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

Minutes of the workshop, 8th November 2018

Executive Summary

Organised by the Swiss Academy of Sciences (SCNAT), the one-day workshop on Open Data and Data Management gathered around 120 participants on Monday, 29 October 2018 in Berne. The day started with presentations on the views of the European Union (EU) and the Swiss National Science Foundation (SNSF), followed by the views of representatives across the science community. The afternoon was devoted to more specific exchanges of ideas in four distinct working groups followed by a panel discussion. Globally, it was a very successful event enabling the sharing of partially contrasting opinions on the topic by all involved parties about one year after the SNSF had introduced mandatory Data Management Plans. The discussions were constructive with an open mind. The push for Open Data in the EU and Switzerland can be seen within the overall scheme of Open Science, making publicly funded research freely available to the public, in terms of (a) Open Access to publications, (b) Open Access to obtained research data (Open Data), and (c) Open Access to software tools (Open Source) used to analyse the data. The general aim is not only to make results/ publications freely available, but also to permit to verify obtained results and possibly even to improve the data analysis in comparison to the approach followed by the original research team. This is supposed to bring more transparency, enhance reproducibility, foster collaborative transnational research, speed up innovation, and improve citizen involvement. These objectives are well recognised by the scientific community, but the implementation of Open Data is not trivial and brings up also questions with respect to intellectual property and proprietary rights as well as fairness in competition if other communities are not following the same approach. While providing the data of published results is simple and often already implemented, the challenge is (a) how to make available extended data to allow reproducibility of the results, (b) to which community should they be provided, and (c) how expensive is this procedure in terms of manpower and costs. The main outcome of the workshop is that a general practice cannot be followed, because each scientific field has its specific needs and limitations. Actually, even within a discipline there would be a large variety of solutions. The quality versus the quantity of the provided data matters. The potential use of the data shall be identified and also the target users: Is it for specialists in the field, scientists in other fields, students, or the general public? Obviously, the volume of the data and their level of processing and documentation depends on this. Another important issue is to balance the benefits with the costs implied by preparing the released data and providing the support to potential users. It turned out from the panel discussion that it remains to be defined how to fund this effort that shall not be at the expense of research. The scientists wish a bottom-up approach, in which each scientific domain can participate in defining what should be an appropriate best practice on Open Data. It has been suggested that SCNAT – possibly jointly with swissuniversities, the SNSF, or other institutions – could help setting up and following this process in the years to come. There is also an identified need to involve in the dialogue other members of the Swiss Academies of Arts and Sciences sharing similar concerns.

Welcome and introductory words

Friedrich K. Thielemann, President of Platform MAP/SCNAT, on behalf of Marcel Tanner, President of SCNAT, welcomes everybody and looks forward to an interesting exchange of ideas and thoughts on this timely topic.

Views and plans from the EU side

Christophe Rossel, member of the European Open Science Policy Platform (OSPP) and of the SCNAT Executive Board, on behalf of Jean-Claude Burgelman, Head of the Unit 'Open Data Policy and Science Cloud' at the Directorate General RTD of the European Commission (EC), presents slides on some of the EU strategies concerning Open Data. He introduces first the paradigm shift in the modus operandi of scientific research, from a closed system to an open and sharing one. Open Science offers opportunities such as a better return on research and innovation investments, a faster circulation of new ideas, more transparency of the science system and more cross-disciplinary research for solving societal grand challenges. In 2016, eight policy priorities were defined, four with regard to the use and management of research results and data – open data, European open science cloud (EOSC), altmetrics, now called next generation metrics and future of scholarly communication, i.e., open access. The other four are in relation with research actors (scientists, institutions and funding agencies), namely rewards for open science activities, research integrity, education and skills for young scientists and finally citizen science. The past milestones and the development of science policies in the successive framework programs are presented. Whereas open access is mandatory for the EC, it is however recognized that the data should be "as open as possible, as closed as necessary", with the possibility to opt-out in specific cases: e.g., when the privacy of data needs to be preserved, when issues with intellectual property rights arise or when access to data might jeopardize the project's main objectives. The process is running with the integrated recommendations of the Open Science Policy Platform (OSPP), which includes 25 organizations from eight stakeholder groups. Rossel then presents the EOSC, its policy milestones and its governance structure. He ends his presentation by showing some results from a survey on Open Science and Career Development conducted this year by the European Physical Society (EPS). The survey indicates that researchers are not yet really aware and concerned by the Data Management Plans. Half of the respondents claim also that the main reasons for not sharing their data are personal data protection and confidentiality. A not so surprising result!

