sc | nat

Swiss Academy of Sciences
Akademie der Naturwissenschaften
Accademia di scienze naturali
Académie des sciences naturelles

Open Data and Data Management – Issues and Challenges

Monday, 29 October 2018, 10:00–17:00
Kaserne Bern

Discussions in the working groups, 28th November 2018

Introduction

In the afternoon of the SCNAT Workshop for Swiss stakeholders held on 29 October in Berne, the participants could choose to join one of the four following Working Groups (WG) for a 1-hour session:

- WG1 on Data Management Plans: Practices and challenges, convenor: Ana Sesartic Petrus
- WG2 on special aspects of small-scale university labs, convenor: Heinz Gäggeler
- WG3 on special aspects of large research infrastructures, convenor: Francesco Pepe
- WG4 on international competition, convenor: Ruth Durrer

The minutes of the short reports by the working group convenors are already included in the overall minutes of the meeting. The aim of this second document is to provide – in a uniform format – additional information collected by the convenors on the discussion in each of the working groups.

WG1 on Data Management Plans: practices and challenges

Ana Sesartic Petrus prepared a series of slides presented during the working group to focus the discussion. She reminds that a Data Management Plan (DMP) is a brief plan written at the start of a project and updated during its course. It shall define what data will be collected or created; how will the data be documented and described; where will the data be stored; who will be responsible for data security and backup; which data will be shared and/or preserved, and how the data will be shared and with whom. DMPs are demanded by SNSF since October 2017 and by the Horizon 2020 EU funding programme. The goal of the SNSF is that research data should be freely accessible to everyone – for scientists as well as for the general public. This is already stated in Article 47 of the SNSF Funding Regulations of January 2016. The DMP is just one of the tools to reach this goal. It allows the planning and documenting of the life cycle of the data from data collection/generation to publishing, sharing and preservation. The data shall be in a format and a repository such that they are Findable, Accessible, Interoperable, and Re-usable (FAIR).

Sesartic Petrus opens the discussion by first collecting issues from the participants. On the choice of the right repository, scientists question how to recognise a FAIR repository and note that the Registry of Research data Repositories (re3data.org) lists thousands of repositories and has no filter to choose only FAIR ones. On the communication between the SNSF and the research community several points were raised. The researchers feel that they were not involved enough by the SNSF when the DMP was created. They question what is the reward of doing a DMP. SNSF should rather present intrinsic motivation, instead of saying "you have to do it because we say so". The research community wants to be treated as equals by the SNSF. Instead of getting imposed new regulations top-down, they wish prior discussion on equal footing in a bottom-up process. Some researchers fear that if the goal is to open as much data as possible, we run into the same problem as with articles – a flood of information nobody can process/read. Therefore, quality shall be prime over quantity.

On the definition of data, it has been noted that, for instance in medicine, it is often difficult to define what information can be considered as data. It is also not clear what are the raw data and what counts as processed data. People are confused about the difference between Open Data and Open Science, there is a widespread misunderstanding of these terms. Another issue is that data become obsolete sooner or later. It is useless and confusing to keep openly available data that are not valuable anymore compared to newer data obtained with better methods. Finally, it is noted that different countries have different definitions and compliances to Creative-Commons licences. How can we make sure, that the user upholds the licence and does not misuse and exploit data commercially?

In a second step, Sesartic Petrus opens the discussion on the way forward. On the choice of the right repository, it is suggested to extend the SNSF shortlist of FAIR repositories. Concerning the communication between SNSF and the scientists, the SNSF should better communicate that data management is useful in itself, regardless of regulations. Besides this, there is the need to better involve the research community in SNSF decisions via an improved exchange of ideas on equal footing. It is advised to start early with the education of future generations of researchers by making Open Science and data management an integral part of university education. Some "wild ideas" came out of the discussion on how to handle open data. It has been suggested that institutions should not handle them, but start-ups, which collect all public data and make sure that they are open and for free. It is feared that big publishers would imprison the data behind a paywall. It is finally proposed that Swiss data should remain in Swiss possession and only be distributed inside Switzerland to protect the intellectual property of local research communities against financial exploitation. The feasibility of such an approach is questionable, however.

