Principles of data governance for research organizations -INRAE's approach

Gilles Aumont, Michaël Chelle, Esther Dzale-Yeumo, Odile Hologne, Olivier Philipe, Hadi Quesneville, Stéphanie Rennes

Principles of data governance for research organizations - INRAE's approach

Gilles Aumont[1], Michaël Chelle[2], Esther Dzale-Yeumo[2], Odile Hologne[2], Olivier Philipe[2], Hadi Quesneville[3*], Stéphanie Rennes[4]

1 : INRAE, CODIR, 75007 Paris, France.
2 : INRAE, DIPSO, 75007, Paris, France.
3 : Université Paris-Saclay, INRAE, URGI, 78026, Versailles, France.
4 : INRAE, DAJ, 75007, Paris, France.
**corresponding author: hadi.quesneville@inrae.fr*

Abstract

INRA and IRSTEA, 2 French research organizations, have joined together to become INRAE, a world class institute for research on agriculture, food, and environment, mainly funded by public resources. INRAE has recently set out its principles of "data governance", to cover all the processes required to manage and enhance data sharing on the basis of ethics, legal, economic, technical, and scientific policy criteria. Roles and responsibilities of the actors are defined to ensure a smooth and sustainable decisional process. A "data governance" charter has been written to indicate who decides, which data, and how they are opened according to good practices such as the FAIR principles. With this document all scientists, administrative support, and data managers involved in the data life cycle have a shared understanding of the rationale guiding the global data framework.

We address here the question of "what can be a data governance framework at an institutional level to support both data management, sharing, and reuse.

We've defined 4 key principles at the foundation of data governance: (1) data must be shared and reused while observing the values of science, (2) data must be managed in order to make it F.A.I.R., (3) data should be "as open as possible, as closed as necessary", and (4) open data contributes to innovation and value creation for society. The key rationale of these four principles is that they build a consistent "system" for guiding the decision, as it requires a careful evaluation of all of them. These principles were complemented with a decision-making chain including the main actors.

This "data governance schema" has been built thanks to a participative approach. It is the result of several rounds of discussion and consultation with different groups of people having different views on data and different levels of responsibilities, including interviews of scientific project leaders on a dozen significant case studies. It took 1 year to establish our first guidelines. This schema is today in the process of implementation. Improvements will certainly come after the first real-life usage. However, we believe that these four principles of governance are generic enough to be applied to many research organisms, only the internal organization or process should differ according to the culture and the regulation context.

***Keywords* – up to five keywords; separated by semi-colon**

Research institute; governance; principles; decision-making chains; decision rules

Introduction

INRA and IRSTEA, 2 French research organizations, have joined together to become INRAE, a world class institute for research on agriculture, food, and environment, mainly funded by public resources. Combining these overlapping themes is of major importance to find solutions to face global challenges, namely climate change, food security, and the loss of biodiversity.

To become a major research player recognized in its domains, INRAE has elaborated a data and knowledge sharing policy in the context of Open Science (Amsterdam Call for Action on Open Science, 2016). It recognizes the opportunities of (i) sharing the knowledge and its validation, (ii) improving the integrity of scientific practices including the reproducibility of results, (iii) reusing the data produced to develop new knowledge from validated results, and better control of research costs, (iv) encouraging collaboration to seek the best skills, and (v) promoting research results to the academic, economic and societal spheres. Consequently, INRAE is committed, in compliance with legislative and contract provisions, to make its data accessible and reusable in order to promote virtuous open science practices.

INRAE has recently set out its principles of "data governance", to cover all the processes required to manage and enhance data sharing on the basis of ethics, legal, economic, technical, and scientific policy criteria. Roles and responsibilities of the actors are defined to ensure a smooth and sustainable decisional process. A “data governance” charter has been written to indicate who decides, which data, and how they are opened according to legal status and good practices such as the FAIR principles. With this document all scientists, administrative support, and data managers involved in the data life cycle have a shared understanding of the rationale guiding the global data framework.

We address hereby the question of “what can be a data governance framework at an institutional level to support both data management, sharing, and reuse?”.

Method

The “data governance charter” has been built thanks to a participative approach with the aim of creating a “shared social construction”. Indeed, the initial observation of the INRAE community was marked by i) very differentiated perceptions according to the functional actors concerned (scientists, IT, innovation/patent, legal, etc.), ii) a strong obstacles with the researchers themselves, with the feeling of being dispossessed of their work in the event of opening of the data, or at the very least that “their” data cannot be correctly re-used, iii) for researchers as for scientific managers, a reluctance linked to the additional workload that represents new requirements (data management, sharing, …).