Views from the Swiss NSF side

Matthias Egger, President of the SNSF Research Council, thanks the Swiss Academy of Sciences (SCNAT) for organizing this workshop. He also mentions the forthcoming SCNAT event on 21 November 2018 on "Beyond impact factor, h-index and university rankings: Evaluate science in more meaningful ways". Both topics are very timely and important. He then starts by showing examples from the SCNAT thematic web portals were Open Science already led to results or has a clear future role to play. He mentions the credibility crisis revealed in recent years by the difficulty or sometimes even impossibility to reproduce published results. Open Science shall aid reproducibility and thus restore the credibility of the scientists. There is a clear rise of retractions in recent years possibly due to more openness allowing to make errors or misconduct more apparent. Egger lists the potential impact of open research data as: accelerate scientific discovery, improve quality and ethical standards, foster reproducibility, speed up innovation, improve citizen engagement in research, and foster collaborative, transnational research and access in developing countries. The SNSF Open Research Data Policy states that the SNSF values research data sharing as a fundamental contribution to the impact, transparency and reproducibility of scientific research. The inclusion of a Data Management Plan (DMP) is now a formal requirement at research project submission and it shall become public at the end of the funding period. The SNSF expects published data to be shared on public repositories, but recognises issues of ethical, confidentiality, legal or technical nature preventing the partial or full release of the data. Data sets and repositories need to follow the principles of being Findable, Accessible, Interoperable, and Reusable (FAIR). The submitted DMPs already provided a census of practices. 150 different repositories were indicated, but only around 60 of them fulfil the FAIR principles and the vast majority is located outside Switzerland. Egger finishes his presentation by a list of future steps at SNSF and presents results of a survey on the barriers to sharing. The main barrier mentioned in 2010 for both code and data sharing is related to additional work: primarily to document and clean up the code and data, and also to deal with potential questions from users. While realizing the existence of these barriers, Egger concludes that he is a big fan of Open Data. The SNSF wishes to be at the service of the community to help do things right for a smooth transition to Open Science.

Views and experiences in Switzerland

Joël Mesot, Director of PSI, calls for a differentiated and careful approach of the topic. While he thinks that we can all agree on the general statements presented before, there is a great diversity of data, with some being pertinent to be released and others not. If we talk about open access to published data, this is not an issue. But open access to all data enabling an independent reproducibility of the published data would include raw data, which are 10–1000 times more voluminous. Open access to not yet analysed and published data ("live-access") is still another issue. He lists arguments in favour of access to published data in electronic format, but notes the need for international regulations. It is also likely that taxpayers may expect first a national exploitation of the published data. He mentions the highly successful example of the Protein Data Bank (PDB) with over 150’000 total depositions and over 650 million downloads per year. Providing access to raw-data for independent reproducibility has also some potential benefits, but Mesot warns of the risk of disruptive top-down regulations, which would sometimes result in enormous efforts in data preparation. He emphases the importance of a bottom-up approach. He then illustrates differences between communities: while astrophysical observations rely on absolute parameters such as time, coordinate systems, wavelength ranges, etc., data in material sciences rely on many more parameters linked to the experimental setup. He notes that Switzerland is already ranked first in the world with 39% of publications in Open Access, raising the question of the reciprocity by other countries, because there is the worry that the additional effort invested for Open Data would lead to a competitive disadvantage for Swiss researchers. Mesot concludes that while open access to published data is desirable, open access to all data for independent reproducibility is difficult to implement and can require huge efforts. He advocates to implement *suitable* measures for *suitable* data (i.e. adapted to individual research areas).

