# Workshop on Open Data in Science
## Minutes

SNSF, Plenary Room, Wildhainweg 21, 3012 Bern
14th September 2015, 9:30-18:00

**Agenda**

**Four parts :**

## 1. Potentials and Challenges: Researchers‘ Perspective

**Sabina Leonelli** (University of Exeter) gives a presentation on the challenges of making data travel.

*"The enormous potential of Open Data within scientific research can be realised by understanding and supporting the specific conditions under which data can be effectively disseminated and re-used. Empirical research shows these conditions to be largely localised and field/application-specific, thus requiring decentralised policies and infrastructures. Attention also needs to be paid to conditions for inclusion in and exclusion from Open Data initiatives."*

While openness can be considered as the key norm for science, different reasons explain why it does not work in practice, e.g. long paths from data generation to discovery, increasing commercialization of the publication regime as well as practical difficulties in disseminating and reproducing data. Another difficulty lies in the fact that making data open means making data mobile and useful across sites, contexts and use in a long-term perspective.

A part of her current research investigates the diversity of data practices across/between disciplines by using different parameters of comparison such as data source, data production mode, publication cultures and ethical concerns. She also focuses on digital databases as sites for data movements (from sites of production to sites of use) such as *tair* which tries to cover the whole history of the data from data warehouse to dissemination.

Data sharing involves four main challenges. First, data sharing needs to be extensive, comprehensive, global and long-term, which requires habitual data donation, standards for data formatting, well-organized databases, sharing of related material and sustainability in time (*challenges of collection*). Second, there are misalignments between IT solutions and research questions, disagreements around data production and interpretation in pluralistic research areas as well as a low participation in the development of data infrastructures (*challenges of re-use*).

Re-using data is sometimes perceived as a risk for encouraging conservatism in research. Third, semantic ambiguity (what openness really means), publication pressure (disincentives for implementing openness in competitive fields) or the confusion about intellectual property and the role of governments in establishing open science policies (UK: OA policies in tension with measures of excellence and impact) are also problematic (*challenges of openness*). Last but not least, some researchers, institutions and laboratories suffer from systematic disadvantages (e.g. lack of infrastructures and online access, funding, government support or transport delays) and are therefore vulnerable (*Open Data Divide*).

The core message is the following: (i) data collections are most of the time partial and difficult to re-use, (ii) we need to promote discussions about what counts as data and openness in each research community, (iii) we have to consider data curation as an integral part of research and (iv) it is important to highlight the role of funders in informing researchers and policy-makers of shifting needs, resources and constraints for each field.

**Daniël Lakens** (Eindhoven University of Technology) describes science as a collaborative enterprise.

"*Science is a collaborative enterprise.*"

His approach especially focuses on behaviors of humans for sharing their data and he presents the (dis)incentives for data sharing through a game theory framework. In this context, a study reveals that most of published papers (overall, more than 80%) confirm the hypotheses tested, which illustrates that competition could lead to perverse effects in scientific progress. According to the paper from Anderson et al. (2007), novel and positive results seems to be considered more publishable than replications and negative results. Lakens also presents some statistics about reproducibility in psychological sciences and concludes that failure rates are high (between 60% and 90% according to the study). For that purpose, he proposes the concept of registered reports, a method which has for objective to introduce the peer-review process after a registered replication report to increase the credibility of published results. By reviewing and accepting preregistered proposals prior to data collection, registered reports can be an efficient way to change incentive structures for conducting replications and reporting results irrespective of their statistical significance. It is worth mentioning that the Dutch science funder NWO will start a pilot project that will exclusively fund replication studies.

He claims that scientists should share more. For instance, we have seen an increase of 35% in data sharing over the 1.5 years by just asking for it, i.e. in the form of recommendation. The NWO wants to go a step further by making data sharing a requirement for all tax funded research.

The rest of the presentation gives an overview of some possibilities for opening science. First, he presents the Open Science Framework which is an interface supporting the entire research lifecycle: planning, execution, reporting, archiving and discovery. Second, he is among the authors who launch the *Peer reviewers' openness initiative* which suggests that reviewers should make open practices a pre-condition for more comprehensive review, i.e. a minimum requirement for publication of any scientific results must be the public submission of materials used in generating those results. He also presents *Curate Science* which is a web application that aims to help for gauging the reliability and validity of published scientific findings by facilitating and incentivizing the independent verification of these findings. For each article, you find via icons if data/syntax, materials, replication studies, reproducibility info, and pre-registration info are available.