A transversal working group has been set up to bring together these different actors, on the initiative of the INRAE head management, concomitantly with the creation of a Division for Open Science (DipSO). This group proceeded step by step, analyzing a dozen case studies of data collections produced (or processed) in a wide range of scientific fields, and in the context of multiple partnerships, through interviews with their scientific managers. These results were presented and shared with all of the scientific managers (heads of scientific divisions), accompanied by several experienced researchers of their choice, and the functional networks concerned. We collectively build a new working framework resulting from INRAE head management to organize data governance at the level of the institute by making significant changes to our practices in this regard.

Step by step, through more than a year of exchanges and meetings, the construction of a reference document on the principles of Data Governance, made it possible to define a common framework, going as far as defining the roles and responsibilities of each. The more technical aspects have been documented through sheets to explain the regulatory framework, good and bad practices, the experts to whom to turn. The whole is intended to be brought to the attention of INRAE scientists and to constitute the common framework of functional networks with local support.

Results

Four key principles are the foundation of INRAE’s data governance

The consultation process identifies four key principles taken as the pillars of the policy.

(1) Data must be shared and reused while observing the values of science. The researcher has to ensure: (i) the positive societal impact of its results, (ii) data are deposited in disciplinary or the institutional repositories in order to assign them a unique identifier (eg: DOI, URI, etc.) to allow them to be cited, (iii) describe data as metadata to make it findable and intelligible to potential reusers, (iv) to be able to share its work while preserving its ability to continue it. Hence, when data is "open" according to the criteria defined by the other principles, it is possible to postpone the opening in order to protect the scientific exploitation of the data by the project partners. The length of the deadline must, however, correspond to ethical and professional rules, the practices of the communities, as well as the requirements of the funders, and (v) the reuse of datasets produced by others is encouraged.

(2) Data must be managed through data management plans which contribute to the quality of scientific projects by addressing technical, legal, ethical, scientific and economic issues. The data must follow the F.A.I.R principles which provide good practices for data sharing and reuse, but also for data handling including their destruction or archival. This good management practice contributes to the reproducibility of results, a key requirement of research integrity and data quality.

(3) Data should be "as open as possible, as closed as necessary". In accordance with the French regulation, opening public data free of charge is the default choice as part of our public research service mission. Some exceptions are allowed and the combination with other types of data must be addressed, in particular via partnership clauses and / or when protected data exist. Research contracts with private partners can also hampered data openness.

(4) Open data contributes to innovation and value creation for society. The creation of value, in particular economic value, can therefore be performed by users from the private or public sector, be it French or foreign. Thus, the data can be used by third parties to create value from (i) the creation of data services by integrating data with other information, (ii) the integration of computer codes in professional or general public applications, (iii) the artificial intelligence applications trained on the data made available. The whole can feed decision support systems by aggregating data, computer codes, and artificial intelligence engines generating or supporting economic activity for large groups, but also small structures such as start-ups.

The key rationale of these four principles is that they build a consistent “system” for guiding the decision, as it requires a careful evaluation of all of them.

A decision-making chain

These principles were complemented with a decision-making chain including all the actors. This process assumes that researchers are generally in the best position to take the relevant and appropriate decisions, leaving decisions as much as possible at a low hierarchical level to insure smooth and rapid decision, avoiding bottlenecks at higher level due to an overwhelming number of requests. Only the most difficult cases in relation with strategic or regulation issues should be addressed at a highest level.

Consequently, we distinguish 4 levels of decisions. (1) The “data producer” that generates or supervises the technical production of the data, (2) the “scientific manager” in charge of a project, a research infrastructure or a research unit, (3) the “head of research division” which has a disciplinary overview, and (4) the head of the institute which defines the INRAE scientific policy.

Each research division has to define the main direction regarding access and reuse, management and development of data in line with its scientific strategy, and to appoint a "data steward". These data stewards are in contact with the INRAE Chief Data Officer and contribute to the adoption and implementation of the data policy within their division.