Florian Altermatt, University of Zurich and EAWAG, is leading a research group on evolutionary biology with positive experience in sharing data. The research is diverse, consisting of modelling, laboratory and outdoor experiments, as well as field studies. He lists the motivations for data sharing, noting that the external forcing by journal or funding agency requirements are secondary compared to serving an internal purpose of sharing an open mind to open access with students and postdocs and in teaching activities. Altermatt then shows examples of data he shares. Videos are their specific raw-data, but they are too big to be shared, so that the sharing is limited to the extracted information in the form of simple data tables plus metadata. His group also shares code, lab protocols and sequence data, all provided on existing databases (GitHub, GenBank, NCBI, etc.) or in supplements of papers. Since the start of the bioRxiv in 2013, there is a big increase of the number of preprints in biology provided in free access. Dryad is a generic data repository compatible with the FAIR principles, while GitHub can be used for sharing code. Altermatt sees many advantages of sharing data as they force group members to structure the datasets well, they increase the visibility of the group and outsource the long-term storage of the data. He admits however that not all data are shared by his group and identifies obstacles for data sharing. Those are either technical reasons – too big datasets, incremental data, a too high time investment compared to the benefits of sharing them – or intellectual reasons – risk of scooping, non-disclosure of sensitive, 3rd-party data and short-term protection of early career scientists.

Marcel Mayor, University of Basel, clarifies that he will express here his personal, more critical view as a synthetic chemist, and therefore will not speak as a member of the Research Council or any other academic position he has. He sees a fundamental difference between large observational datasets (astronomy, earth observation, etc.) and data of small labs. In the former the sharing of huge amounts of data that cannot be fully exploited by small teams makes a lot of sense. On the contrary, smaller local infrastructures develop individual ideas and approaches that they do not wish to share before success. At this point, the researchers are happy to share the data of this successful experiment, but not the many missed preceding attempts. This is particularly true in chemistry, because the lab equipment is rather standard, so that it is the expertise of the group that is the key to success. Submitting interesting ideas, research objectives and describing the methods in grant proposals to be distributed worldwide to experts is already perceived as risky and scary. Too much openness is therefore not in the interest of the researchers, except after a successful breakthrough. He takes as an example one of his study that lead to a publication of 8 pages with 112 pages of supporting information. In summary, Mayor thinks that there are very different situations between disciplines and research directions, such that no general regulation can apply to all. In his field, the achievements and findings are the intellectual property of the scientists, preventing the disclosure of raw data, conceptual ideas and intermediate results. The complete data set shall only be publicly available after publication. He is glad that these limitations are apparently fine for the SNSF, since his submitted DMP was accepted.

Coffee break

Robert Jones, IT Department of CERN, presents via video conference the challenges of Open Data at CERN. The four main experiments along the 27-km underground particle accelerator ring of the Large Hadron Collider (LHC) at CERN, Geneva have approved open access policies whereby fractions of their data are made available after suitable embargo periods. These refer to "derived data" with supporting documentation and software. The LHC data volume is already more than 200 PB and will eventually reach of the order of 10-100 EB with the upgrade to the High Luminosity LHC (HL-LHC). The first challenge is to preserve the data, but not everything will be open. Openness is important for the reproducibility and hence the reputation, but it needs a lot time and effort. It is also of interest for research and education. CMS – one of the four LHC experiments – has made public in November 2014 a set of 1 PB of data acquired in 2010. A couple of peer-reviewed publications from outside the CMS collaboration have been published with these data in 2017. Two new developments are the CERN Analysis Preservation (CAP), a command-line client, and the REusable ANAlysis (REANA) to run research data analyses on containerised compute clouds. CERN as a trusted digital repository with ISO 16363 certification will allow it to implement best practices ensured for the long-term. Jones mentions in this respect a very recent publication based on data from the Large Electron-Positron (LEP) collider, the predecessor of the LHC that ended 18 years ago. He notes however that the steadily increasing amount of data in the years to come is above the expected technological capacities with reasonably constant costs. There is a need for technology revolutions and steps in this direction are science clouds. CERN is a procurer of the Helix Nebula Science Cloud. Such a hybrid cloud platform brings together research organisations, data providers, publicly funded e-infrastructures, and commercial cloud service providers. CERN is also participating to the development of the future European Open Science Cloud (EOSC).