WG2 on special aspects of small-scale university labs

Heinz Gäggeler made a short survey among the participants that shows that the DMP is well accepted by about 80% of the people and often already implemented in new research projects. The idea to follow the FAIR data principles is also well accepted as an ethical guideline. The request for data storage in repositories has been extensively discussed with controversial opinions. The issue is clear for published data; all agree to archive them in a repository and this is mostly already implemented. However, processed and especially raw data have different meanings for different research fields. Some groups (e.g. biology) fully agree to store raw data as well. Other groups (e.g. those working with large infrastructures such as accelerators) deny usefulness of storing such data. The reason is that the effort to store them in a valuable and useable way for external people is far too high. There is no consensus found, but all agree to make raw data available upon request. In general, this would however be limited to colleagues with competence in the field. Another point of discussion is how to proceed in international collaborations with partners from outside Europe. Some standardization would be welcome.

WG3 on special aspects of large research infrastructures

Francesco Pepe reports from the discussion that it was strongly questioned what Open Data is useful for and whether it is really needed and for which public. Several people expressed their doubts about the universality of Open Data policies. The question of costs-to-benefit must be addressed. Even the FAIR principles are questionable and should be subject of debate. There is an issue with large research infrastructures having often very specific policies, which are binding for the researchers: embargoes, data property times, etc. These rules cannot be overruled. It is certainly true that large organisations have already open data approaches. However, the objective is often data preservation and efficient access. Such facilities have addressed these questions because of necessity and common sense, rather than by imposed rules. Large research infrastructures, in particular, have often to deal with extremely large and complex data sets. The level of data to be made open, the target public, and the access tools are subject to strong optimisation needs. If well addressed, this will cause huge costs for data preparation. Who will pay for all this? The money invested in data management is at present taken from the same pool as research funding.

Hans Rudolf Ott proposes that data should be accessible but in a controlled way, like when applying for funding, beam time, or observing time. A researcher who wants to make use of available data would have at least to make a formal request for this data, with an exact description of what his/her goals are, what will be the methods, etc. This would not be in contradiction with the Open Data approach, but would allow a better customisation and control of the data flow. As for data generation, one could imagine that accessing archived data is associated with a certain cost, which has to be covered. Some persons found that this approach would be a reasonable (less costly, better controlled, more adapted) alternative to a blind, indiscriminate open-data policy. There was a general feeling that goals of the SNSF DMP must be better clarified. While on the one hand the DMP is mandatory and must be approved prior to the release of funds, it is said not to be part of the evaluation of the proposal. While it is said that everybody can describe his/her own approach, it is required to commit to the FAIR data principles. Many participants do not understand why the SNSF did not inquire about the opinion and needs of the researchers well before. A strong wish for a bottom-up and discipline-specific approach has been expressed, because policies and rules cannot be applied universally across all disciplines and organisations.

WG4 on international competition

Ruth Durrer structured the discussion of her working group with a brainstorming approach collected on a flipchart on advantages, disadvantages, and suggestions concerning Open Data in the context of international competitiveness.

The listed advantages of Open Data are to: encourage international collaboration; increase the visibility of own data, research and institution; enable research in financially less strong countries; help the recruitment of talents/skills; improve data quality and structure; foster cultural diversification of interpretations; enable small and medium enterprises to profit from research for speeding up innovation; provide access to results from expensive/unique research infrastructures; allow learning from mistakes of others; open a new and highly transferable field of skills that can lead to a professionalisation of data management; enable thinking outside the box by combining datasets in a new way.

The listed disadvantages are that: costs associated to preparing Open Data go away from research funding; Switzerland may become a data provider country; there is a risk of getting scooped after a lot of work; there is a risk of monopolism by profit repositories; it can lead to demotivation for taking cumbersome data; it may lead to two categories of scientists: the ones who produce data and the ones who work on data collected by others; there is a risk of scientists loosing creativity by preferring an easy reuse of existing data; it is a too big effort compared to the benefits; it is a task with little valorisation for young scientists, thus weakening their career path; the bias of the evaluation system favouring quantity over quality will go on with Open Data, unless this gets properly adapted; there is a need to get acknowledgement of data usage or some kind of reward for sharing the data/code; there is an important issue related to the preservation of the intellectual property.

The proposed suggestions to keep the advantages and reduce the disadvantages are to include data-handling and Open Data practices in evaluation and hiring criteria, as well as in university rankings. This, however, with a differentiated assessment depending on the research fields. There should be more pressure to Open Data on costly, big experiments than on small research groups, especially for data/results relying strongly on intellectual accomplishments. The development of repositories shall be bottom-up and making sure that they are non-profit. A standardisation of data formats per field would help, and the FAIR data principles should be followed by everybody to enable interdisciplinary research.