The core message is the following: (i) competition may have perverse effects on science and data sharing, (ii) failure rates for reproducibility are high, (iii) the NWO will make data sharing a requirement for funding and (iv) an increasing number of websites propose services for opening data and fostering reproducibility in the scientific community.

**Robert Terry** (WHO) presents the potentials and challenges for open data in global health research.

*"We need to create an environment for research where openness and sharing is the norm and the full potential of all the data generated is realized."*

The objective of the TDR (Special Programme for Research and Training in Tropical Diseases) is to develop improved tools for the control of tropical diseases and strengthen research capacity of affected countries themselves. One difficulty is to deal with different kinds of data, i.e. individual data (project-based data), metadata (clinical trials) or genome studies and cohort (community resource projects). Data sharing has several advantages – validation for quality, efficiency (e.g. emergencies), associations (e.g. links between disease and environment) or open innovation (e.g. drug discovery) – but some reasons may explain why it is not often the case: (i) incentives are not strong (no credit, too busy, scared of being scooped), (ii) there are capacity problems (curation, expertise, infrastructures), (iii) ethical issues (consent adequate, confidentiality) can be problematic, (iv) we face technical difficulties (How to share? How to cite? Licenses?). How can we solve these problems? (i) funders and publishers should ask for data sharing, improved citation system, including data sharing "credits" in research assessment, (ii) support for training and guidelines, reinforcing infrastructures, (iii) clear definition of secondary use, creation of oversight committees, (iv) DOI for citation, creative commons.

The Public Health Research Data Forum – led by Wellcome Trust – brings together more than 20 funders of global health research from around the world who are working in partnership to increase the availability and use of health research data in ways that are equitable (e.g. balancing the needs of the researchers who generate and use data), ethical (e.g. data sharing should protect the privacy of individuals and the dignity of communities) and efficient (e.g. any approach should improve the quality and value of research, for instance, by reducing unnecessary duplication and competition), and will accelerate improvements in public health.

The WHO recently held a consultation with different stakeholders to advance the development of global norms on data and results sharing in public health emergencies. It was recognized that epidemiologic data belong to the countries where they are generated but there was consensus that the default option is that data should be shared to ensure that the knowledge generated becomes a global public good. It was agreed that pre-publication information sharing should become the global norm in the context of public health emergencies. Researchers should take the responsibility to ensure that results – even when preliminary – are adequately robust and have undergone quality control, prior to public disclosure to enable an evidence-based dialogue with the media and communities. There was a consensus that the risks and potential harms to individuals of non-disclosure of important information provide a strong ethical rationale for rapid sharing of data.

The core message is the following: (i) research funders have an important role to play with regards to data sharing incentives, (ii) we need to find relevant solutions for a rapid sharing of data in contexts of emergency and (iii) the use of health research data should be done in ways that are equitable, ethical and efficient.

**Benedikt Fecher** (German Institute for Economic Research) proposes an overview of data sharing in academia.

*"Successful data sharing policies need to take the primary researcher's perspective into account and need to be embraced by the scientific communities. Researchers should be motivated to archive and reuse data, not forced to. Only if researchers recognize the benefits of open research data for themselves, will data sharing be practised."*

The first part of speech consists in presenting some facts and statistics about data sharing. First, he reports the results of a study in psychology which concludes that 75% of social psychology experiments failed replication test. He also mentions the case of Reinhart and Rogoff who made false analyses (data omissions, questionable methods of weighting and elementary coding errors) which influenced political and economic decisions. Overall, it seems that modern scientists are doing too much trusting and not enough verifying. He insists on the fact that data sharing culture is still poor. For instance, Tenopir et al. (2013) report that less than 6% of the researchers make their data openly available. Among the 50 journals with the highest impact, Alsheikh-Ali et al. (2011) observe that only 47% of 500 studies make their data available.