Nicolas Thomas, University of Bern, presents a talk entitled "How far dare you go?". He reports from his experience in planetary space missions, in particular the CaSSIS camera built in Bern and taking continuously (24/7) images of Mars aboard the ExoMars mission. NASA's Planetary Data System (PDS) has been in existence for more than 35 years and is still the mandatory repository for all NASA missions. The European Space Agency (ESA) has its own Planetary Science Archive (PSA), but adopted the PDS formats. The latest version 4 of the PDS is not backward compatible, the Data Provider's Handbook contains 129 pages and the data dictionary has 644 pages… This is just to illustrate how complex such standards can become over time. When submitting the CaSSIS proposal, Thomas had to sign three years before launch that he will make all data public 6 months after data taking in this PDS format and without specific ESA support to do so. In addition to the data, it is important to have the proper calibration, processing and storage of the data. PDS/PSA tries to insist that the calibration software is also delivered to allow the community to re-calibrate the data with potentially new techniques and calibration files, but this requires a huge additional drain on resources. Providing analysis software is essential – according to the high-energy astronomy group of the University of Geneva – to avoid a "data mortuary", where data are physically available, but the community cannot make a meaningful use of them. However, the building and long-term maintenance of software in long-lasting open source formats (e.g. FORTRAN, C and C++) is very time consuming. It must be seen how resources and time can be saved by also admitting the use of higher-level coding languages. All this has a clear cost in terms of resources and Thomas estimates that 15% of the current spending on CaSSIS is archive and "open data" related (~1 FTE). The support to the community to use the produced data and software is an additional cost, who pays for it? In conclusion, he sees the high value of archiving data for the community (more than 1000 publications have been generated over the years by using the PDS archive of data collected by US instruments), but doing this properly has an important cost in terms of manpower. This shall not be underestimated by the funding agencies, because underfunding the Open Data process is a recipe for exploitation of junior staff. One shall ask himself, why and to what extend one wants to preserve the data sets and whether one is prepared to pay for it.

Ana Sesartic Petrus, ETH Zurich, gives a talk entitled "Data Management Landscape from the viewpoint of climate sciences and libraries" based on her experience on research data management and digital curation at the ETHZ Library. She refers to the Nature article by Baker et al. (2016) showing results of a survey, where more than half of the researchers identify a significant credibility crisis and most others a slight crisis. Furthermore, a substantial fraction of research in general is perceived as irreproducible. Possible explanations are: the pressure to publish, a publication bias towards positive results, unavailability of methods and code, and unawareness of implicit knowledge and hidden dependencies. In parallel to this there is a strong increase of the volume of collected data, as illustrated in the field of Earth observation. An improved data management can increase reproducibility at least by the researchers themselves and – if the data are open – by others. There is therefore a demand for Open Data and data management by the higher education institutions, by funders, and also by research journals. Sesartic Petrus then presents the challenges of preservation and sharing of climate models and data. It is important to always refer to community best practices, while external requirements by the institutions can usually only be generic. She also lists the existing services for helping the researchers to make the shift to a better data management and ends with many remaining insecurities and questions, in particular related to what should be preserved to best suit the community and on the associated costs. She concludes that each community has different needs and wonders if the current guidelines take this sufficiently into account.

Global discussion on presentations

Felicitas Pauss comments that the two publications on CMS Open Data are valuable achievements enabled by Open Data, but it is challenging and it required help from people inside the CMS collaboration. Asked about the conditions to opt-out from data sharing, Egger answers that researchers (e.g. in mathematics) having no data to share, obviously do not need to share. For the others, at list the publication results are to be shared, except for special cases (e.g. patents). Asked why this discussion with the community has not been done before the implementation of the DMP, Egger notes that in the US this has been pushed already some 10 years ago, the SNSF is following the EU developments. In his opinion, the DMP is very mild and general, it is up to the communities to define best practice in their field. Rossel clarifies for the audience that Swiss researchers will be allowed to use the EOSC. It is noted that the 10 kCHF by the SNSF for additional support for Open Data is much too low in many research fields. Egger answers that this amount is the maximum that can be asked for without a specific justification. One can get more money by proper justification of the needs in the research plan. The issue of the staff for preparing and providing the data is also mentioned. It is not a work for a PhD student and needs adequate skills. Egger gives the example of EPO, where putting together all the available data allowed scientists to identify an increased mortality induced by the use of EPO in cancer treatment. He is surprised that negative results are not published in chemistry, since in medicine it is important to publish also inconclusive attempts. Somebody from the audience notes that publishers are not willing to publish negative results in fields like chemistry or molecular biology. Thomas agrees with Mayor that chemistry and space science are totally different fields. In space, not all data can be analysed by a small group and therefore opening them to the community maximizes their scientific exploitation. This is common practice since many years for NASA and ESA space missions. He repeats that because of these differences, the Open Data process shall be bottom-up. The scientists must be involved in preparing the standards for their own field.

