

2nd LEARN Workshop | Vienna | 6th April 2016

Research Data Management towards Open Science – Round Tables

Notes of Round Table 3:

How might setting policies ensure research data can be shared and be re-usable?

Moderator: Paolo Budroni (UNIVIE)

Rapporteurs: Martin Moyle (UCL), Imola Dora Traub (UNIVIE)

Round Table 3 was opened by discussing the following:

Who among the attending institutions has released/published an RDM Policy? | When? | Who was the author of the policy? | Who is planning to release/publish an RDM policy? | When?

Main questions:

Where should data be stored? | Who decides on it? | Who controls the integrity of data? | Who decides about planning horizons and periods? | Who controls the release of data? | Who controls the publication processes of data? | Who should not control the publication processes of data? | Who manages the data throughout the research data lifecycle? | Who defines “the research data life cycle”? | How long should the data be retained? | What is the minimum retention period for research data and records? | What are the exceptions for the minimum retention period in a policy? | What kind of data deserves the right to be “forgotten”? | When should research data and records be deleted or destroyed? | Who decides this? | Who carries this out?

Part I. – Discussions

- Who decides about policies in the organisation?

Due to international settings or to the collaborative nature of research, it is not always clear where authority rests. Relationships reaching beyond the institution should be taken into consideration. UCL example: senior committee sign-off; but it drew partly on stipulations made by research funders.

- The body making policy also has to make decisions. Should these be delegated or left to third parties?

It is always necessary to consider how data is going to be used. The problem is that infrastructures are fragmented and some countries lack a national research strategy. Sometimes there is no central funding council.

Example: In Italy there is a centralized facility (Cineca, for storage and deposit software), but it is up to the individual University to determine how to use it. Cineca does not have a policy, just guidelines. Underlying universities need policies, definition of unified common elements, Open Science commons, guiding principles, and standards.

- **So who decides?**

This should be the university.

Within the university, the guiding principles should be to combine top-down requirements and bottom-up needs. It is also important to consider the variety of the different research disciplines. Some already have well-established structures. Policy needs to be flexible. Also to be considered are the needs of all stakeholders: institutions, data producers, research funders, data curators.

Publishers also influence policies and therefore they should be considered as stakeholders.

Finding common ground is very important. That is, finding the parts that fit for all stakeholders. Also important is the consideration of the impact of policy decisions on different disciplines. Ergo, a lack of buying-in will mean that the policy will probably fail.

- **Who decides what sort of data is retained / preserved / in scope of policy?**

Generally, policies should be created by institutions and funders. The policy of the **institutions** should be aligned with research funder policies. Considering the role of **research funders**: research funders should have a policy to push researchers in “the right direction”.

In a first step, the main stakeholders of a policy were identified, in a second step, roles and duties were assigned.

Prioritized answers about who decides about retention were then: **scientists / data producers / scholars**.

Service provider / IT lead are not deciding entities. They decide on the mechanism of preservation, formats, types, and so on.

- **So who decides what kind of data is stored?**

Policy-making body. But researchers should have the right to say which data are / are not important. Also research funders have a right to propose stipulations. But funders do not always meet the costs of their stipulations.

- **Who should control the release of data?**

Funders should determine when data is released. The end-of-project release should be a condition of funding.

Could have different policies for different subject areas. Exceptions are important.

- Who controls the integrity of data?

Provider of storage is responsible for physical integrity. Content is not the responsibility of repository. The data producer has the responsibility for content, meeting standards agreed with the research funder.

- How long should data be stored? Who decides who discards?

Preferably as long as possible.

Policies should be time-limited. 10 years? Funder should help to define, but what about costs? 10 years after last access?

Part II. – Remarks

Interested attendees sent us the following suggestions after the workshop:

a) Where should data be stored?

Suggestion 1: Big data should be stored in disciplinary repository, microdata should be stored in Institutional Repositories provided that the repositories are Openaire compliant. Preferably they should be certified.

Suggestion 2: The decision about the storage of the data depends on their volume. Concerning big data the principle of "algorithm to data" should be applied.

Example: The Sentinel 1 satellite data (1 Tb per day) are archived in a facility at Arsenal (part of the Earth Observation Data Center EODC) near the Vienna Scientific Cluster 3 (VSC3) so there is a direct link to the high performance computer and no need to copy huge amounts of data between infrastructures. (see also CERN network with their distribution between computing centers)

The other infrastructure at ZAMG the CCCA Data Center is hosted here and collects the climate data for Austria and provides data access for everyone interested in a local infrastructure.

So the answer would be: it depends.

b) Who decides on it?

Suggestion 3: Right now the repositories are organized around scientific communities who define their needs and find a "host" who can provide the services. The institution decides to do it, but a consensus has to be reached by all parties involved for every single step in the process.

c) Who controls the integrity of data?

Suggestion 4: The data provider controls the integrity of the data but allows checking and

feedback from the repositories and, in a further step, the users of the data.

d) Who decides about planning horizons and periods?

Suggestion 5: The institution providing the repositories decides on planning horizons and periods but this depends on the opportunities available in the near future (funding).

e) Who controls the release of data?

Suggestion 6: This depends on the data and how the organisational context the data was collected/generated. In some cases, the copyright lies with the funding bodies, whereas in other cases it lies with the institution, and above that some data is commercially interesting and has its own copyright. So it would be specified by rules and regulations in research contracts and/or by the ones binding the institution.

Part III. – Key outcomes

1. Universities should define institutional policy

Not so obvious. Complex environment. International / collaborative nature of research – always connections outside the institution. Not always clear-cut where authority resides. Existing infrastructures are fragmented. Some countries lack national research strategy e.g. where there is no central funding council. All the more reason for universities to take ownership/leadership.

2. Research funders should also have a policy

May wish to make stipulations about data. Also have a (moral) obligation to push researchers in ‘the right direction’. Institutional and funder policies should be aligned.

3. Policy development should involve all stakeholders

Policy development should involve all stakeholders who are involved. Institutions, data producers, research funders, data curators; publishers also influence policy.

Need clear stakeholder definition to define roles and responsibilities in the policy framework.

4. Find the common ground across disciplines

Some have well-established structures already. Policy needs to be flexible and sometimes allow for exceptions. Finding the common ground is very important: find the parts that work for all. Guiding principles. Are there many? Need buy-in / credibility or policy will fail.

5. Policy-making body should determine scope of policy / types of data

But researchers should have right to say which are / are not important in any given project. Also research funders have a right to make stipulations. (But, funders do not always meet the costs of their stipulations.)

6. Storage provider is responsible for physical integrity of data

The policy should consider that the storage provider is involved in issues like: safeguarding identity and integrity of data, security in order to ensure permanent access.

7. Data producer responsible for quality of content

Meeting standards agreed with research funder.

8. Funder should define when data is to be released

Ideally, end-of-project release should be a condition of funding. Could have different policies for different subject areas. Exceptions are important.

9. Retention policies are essential

Retention should be time-limited. 10 years? Funder should define, but what about ongoing costs?