In 2014, Fecher and his colleagues led an online survey to analyze data sharing in the academia (sample of 1600 individuals with different positions and from different fields). 83% say that open research data is a major contribution to scientific progress, with some differences in terms of culture (31% of researchers in natural science say it is common to share data, 22% in social sciences and economics, 15% in medical research) and knowledge (61% of the natural scientists say they know how and where to archive data, 42% of social scientists and economists). The survey also reports that 58% of the participants have shared their data with researchers they know personally, 13% have shared publicly and 32% are generally willing to share publicly. The most important motivators for data sharing are "Data citation" and "to publish before sharing". The most important impediments are "If others can publish before me" and "archiving efforts". Interestingly, "Criticism for falsification" is the least important impediment.

The core message is the following: (i) poor data sharing culture, (ii) strategic publications considerations, low level of knowledge, efforts, (iii) requirement of delivering a data management plan in the acceptance letter (data should be mandated) and (iv) considering that the default position for research data is open science (ESRC: data have to be submitted to the UK Data Archive).

## 2. Policies and Measures: Positioning of Science Funding Agencies

**Paul Ayris** (LERU) presents the results of Science 2.0 and how the movement towards open data in science should be fostered in the EU research landscape.

*"All stakeholders in research data management need to work together to support researchers' needs for RDM. This includes advocacy, guidance, provision of technical platforms and tools, policy development, adequate funding for this provision and sustainable services. The LERU Roadmap for Research Data provides guidance on what researchers, universities, policy makers and research funding organisations can do. With growing interest in the EU Commission's Open Science agenda, now is the time for universities and research funders to act."*

Ayris started by presenting the main findings of Science 2.0, a public consultation carried out in 2014 by the European Commission, targeted to universities, research funders, research institutions, scientific libraries, academies, and publishers among others. LERU, advocating the standpoint of 21 leading research-intensive universities in the EU, was participating in the survey. The main objective of the consultation was to picture the current state of the open data movement in the EU by: (1) assessing the degree of awareness amongst the stakeholders of the changing mode of data sharing; (2) assessing the perception of the opportunities and challenges and (3) identifying possible policy implications and actions. Ayris pointed out that the participants of the consultation were aware and agreed on three new trends in science: an increased scientific production, new ways of doing science (e.g. data intensive science) and increased number of actors and stakeholders in science. The availability of digital technologies was regarded as the main driver, and concerns about quality assurance and lack of credit-giving as the main barriers for open science. The participants regarded the increase of reliability and efficiency of science as the main implications of open science. Policy interventions were widely deemed necessary for open science and participants estimated that interventions by the EU/funding agencies/institutions would speed up the implementation of open science. The most important measure to be taken according to the survey was to promote open access to publications. Further suggestions ranged from raising awareness for open science, to unblock funding for infrastructures and research programs on open science, and to provide support for data sharing, management, curation and storage.

Ayris continued by outlining the role of research institutions and science funders in the process of taking forward the open science debate. Research institutions have an important role in the open science movement and according to Ayris it is imperative that they should adopt an open science strategy. This includes research data management but also data management plans, which have an educational aspect and will engage researchers to plan data management already at early stages of their career/projects. The role of funding agencies in the open science movement is equally important. Ayris points out that a joint effort between the different funding agencies would have a significantly higher impact in that a consistent policy would facilitate the open science movement in the EU research area. Research funders could implement guidelines for data management plans and define research areas where the default situation for data is open. Furthermore, they can require research data management policies from institutions and elaborate plans to fund infrastructure.

Research funders are in a position to lead the movement, like for the open access debate, but a global interchange of ideas and measures is seen as key to successfully install a continuous change towards open data in science.

The core message is the following: (i) stakeholders in the EU research area agree that policy measures are needed to foster the open science movement, (ii) research institutions should adopt an open science strategy and (iii) research funders should lead the movement but should take actions collectively.

**Daniel Mietchen** (NIH) presents the vision for data at the NIH.

*"Funder policies should be accompanied by appropriate funding, supported by technical infrastructure and reduce red tape, not add to it. For instance, if data management plans were machine readable and public, they could serve as a discovery tool, and policy compliance could be monitored automatically."*

The National Institute of Health constitutes of 27 institutes and centers. Open data has been identified as a major challenge for the NIH and as a response the Associate Director for Data Science Office (ADDS) was founded in 2014. The primary responsibility of the ADDS is to develop an open science strategy and lead data science activities and funding initiatives. The long-term vision of the NIH is (1) to create a common storage space for data where researches can access and work on the data without actually download the data (“the commons - compute platform”), (2) to identify a citation system where credit is given to scientists for making reliable, high quality data sets available and (3) to ensure the highest level of discoverability.