Lunch offered by SCNAT

Repartition in four working groups

Separate 1-hour sessions on the four themes:

- WG1: Data Management Plans: Practices and challenges
- WG2: Special aspects of small-scale university labs
- WG3: Special aspects of large research infrastructures
- WG4: International competition

Report by working group convenors

Ana Sesartic Petrus reports on the outcome of WG1 on Data Management Plans: practices and challenges. Some participants regret that the SNSF did not consult enough the scientists before imposing the DMP and call for an equality between SNSF and researchers, when discussing the way forward. SNSF shall trust the scientists that they know best the specific issues of their field and shall maintain flexibility in their decisions. Researchers should see the benefit of structuring their data and act in a bottom-up approach. They should make data management an integral part of scientific education. More practically, there is an issue in choosing the right FAIR repository and also on the aging of the stored data sets, which can become obsolete pretty fast. Linked to this data life-cycle, there is the question of the quality of the data. It is not useful to have a flood of data of little or no use. Finally, there is a question on the handling of Creative Commons (CC) licences in order to prevent data-theft for economic reasons.

Heinz Gäggeler reports on the outcome of WG2 on special aspects of small-scale university labs. A short survey among the working group participants has shown 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 and the amount of data is also very diverse. Another point of discussion is how to proceed in international collaborations with partners from outside Europe. It is concluded that it should be a task for the SCNAT to define what strategy is appropriate for each discipline/field and then one shall discuss further in the research groups what it implies and how to handle this effectively.

Francesco Pepe reports on the outcome of WG3 on special aspects of large research infrastructures. The question of the audience – is it for colleagues, students or the general public? – and of the usefulness of open data for large infrastructures was discussed. It is felt that Open Data cannot be a goal by itself, but the benefits shall be balanced against the inherent costs – which can be large – of preparing truly usable data sets. In some fields (e.g. astronomy) the data of large international infrastructures are already open and freely accessible. But the data remain extremely large and complex, often limiting their use to within the community. There is also the question of the data policy inside an international collaboration sometimes preventing any disclosure of data. It is concluded that the quantity and processing level of the data to be accessible shall be discussed on a case-by-case basis, since even within a research domain it can be very diverse.

Ruth Durrer reports on the outcome of WG4 on international competition. The participants were asked to identify advantages, disadvantages and suggestions concerning Open Data. The main advantages are to: encourage international collaborations, enable research in financially weaker regions, increase visibility of own research, enhance data quality, foster innovation and technology transfer, access data from expensive/unique facilities. On the other hand, the main disadvantages are that the associated costs go away from research, that Switzerland may become a provider country doing additional work for enabling others with the risk to get scooped and a not-yet-defined acknowledgement of data usage. How to keep the positive aspects without the negative ones? The suggestions are: to include data-handling & open data in evaluations/hiring criteria and/or in university rankings, to allow for different practices in different communities, to use bottom-up non-profit repositories, to standardise data formats as far as possible within a field, and to ensure that the FAIR principles are valid for all.

Panel discussion with representatives from SERI, EU Open Science Policy Platform, swissuniversities, SNSF

Moderated by **Gerd Folkers,** President of the Swiss Science Council, the Panel is composed of **Gregor Häfliger,** Vice-director of the State Secretariat for Education, Research and Innovation (SERI), **Christophe Rossel,** EU Open Science Policy Platform/SCNAT, **Patrick Furrer,** swissuniversities, and **Angelika Kalt,** Director of SNSF.