The scientific project leaders, research infrastructure managers and heads of research units are in charge of solving data management issues. They guarantee an effective implementation of the good practices and decide the opportunity to share and / or open data in line with the research division’s strategy. Direct contributors who produce, process and manage live data on a daily basis in research projects must comply with the principles of data governance and actively contribute to their proper management. In particular, they should write Data Management Plans (DMPs) and contribute to the actual implementation of the FAIR principles with respect to possible legal limits.

The multidisciplinary data-related issues (scientific, technical, legal, ethical) as well as the need for an animation of the community over time at the level of the research institute, require a national committee chaired by a Chief Data Officer. This committee coordinates the various INRAE actors, proposes update on the Data Governance

operating rules, and investigates unsolved situations at the scientific division level possibly requesting the decision of the Head INRAE management for some specific cases. The Chief Data Officer relies on the « Open Science division » and its data competence center to carry out operations at national level including data awareness-raising, relations with IT infrastructures, and partnership with other national or international players.

Discussions

A system of 4 principles leading to a decision

The 4 principles form the basis of the institute's policy and govern the decisions to be made during the data lifecycle. These principles form an overall logic. They must be examined in turn to allow an informed decision. They allow us to grasp the subject of openness on all of its facets, taking into account the points of view of the actors: (i) the scientific dimension with the ethical values it carries, (ii) the data management dimension highlighting good practices to be observed, (iii) the regulatory dimension with respect for legislation and contractual commitments, (iv) the dimension of innovation for society with the creation of value that it produces. This is by examining the opening conditions according to these 4 dimensions, that the final decision can be obtained. Each principle posing or not possible restrictions and indicating a possible degree of openness.

Responsibilities

Through these principles, the individual responsibility of researchers from a public institution like INRAE is enforced. Indeed, even if the data belong most of the time to the institution which has released the resources and skills necessary to produce them, only the researchers at the origin of the production of these data can guarantee their quality, their good use, and correct FAIRification. Their role is central for choosing an appropriate broadcast channel for their community. The risks of erroneous re-use of data, in good faith, even with malicious intent or deliberate misinformation, would be limited when the choice of open data supports has been rigorous (recognized and known support, reviewed by peers, complete and solid contextual information).

But the collective dimension of ethical, deontological, political, strategic reflection remains necessary. This is the role of head management. The research staff of a research institute such as INRAE is composed by a very large diversity of researchers, engineers, and technicians. Scientific domains cover biology, environmental sciences, computer sciences, sociology and economy. There could be large differences in research practices and the feeling of scientific values of researches. A common vision is needed to promote the same data culture across the institute aligned to the institute scientific strategy, but also defined ethical and deontological values. Head management must provide this general guidance by proposing a unified framework.

Some data are more related to an institutional vision such as long time series or patrimonial data which have to be conserved over long period of time which could exceed a researcher career in a given laboratory, or even the existence of some structures such as laboratory, platforms, research infrastructures. The responsibility of the head management is in this case strongly engaged.

Data sharing modalities

From our analysis, it results from the different constraints that the data opening takes place on a spectrum from the most closed to the fully open (Figure 1). Thus, we distinguish:

Closed data: These are data for strictly internal use at the institute. They support work in progress and do not correspond to a final result which have to be delivered. Some special case such as personal data or sensitive data can also be considered in this category.

Shared data: This is data whose use by persons external to the institute is possible under certain conditions. They might be used by partners of a project, or by a consortium for instance. This use is therefore restricted and can be controlled via authentication or a device that identifies the re-user. Unless restricted by a contract, these data should be opened at the end of the project or after a given embargo.

Open data: This is data that can be reused by everyone, free of charge, and without any control. They correspond to final data associated with a scientific result.

A content modality: It is possible to open only part of the data, or to open only after processing of the data (noisy effects, scale modification, anonymization, etc.), if this information keeps scientifically.