The NIH already installed open data policies, starting first with recommendations and later making it voluntary for researchers to allow public access to their data. After gaining experience, the NIH (1) identified the need to elaborate different data sharing policies for different scientific fields (genomic data vs. environmental data, etc.), (2) deemed it necessary to make data sharing mandatory for NIH grantees, (3) to enforce a data sharing plan which is machine readable and (4) to require repositories to include grant numbers. A key to success of open science is a valid data citation system, with the goal to legitimize data as a form of scholarship. The NIH is working towards this goal by providing machine readable standards for data citations and the endorsement of data citation for inclusion in NIH grants, bib sketches, reports etc.

The NIH urges other funding agencies to start with open data pilot phases in order to be in a better position for the installment of general open data policies. In general the NIH suggests the immediate sharing of data as the default and not to allow for embargoes but define the areas of exceptions. Machine readable data management plans were an important measure to implement the NIH open data policies and science funders should explore the possibility to adapt such plans as part of their actions towards open science. When outlining open data policies science funders should pay attention that their actions are coupled with technical means facilitating their implementation and that the focus is on empowering the researchers, rather than regulating them. Furthermore, “data” should be viewed in a broad sense and not necessarily only encompass e.g. digital machine readable data.

NIH regards research funders as key players in the open science movement but identifies a collective action problem. Open science can only be truly achieved by international collaborations between funding agencies. The NIH recommends science funders to lead by example and be more "open" themselves and explore possibilities to increase transparency.

The core message is the following: (i) open data policies should be adapted to different research areas, (ii) machine readable data management plans are an instrumental part of open data policies at the NIH (iii) the default of data should be open and (iv) funding agencies should lead by example by being more transparent.

**Roar Skalin** (RCN) presents the experience of the Research Council of Norway with its policy on Open Access to Research Data.

"*The scientist is the key to succeed with sharing and reusing research data.*"

In 2013 the RCN received a mandate from the government to implement an open data policy. As a first step, the RCN launched a survey to identify the needs/concerns/practices of scientists regarding data management and data sharing. Researchers are open to data sharing but identified three main challenges: (1) lack of time, (2) lack of infrastructures, and (3) reduction of the chances of future publications. The researchers considered (1) better infrastructures, (2) a citation system for data sets and (3) guidelines, training and standards as key for data sharing.

The RCN proceeded to implement open data policies as recommendations first. This decision was based on the current lack of infrastructures for data storage and an accepted data citation system, and on the idea to strengthen the positive attitude of scientists towards data sharing, rather the install another requirement. Similar to the NIH, the RCN recommends to make data open by default and to define areas of exception (e.g. confidentiality obligations, protection of personal data, etc.), while on the other hand the RCN allows for embargo periods for data sharing. Furthermore, the RCN started pilot experiments and introduced mandatory data management plans (DMP's) for certain funding instruments and calls. They identified two important obstacles: (1) a lack of knowledge in the scientific community about writing DMP's and (2) a lack of knowledge among the experts about evaluating DMP's. The RCN plans to modify submission procedures concerning DMP's, requiring researchers to submit a detailed DMP only in case their project is funded at which point they can get support and guidelines. As a consequence, DMP's are not yet part of the evaluation process, reducing drastically their impact and importance. The RCN considers further measures to implement their open data policy by (1) making DMP's mandatory for all relevant applications, (2) providing funding related to archiving and publishing open data, (3) fund research infrastructures that facilitate open access to publicly-funded research data and (4) establish a platform for dialog and learning between universities and research institutes.

The core message is the following: (i) researchers are willing to share data, but measures have to be taken to engage them since they are the gatekeepers, (ii) data management plans are important but scientific community might not be ready yet and training is needed, and (iii) the default of data should be open.

**Stefan Winkler-Nees** (DFG) presents the strategies and funding approaches of the German Research Foundation concerning Research Data Management.