Folkers wonders if we are not making things over complicated nowadays, since sharing data is the essence of research. He gives as one of two examples Alexander von Humboldt who made his collection publicly available. Kalt sees many differences compared to the 19th century, in particular that research is now publicly funded and the amount of data increased a lot. Häfliger adds that the world and research has changed, while Rossel notes that scientists were a very small community at that time. Nowadays, science is very multi-disciplinary and knowledge is generating a lot of business (start-ups, etc.) via innovation. Furrer thinks Open Data is important to re-appropriate science, to bring back the science debate in the hands of scientists. Haefliger acknowledges that there is now a strategy towards Open Science. It is defined what has to be done in the next years. There will be a large diversity of different practices and repositories depending on the field. The challenge is not only archiving data, but really reusing it. He notes that astrophysical data centres are pioneering this approach, but we have to discuss at which level repositories have to be build and how we deal with the costs. Rossel agrees that a coherent view on best practice needs to be found. He notes that only in physics, as many as 271 different repositories are currently used worldwide. Kalt agrees with Haefliger that the new services have to be paid. Open Data implies an increase of resources that shall not be at the expense of research. Folkers takes the example of museum curators highlighting part of their collection to give a new perspective on objects that have already been shown before. Similarly, scientists can find new viewpoints on existing data sets. He also notes that companies are generating a lot of data as well. For instance, medical imaging is done by essentially four big companies in the world. He wonders if there is a link to be found with science data. Rossel answers that social media are also generating a lot of data. Data mining consists of finding trends in these unstructured data sets, but our scientific data are of a different nature. Kalt fully agrees, we are not talking about any data, but about science data. In her view, sharing data is on request, while open data shall be open to anyone, but, indeed, one needs to define the target audience. Häfliger agrees with Kalt that the publicly funded research shall be accessible to the public. The question is how to prepare the transition in Switzerland and who will pay for the extra costs. Furrer notes that neither SNSF, nor swissuniversities are really prepared at this stage for funding long-term maintenance of data. Kalt answers that according to the submitted DMPs, most of the data are stored outside of Switzerland. We will learn over the years what the costs are. From the audience, Francesco Pepe notes that most of the induced costs are not for the repositories themselves, but are for the preparation of the data in a coherent and usable way, while Nicolas Produit wishes the SNSF to fund initiatives for data preservation coming from the scientists. Pepe also notes that if private companies store data, there is a risk that they get bankrupt and the data will get lost. Hans Rudolf Ott wonders if we would still be able to read data after 20 years. Rossel answers that there should be a free choice of the researchers to remove their data after some time. But Ott questions the benefits in view of the costs for data preservation and finds it not right that people applying for data need to justify this carefully, make the work to provide them and then people can simply get them for free without any effort. Haefliger notes that this is fine as long as it is a scientist to scientist interaction. What is needed now is to define the different needs of the various fields. Citizen Science is something else. Rossel is also cautious with Citizen Science, due to the risk of misuse or misinterpretation of data by people without the needed knowhow. Kalt sees in this respect the importance to define what Citizen Science is. For instance, it can be very useful to involve the society for getting observations of meteorite strikes over large areas. If the task is well defined by scientists it is appropriate. Someone from the audience thinks, however, that there are a lot of people who left academia and have skills to do valuable science on existing data in their free time, but Marcel Mayor doubts that people got globally smarter, even if the level of education increased…

Conclusion and next steps

Friedrich K. Thielemann keeps it very short. He concludes that the workshop has demonstrated the need of a diversity of approaches to be followed by the various communities. They shall define the best practices for the different fields of research with the help in this process of SCNAT and possibly other institutions, such as swissuniversities and the SNSF. There is also a need to involve in the dialogue other members of the Swiss Academies of Arts and Sciences, which share very similar concerns. He thanks all speakers, the working group conveners and the panel members, as well as all participants for coming and sharing ideas. The presentations will be made available on the SCNAT website in the coming days.

Christian Preiswerk retains the audience for a second to let it also thank Thielemann for organising the meeting and contacting all the speakers.