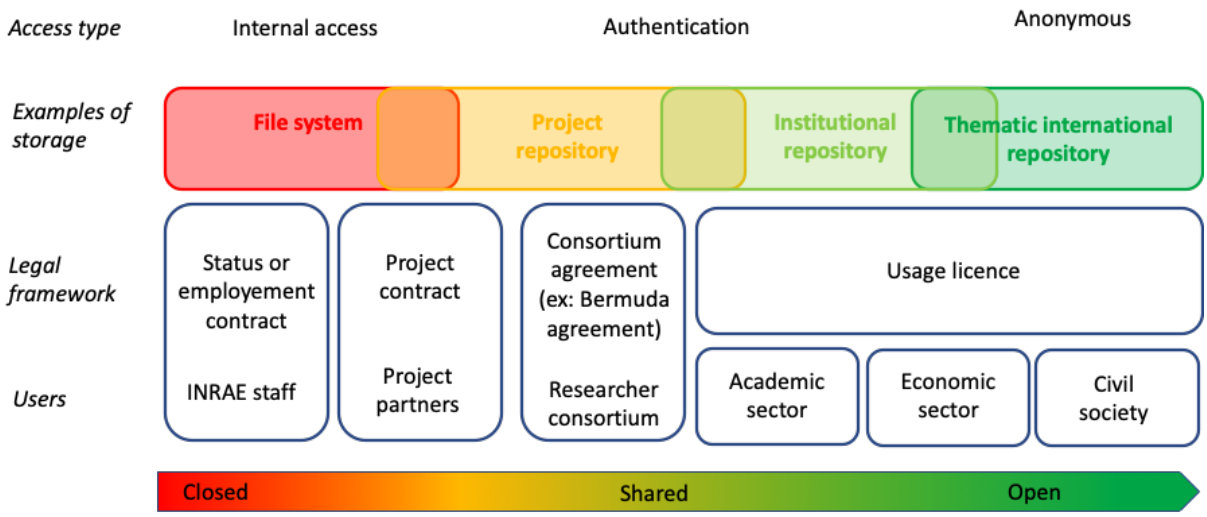

Figure 1 : The data opening spectrum

Monitoring the usage to improve

A particular attention will be paid to understand the impact of this opening on the scientific, economic and societal spheres. Analyzing the positive and negative aspects will be important to develop the system. Correction must be brought if the impact is not sufficient or has negative effects.

We consider that monitoring the usage and the impact of the shared data needs to define good numerical indicators. Finding indicators such as counting downloads or views of a dataset is often proposed as they are straightforward to obtain. But are they good indicators that will allow to have a correct look on the data usage ? In other words, Does the interpretation of the indicator value give any clue about the positive or negative impact of data opening ? We think that finding good indicators is the result of a rigorous approach to ensure that they will be efficient to follow the desired trends.

First of all, the user target has to be defined. It could be the scientific community, the civil society, or the economic sphere for example. All these targets might involve different indicators, and if we target several user-types, they must be a concatenation.

For each targeted user-type, objectives have to be determined. Objectives can be to promote innovation, develop collaborations, increase scientific trust, accelerate research, etc … Once defined, we need to determine if we search for an increase or a decrease.

Only after this, indicators can be chosen according to their ability to show the evolution towards the desired objective goal. Then the condition of success has to be determined. Is an increase (or decrease) of the indicator value is enough, or do we want to reach a given value.

Traceability of data will be key for any indicators. Using permanent identifiers (PID, DOI, URI …) as recommended by FAIR principles (Wilkinson et al., 2016) is a good start to follow the data set usage, but relies on the good will of the user to cite correctly when they reuse them. We clearly need to improve information systems for a better traceability, using for example blockchain technology (Crosby et al., 2015).

Conclusion

We are today in the process of implementation. Improvements will certainly come after the first real-life usage. However, we believe that these four principles of governance are generic enough to be applied to many research organisms, only the internal organization and process could differ according to the culture and the regulation context. Monitoring all the process will be our next challenge to improve our governance organization and rules.

Acknowledgements

We warmly thank our INRAE collaborators for their active involvement during interviews, working group meetings and general discussions, and for the richness of the exchanges. This is a collective work that involved more than the authors of this paper.

References

Amsterdam Call for Action on Open Science (2016). This document is based on the input of many participating experts and stakeholders of the Amsterdam Conference 'Open Science – From Vision to Action', hosted by the Netherlands' EU Presidency on 4 and 5 April 2016. It is a living document reflecting the present state of open science evolution. https://www.government.nl/topics/science/documents/reports/2016/04/04/amsterdam-call-for-action-on-open-science

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

Crosby, Michael; Nachiappan; Pattanayak, Pradhan; Verma, Sanjeev; Kalyanaraman, Vignesh (2016). BlockChain Technology: Beyond Bitcoin. Sutardja Center for Entrepreneurship & Technology Technical Report. Applied Innovation Review 2, June 2016. University of California, Berkeley. http://scet.berkeley.edu/wp-content/uploads/AIR-2016-Blockchain.pdf