*"Free and open access to research data has an enormous potential for new scientific findings in any discipline. It requires amongst some other things specific infrastructure, information specialists and the willingness to provide funds. However, the true key to a successful digital science environment is the awareness of its potential combined with the willingness to share and to reuse data of those, who produce and work with them – the scientists."*

What are research data? Before implementing a policy for research data management, the DFG attempted a definition: "*…research data are digital and machine readable data, which were generated by a scientific research endeavor e.g. through the study of sources, experiments, measurements, data collections, surveys and interviews. …*". Winkler-Nees states that such a definition is necessary, but stresses the fact that it is also the first crux in the process of outlining an open data policy. The definition already reduces the focus and potentially excludes a lot of data, which are considered by many as "research data" (e.g. software, samples, objects, etc.).

The vision of the German Science Organisation is to achieve the goal where research data are freely accessible, easy to obtain and professionally curated on a long-term basis. The DFG will take measures to achieve this goal but stresses that a successful move towards open science is only possible if other stakeholders get involved as well. Publishers need to integrate data sets more into publications, in order to make data sharing an integral part of data publishing. From the point of view of the information services (infrastructure view) actions should be taken to reduce or eliminate the differentiation between data "management" and scientific analysis. According to Winkler-Nees, repositories have been often to a large extent an IT project and the needs and perspectives from the research side, which are instrumental to ensure long-term success and re-usability of data, have not been taken into account enough. The DFG is taking action to work towards an academic culture change among scientists. They see it imperative to introduce a system where formal recognition for data sharing becomes the standard, which would lead to a faster adoption of open data practices. To achieve this goal the DFG started to offer support and incentives to researchers (information specialists, funding for research data management, etc.). Furthermore, the DFG wants to include scientists more into the policy making process. Scientists should contribute to defining requirements and policies for their specific disciplines, to identify best practice examples and data management plans. Another strategy of the DFG to tackle the challenge of open data is to initiate pilot and exploration projects on a smaller scale. Furthermore, the DFG launched calls focusing on open data/repositories/re-usability (e.g. call "Information Infrastructures for Research Data" in 2010 and 2013, call "Research data in practice" 2015), which they consider an optimal way to initiate collaborations and idea exchange between researchers and infrastructure services and improve data storage infrastructures.

The core message is the following: (i) establishing a system which gives formal recognition for data sharing is a necessity, (ii) involvement of researchers into the process of policy making is important, and (iii) funding agencies should launch pilot phases or call focusing on open science/re-usability/research data management

## 3. Solutions: Effective Data Storage and Curation

**Tim Smith** (CERN) presents the challenges of storing, managing and opening data in the "big data" field of particle physics.

*"In the digital era it is no longer possible to meaningfully summarize and transmit scientific methods or results on paper alone. It is necessary to share the digital artefacts for the machinery of science to turn correctly. But sharing alone isn't sufficient, the data and software must be described and prepared for reuse. Otherwise we will continue to slip slowly into the digital dark age."*

Experiments at the LHC (Large Hadron Collider) at CERN typically generate petabytes of information per second. However, most of the events produced at a collider are not interesting. Using HLT (High Level Trigger) algorithms and filters allows to reduce the produced data to gigabytes/s.
The Worldwide LHC Computing grid project is a global collaboration of more than 170 computing centers in 42 countries, linking up national and international grid infrastructures. It aims at providing global computing resources to store, distribute and analyze the gigantic amount of data annually generated by the LHC.

Storing 120 Petabytes is a good challenge but it is a static task. The dynamic task of analyzing it is a much bigger challenge. Analyzing means transformation, reduction, transport, replication and regeneration of data. Managing a large data archive also means dealing with storage media verification and migration (due to obsolescence) in order to keep the data accessible.

The CERN sees open data as a service and has put forward data access policies. The CERN Virtual Machine is a tool that allow to "hunt for dark matter on your sofa". More largely, the CERN is confronted to the question of how to transform existing distributed computing infrastructures into a service-oriented platform for research and educational purposes.

Only a small number of scientists have access to big storing facilities. To address the data challenge for the "long tail of science" (the large number of scientists that do not have access to big in-house computing resources), the CERN, together with other intergovernmental research organizations, has created the community concept of Zenodo.org, a platform that aims at answering the very different needs of the various research communities.

