# Rencontres Data

MINUTES OF JANUARY 2019
SESSIONS - CeRIS & DSI

Institut Pasteur

The Rencontres Data were initially set up by DSI and CeRIS to obtain feedback from the field about issues associated with data management, and to allow people manipulating and managing data at the Institut Pasteur to meet up and discuss subjects of common interest.

The first Rencontres Data took place on 17, 23 and 29 January 2019. These meetings were essentially for researchers and engineers who manipulate scientific data on the Institut Pasteur campus (data managers, bio-informatics experts, bio-statisticians as well as “data advisers” for a unit, etc.).
In all 36 people participated in the first Rencontres Data: 28 scientists and 8 people from support services.

The objectives were as follows:

- **Provide an overall view of management of scientific data**: data life cycle, data management issues at the Institut Pasteur, challenges and advantages of data management, requirements of funders, etc.
- **Present the data management plan template** of the Institut Pasteur and explain how to use it in practice via the REDCap tool.
- **Discuss data management practices** and requirements in the areas of information, training, support, etc.

Extremely useful discussions were held during these 3 sessions, highlighting a number of questions and expectations regarding data management at the Institut Pasteur.

## An expectation for instructions from the management of the Institut Pasteur.

In each session, certain participants expressed a need for recommendations or even obligations in order to harmonise and/or supervise data management practices at the Institut Pasteur:

- **Recommendations for data management best practice and rules** to ensure the data are FAIR (Findable, Accessible, Interoperable and Re-usable). Projects are becoming increasingly collaborative, with frequent exchanges of data, especially internally between research units and platforms. Data management rules would save time and improve efficiency, since data managers would receive “clean” data that can be analysed more efficiently.
- **Obligation to deposit data** (raw and analysed) **and software produced at the Institut Pasteur in a centralised storage space at the Institut Pasteur.** These data and software must be accompanied by standardised metadata, so that the data can be understood and re-used. Several institutes have set up an obligatory data deposition policy: the Max Planck institute, NIH, CERN, NASA, etc. To put this obligation into effect, it was suggested that the Institut Pasteur could follow the example of the practices put in place by these institutes, such as stipulating the deposition obligation in the employment contract of any researcher arriving at the Institut Pasteur.

- **Recommendations for tools and methods to use to ensure effective data security.** E.g.: which mobile phone to use, how to make sure the VPN activates before the Wifi, how to share data securely, which tools and software are reliable.

## Positive feedback about the data management plan (DMP)

**Several funders now require a DMP to be drawn up for each research project they fund.** This DMP is drawn up at the start of the project. It facilitates data management for the duration of the project. It is also used to estimate the costs for management and opening of the data, which can occasionally be paid by the funder. This payment by the funders raises some questions. For instance some participants are not sure whether the costs of recruiting a data manager are eligible. Since each funder has their own policy, the Grants Office is here to help you answer this kind of question.

Many participants have realised the **potential that the DMP would have as a tool to optimise data management and sharing**. They recommend that it should be made visible unless it contains confidential information, to inform other researchers whether or not the data are shareable, and to increase the potential for collaboration within the Institut Pasteur, and also outside it.

The Biomics platform and the Bio-informatics Hub plan to **set up a DMP** that is common to all the projects on their respective platforms. This DMP would be adapted to suit the platforms, with less specific questions than for a project DMP. A working group is going to be created on this subject.

For the DSI, the DMP would make it possible to **view data recently produced at the Institut Pasteur.** In particular, to differentiate:

- between data that must be secured, and data that can be stored in a less secure space,
- between data that requires long-term storage, and data that can or must be deleted (regulatory requirements).

## Positive feedback about the usefulness of the REDCap tool to fill in the DMP

The DMP has recently been implemented in REDCap, and the REDCap form will be enhanced according to feedback from users.

During the 3 sessions, several functionalities offered by the tool were warmly welcomed:

- **Capacity for several people to work on the DMP**, including external partners.
- **Detailed management of user rights**: e.g. to make the DMP visible to the person responsible.
- **Capacity to duplicate the entire DMP with its answers,** to avoid having to re-enter similar information again.
- **Capacity to version the DMP.**

Certain participants are asking for additional functionalities which are going to be examined by the REDCap team at the Institut Pasteur.

## Questions about data conservation

**Observation: at the Institut Pasteur, some data are stored on the DSI Gaïa server (previously called Atlas) but have not been accessed for 2 years (33% of the total data volume).** These data include:

- data awaiting publication,
- data which are worth conserving, because they are unique or were costly to produce (e.g. sequencing data),
- data we know nothing about: author, type, etc.

These data are all conserved in the same way (with the highest possible level of protection). Yet some of these data could be deleted or stored in a less expensive manner: "cold" storage. However, the DSI is not able to sort these data (it does not know what kind of data they are, who their author is or their level of sensitivity). A survey of the data in collaboration with researchers would make it possible to optimise management of data storage.

**Some participants also have questions about long-term storage**:

- Which formats are the most durable?
- Which repositories would allow long-term storage of the data?
- What are the regulatory requirements pertaining to data storage?

**The Archives department of the Institut Pasteur is available to help you answer these questions** and provides individualised support so that you use best practice when managing documents and data.

## Technical questions about the Institut Pasteur's IT infrastructure

**During the Rencontres Data many questions were asked about the functioning of the DSI technical infrastructure**:

- In Gaïa, is it possible to manage user rights (read, write) within the unit itself?
- Are there French cloud solutions we could use at the Institut Pasteur?
- What is the best way to share data within a department (e.g. to share data produced by an instrument)?
- How should we make daily use of the computation cluster in our analyses?
- Data security: what are the solutions and recommendations?
- When a unit member leaves the Institut Pasteur, what happens to their Gaïa space?
- Is it possible to centralise sequence data in a server at the Institut Pasteur? These data could be opened after a certain time if they have not been published. Can we do the same thing for images?

**DSI has taken note of these questions** and will soon provide concrete answers.

## Legal issues regarding the data

During the Rencontres Data, several legal themes resulted in questions being raised:

- Anonymising of data, particularly for data sharing purposes
- GDPR legislation
- Data ownership
- Conditions for use of external data
- License to assign to data and software
- Legal aspects concerning data collection

**Sometimes there is a contradiction between the publishers or funders who require the raw data to be made available, and legal obligations** (e.g. personal data). The participants are not sure how to share data when they are subject to regulations.

## Questions about the way in which data are opened

**Some participants felt it was regrettable that not all the data that are produced and are shareable are effectively opened.** If there is no obligation, they believe the Institut Pasteur researchers will not open their data, which is prejudicial for the scientific community since it will not be able to re-use these data.

Some participants are not sure how to concretely make their data FAIR (Findable, Accessible, Interoperable, Re-usable) and would like to know which repositories are considered to be reliable for publishing data outside the Institut Pasteur.

**CeRIS and DSI have taken note of these questions** and in particular will soon provide a list of trustworthy life science repositories.

## Conclusion

The Rencontres Data provided an opportunity for numerous discussions as a result of which several themes have emerged. The objective now is to examine in more detail the questions raised at the next Rencontres Data. The idea of creating working groups on the different types of data was suggested and will soon be considered.

**CeRIS and DSI are available to answer your questions:**

- about IT questions relating to data: data-dsi@pasteur.fr
- about the data management plan: pgd@pasteur.fr
- about REDCap: redcap@pasteur.fr

Don’t hesitate also to consult the CeRIS webcampus page about management of research data.