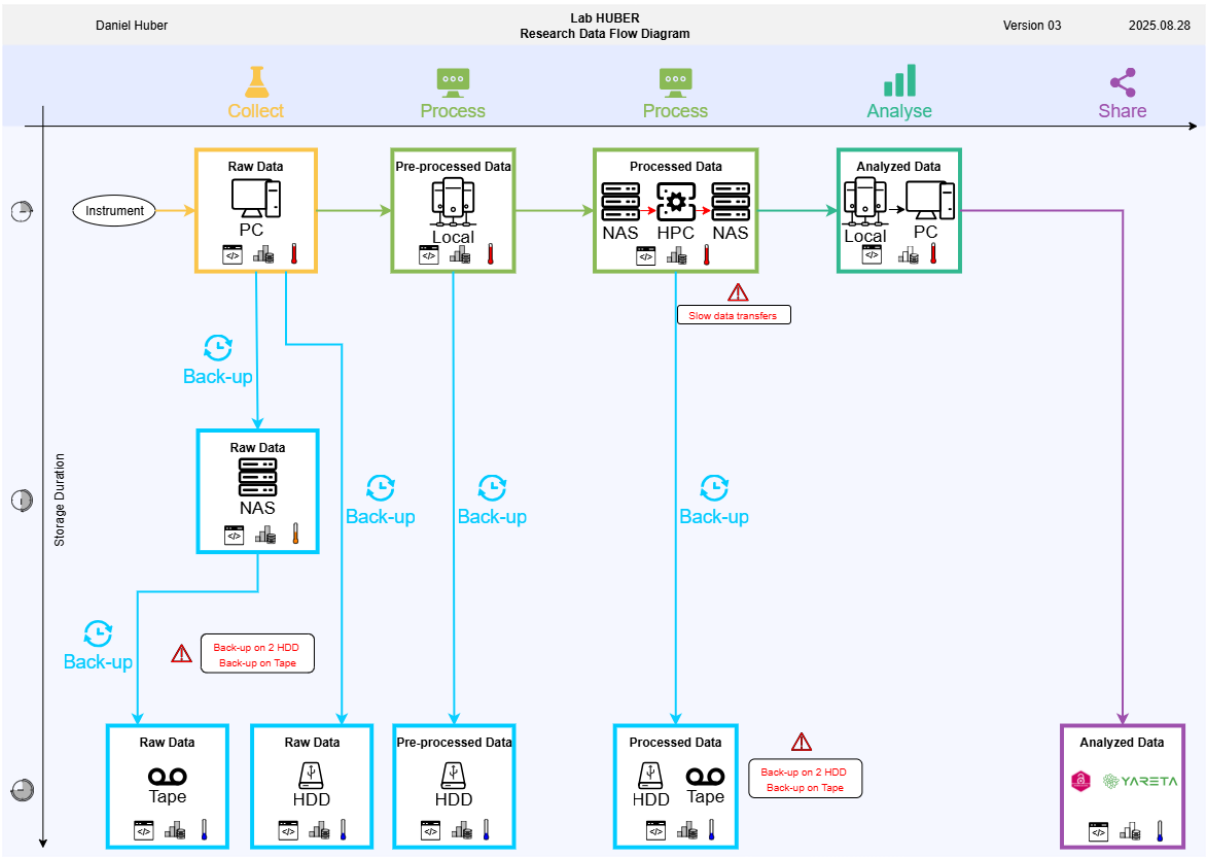
- Metadata: “data about the data”, they describe everything that a new “data user” would need to know to find, understand, reproduce and reuse the data.

There are some terms that you will need to define:

- Storage duration: you should define what you consider short-, medium- and long-term storage.
- Data volume: you should define what you consider low-, medium and high-volume data.
- HPC: you should specify which cluster is used: Baobab, Yggdrasil or Bamboo

Examples

- **RDFD without sensitive data**



- **RDFD with sensitive data**