The core message is the following: (i) CERN and the LHC program have been among the first to address "big data" challenges. Solutions have been developed and important results obtained (ii) CERN is preparing for future needs in common with many scientific domains (iii) Zenodo is an open data platform for the "long-tail" of science, enabling outputs in any size, any format and from any science.

**Emma Ganley** (PLOS biology, Cambridge) gives a talk on the lessons learned from PLOS on open data policies.

*"Although not without challenges, aligned efforts from stakeholders who back Open Data should be able to provide the support (infrastructure, guidelines, etc.) required to set expectations and result in successful new global standards and policy for data availability."*

PLOS is a non-profit publisher that aims at making scientific information available to a much wider audience, including millions of potential readers that do not have access to journal publications which one must pay for.

Emma Ganley states that data availability declines over time: 10 to 15 years after publication, almost all data are lost, which means that these data are not any more available for replication, reuse, reanalysis or reproducibility check.
When publishing in PLOS, authors must provide a data availability statement (DAS) that describes the compliance with PLOS' policy. *PLOS requires authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception.* Before implementation of this policy, many questions were discussed like: What to do with massive datasets, what if the researcher plans to publish additional studies using the data, what if competitors take advantage, etc.

The policy of PLOS requires that not the entire datasets are available, but the specific data useful to support the conclusions. It also define exceptions (ethical or legal, for example patient privacy) and data-sharing methods (data must include DOIs and be in a format that can be extracted). Unacceptable restrictions are for example the refusal to share due to personal reasons, such as patents or future publications. Some of the issues encountered by the editors are: How much data checking should the editors / reviewers do? Which data are required? Which repositories? But also problems like un-extractable data, proprietary file-types or lying authors.

Emma Ganley then addresses the challenge of big data in field such as genetics and genomics. We still do not know how to handle such large amounts of data. Practical solutions are needed at the institute level too. Other questions we do not know the answer yet, among others, are: How long should researchers store data, how to preserve obsolete formats, how to cite the use and reuse of data, how to handle restrictions in particular area (human and clinical data, MRI images, government data, etc.).

An important finding is that just by requesting a DAS, a massive increase in data sharing has been observed. However, considering the large amount of papers published in PLOS, a manual checking of data availability is very time-consuming and complicated.

The NSF funds the project "Make Data Count" which aims at exploring and testing data-level metrics. In this context, PLOS co-organized a meeting, emitting recommendations for publishers to increase access to data.

The core message is the following: (i) The implementation of a data policy is a good incentive for the researchers but it still contains many issues and unknowns (ii) Just by requesting a data availability statement (DAS), there was a massive increase in data sharing (iii) The road to data access requires the coordination of the whole community (researchers, funders, publishers, institutions, policy-makers, data centers).

**Richard Mount** (SLAC, Stanford University) makes a presentation on preserving and opening access to data from a particle physics viewpoint.

*"From the viewpoint of a "big data" scientist the key goal of Open Data is to enhance overall scientific productivity of the nation and the world within available financial resources. Preserving the ability to analyze data beyond the tenure of a graduate student is challenging and costly but*

*necessary. Broadening meaningful access to data both in time, and in the size of the target audience, is desirable in principle but must pass cost-effectiveness scrutiny. The costs ascend in the order: curation and communication of bits; curation and communication of metadata; curation and communication of human knowledge and process."*

Data preservation in long running experiments is a long term task: Data analysis will continue at least 10 years beyond the end of data taking. There are two approaches possible: to freeze the software to be used indefinitely or ensure that current versions of the software can process all old data.

Open data can be a solution: Long term data access can ensure the ability to support analysis of the data. It is important to provide a stable environment, use open formats and store data on adequate medium. For the BaBar experiment, funding for data preservation equipment has been provided but there is no funding for supporting open data. In the frame of the LHC experiments, policy on open data is the responsibility of the collaborations. Generally, the release of data occurs after the end of the intense analysis by the collaboration (3 to 5 years). There is no benefit by releasing unintelligible, unsupported data. In addition, for serious analysis, simulated data is required. This means that open data requires open software.

Funding agencies are keen to see open data release, but not always to provide funding for it. The CMS experiment is a pioneer in open data release: In 2014, it has released a large set of reconstructed data for public use. The feedbacks were very positive and the high quality of data was appreciated.

In particle physics, use of open-access journals is becoming the norm. Particle physicists make wide use of bibliographic databases and electronic platforms.

In conclusion, the main challenge is to extract knowledge from data. This requires a large amount of effort, time and computing resources. The following question remains open: Can we make (at reasonable costs!) data sets from 2015 fully available to scientists and the public in 2115?

The core message is the following: (i) Data preservation in long running particle physics experiments is a complex and long term task, but open data might be a solution (ii) In the field of particle physics, the main challenge is to extract knowledge from data. This means that open data requires open software (iii) Funding is a limiting factor: Long term preserving and opening access to data for scientists and the public requires a lot of effort and resources.

**Micha Rieser** (Wikimedia, University Library of Basel) gives a presentation on the use of open access and open data in Wikipedia and partner projects.

*"Free knowledge is a benefit for all. Open Data, Open Access and Open Licences are the key factors for the liberation of knowledge."*

The concepts of open access, open data and open content have common features. However, definitions are needed, in order to know what we are talking about. For Wikipedia, "free" means freedom of reading, copying, modifying and redistributing. Wikipedia is a collaboratively edited, multilingual, free access and free content encyclopedia.

M. Rieser gives some examples of his own contributions to Wikipedia in the field of spiders. A large amount of work remains to be done. In this specific field, there are only 300 articles about spiders, while about 900 new species are described per year. Creative commons (CC) licenses define the conditions under which works of others can be used and distributed. There are 6 types of CC licenses, going from least to most open, comprising a selection out of four conditions:

- Attribution: The work may be copied, distributed, performed and modified if proper credit is given to the authors.
- Share-alike: Derivative works may be distributed only under a license identical to the license of the original work.
- Non-commercial: The work may be copied, distributed, performed and modified only for non-commercial purposes.
- No derivative works: The work may be copied, distributed, performed but not modified.

Wikipedia requires free availability on the public internet, without any barriers. The only constraint should be to give authors control over the integrity of their work and the right to be properly acknowledged.

The Wikimedia foundation has an open access policy, which covers not only publications, but associated data, software and multimedia. Being itself available under an open license, it stresses the importance of open licensing, which facilitates and broaden the scope of reuse. Furthermore, it avoids embargo periods and allows only for limited exceptions. The exceptions are to be documented in public, helping later refinements of the policy.

The core message is the following: (i) Wikipedia is a collaboratively edited, multilingual, free access and free content encyclopedia. For Wikipedia, free means freedom of reading, copying, modifying and redistributing (ii) Creative commons licenses define the conditions under which works of others can be used and distributed (iii) The Wikimedia foundation has very high standards of openness: It recommends the free availability of all kind of contents (publications, data, software, media, etc.) without financial, legal or technical barriers. The only constraint should be to give authors control over the integrity of their work and the right to be properly acknowledged.

## 4. Roundtable discussion

The discussion of the roundtable focuses on the following points/questions:

1) Barriers to open data?
2) Is the requirement of delivering a research data management plan in the submission procedure relevant?
3) Which kind of policy for the SNSF?

Conservatism and risk aversion in the research community as well as competition between scientists are considered as one of the main barrier to the movement towards open science. Moreover, the culture about data sharing is still poor (nevertheless, with some differences between the disciplines). Overall, data sharing is time-intensive, unclear about accepted standards and suffers from a lack of knowledge and training among the stakeholders involved in the data lifecycle. It is also mentioned that the project-based funding is not sufficient to deal with this data storage, preservation and curation (e.g. data in the field of humanities – which are generally complex (image, films, pictures) – have to be stored and curated for decades and even more).

Despite these obstacles, the open data movement is gaining ground. In particular, young scientists are realizing themselves that data sharing is an added value and are aware of the trend towards open science. Scientists coming from disciplines which are deeply data-intensive are also more used to share their data. We have to give more value to research data to foster data sharing. One possibility could consist in establishing a formal recognition system for data sharing through, for instance, data citations. Another possibility would be to fund researchers for research data management activities and to provide training and support. Finally, long-term funding is needed for developing appropriate infrastructures which are currently insufficient.

Several persons agree on asking for research data management plans (DMPs) in the applications. However, the main issue with DMPs is their objective, i.e. researchers should not perceive these plans as an additional administrative workload, without further use. More value should be attributed to DMPs. For that purpose, they should be included in the evaluation procedure, more precisely in the peer-review process. This will motivate the researchers to write them in a good, effective and efficient way. The DFG, NIH and RCN have already implemented such practices. The NWO is the next on the list. It is nevertheless worth mentioning that the preparation and evaluation of DMPs require data management training.

In terms of policy implication, it is pointed out that the SNSF has a good position to develop and propose new business models regarding open science. Overall, it is agreed that funding agencies have to lead the open science movement. The SNSF is a small agency, without concurrence, and can have a quick and important impact on the scientific community ("The scientific community will praise you"). However, the SNSF cannot work alone. A cooperation with other stakeholders is necessary. A diversified policy is encouraged to account for disciplines' specific needs. It is also important to include the scientists' perspectives in the policy implementation. For example, a proposition for developing a pilot project by the SNSF Division "Careers" in which the young researchers could express their opinions about open data is also mentioned. Last but not least, it is agreed that data should be open by default with few exceptions.

As concluding remarks, Martin Vetterli says that (i) a strong cooperation between stakeholders is necessary; (ii) a pilot project by including DMPs is a possibility; (iii) we need to consult the external reviewers for advices regarding the implementation of data policies; (iv) finding data scientists is not an easy task and (v) open data has the same risks as open access publications, i.e. the risk to remunerate the private industry (in the present case for realizing the DMPs and offering data repositories).

*Minutes have been written by Martin von Arx, Cornelia Sommer and Lionel Perini.*

## Annexe 1: Programme

**Annexe 1**

## PROGRAMME

9:30 Registration and Coffee

10:00 Welcome Address and Opening Remarks **Martin Vetterli**

### Part 1 Potentials and Challenges: Researchers’ Perspective
**Dominique Soldati-Favre**

10:10 **Sabina Leonelli** “The Challenges of Making Data Travel”
Department of Sociology, Philosophy and Anthropology, University of Exeter, UK

10:40 **Daniel Lakens** “Science as a Collaborative Enterprise”
Human-Technology Interaction Group, Eindhoven University of Technology, The Netherlands

11:00 **Robert Terry** “Potentials and Challenges: Researchers’ Perspective – Health”
TDR – The Special Programme for Research and Training in Tropical Diseases, WHO, Geneva, Switzerland

11:20 **Benedikt Fecher** “Data Sharing in Academia: The Researchers’ Perspective”
Internet-enabled Innovation Department, Alexander von Humboldt Institute for Internet and Society, Berlin, Germany

### Part 2 Policies and Measures: Positioning of Science Funding Agencies
**Robert Terry**

11:45 **Paul Ayris** “ Open Science, Research Data Management and Research Funding Agencies”
Library Services and Copyright, University College London, UK

12:15 **Daniel Mietchen** “The Vision for Data at NIH”
National Institutes of Health, Bethesda, Maryland USA

12:35 **Roar Skalin** “The Research Council’s Policy on Open Access to Research Data”
Department of Research Infrastructure, The Research Council of Norway, Norway

12:55 **Stefan Winkler-Nees** “Research Data Management - Strategies and Funding Approaches of the German Research Foundation”
Scientific Library Services and Information Systems, Deutsche Forschungsgemeinschaft DFG, Berlin, Germany

13:15 Lunch

## Part 3 Solutions: Effective Data Storage and Curation
**Antonio Ereditato**

14:30 **Tim Smith** "One Size Fits All?"
Collaboration and Information Services, IT Departement, CERN, Switzerland

15:00 **Emma Ganley** “Open Data Policies; Lessons Learned from PLoS”
PLoS Biology, Cambridge, UK

15:20 **Richard Mount** “Preserving and Opening Access to Data: Challenges and SOLUTIONS - from a Particle Physics Viewpoint”
SLAC National Accelerator Laboratory, Stanford University, USA

15:40 **Micha Rieser** “Use of Open Access and Open Data in Wikipedia and Partner Projects”
Wikimedia, University Library of Basel, Switzerland

## Part 4 Discussion and Conclusions
**Martin Vetterli**

16:00 **Round Table**
Paul Ayris, Huldrych Günthard, Brian Kleiner, Sabina Leonelli,
Christian Leumann and Tim Smith

16:30 **General Discussion**

17:00 **Conclusions**

**Aperitif**