⟨S|P|S⟩ swiss physical society

Nr. 57
März 2019

SPG MITTEILUNGEN
COMMUNICATIONS DE LA SSP

AUSZUG - EXTRAIT

Physics and Society

Let's Talk About Open Science

Martin Vetterli, President of the EPFL, presidence@epfl.ch,
Luc Henry, Scientific Advisor at the Presidency, EPFL, luc.henry@epfl.ch

Physics and Society

Let's Talk About Open Science

Martin Vetterli, President of the EPFL, presidence@epfl.ch,
Luc Henry, Scientific Advisor at the Presidency, EPFL, luc.henry@epfl.ch

This article is the written version of the public lecture given by Martin Vetterli at the SPS Annual Meeting 2018 in Lausanne.

In 1989, English telecommunication engineer and CERN fellow Tim Berners Lee wrote a 'vague, but exciting' proposal to build a hypertext project called "WorldWideWeb" [1]. The rest is history. As of June 2018, an estimated 55% of the 7.5 billion people living on this planet have access to the Internet via the Web. It has become the backbone of the Information Age we all live in, the cornerstone of our modern digital culture and of its communication revolution.

What is well known is that a CERN scientist invented the Web. What is lesser known is that it was initially invented *for* scientists. More precisely, the goal was "*to allow high-energy physicists to share data, news, and documentation*" as Tim Berners Lee explained on an internet forum in August 1991. Anticipating the enormous potential of the Web technology, he insisted that it should not be restricted to his discipline. "*The Project started with the philosophy that much academic information should be freely available to anyone. It aims to allow information sharing within internationally dispersed teams*", he wrote in another memo addressed to the early adopters and supporters of the Web.

Open Science and Digitalization

Open science is based on the assumption that the discovery process should be carefully described and ultimately shared so that everyone can critically assess the results and build on them or make new discoveries [2]. It can be debated whether the Web is the cause or the consequence of the open science movement we witness today. What is clear is that it was conceived as a tool to support global collaboration in science. It has dramatically changed, and is still continuously changing, the way we discover, produce and share knowledge.

Beyond the Web, it is digitalisation in general that has triggered a transformation in research and education. The academic institution is in a transition, much like the rest of society. Computers are used in virtually every step of the research process, from generating a hypothesis to disseminating results. Scientists use search engines to find relevant documents in the ever-expanding volume of literature, store their observations, measurements and models in digital files on servers, craft algorithms to sort, analyse and visualise the data, and use the web again to share all of this with colleagues across the globe. Some compare the impact of computers and networks to what happened during the first scientific revolution when the printing press was invented. One can wonder: is digitalisation triggering another scientific revolution?

What is surprising is that, although our means of communication have evolved so dramatically, the format used for the dissemination of scientific knowledge within the academic community has remained relatively unchanged. In 1665, German theologian and natural philosopher Henry Oldenburg joined forces with Robert Boyle, considered the first modern chemist and one of the pioneers of the experimental scientific method, to launch the *Philosophical Transactions* of the Royal Society. It was the first journal in the world exclusively devoted to science and the first to adopt peer review. The publication process was put in place to fulfill five main roles: documentation (detailed description of the experiment and the results), registration (list of authors and date stamp), certification (through peer review), dissemination (via subscriptions) and archiving (in libraries). With the advent of scientific journals, the reputation economy of academic scholarship was born.

Some voices say that the format adopted by the *Philosophical Transactions* in the 17th century is out-dated today. The complexity of contemporary research methodologies and results cannot simply be described in words, tables and static figures anymore. The problem started to appear over two decades ago with the advent of computers in research, leading to the apparent expansion of the scientific method [3]. Inspired by Stanford professor of geophysics Jon F. Claerbout [4], James Buckheit and David L. Donoho expressed their concerns in the following statement: "*An article about computational result is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result.*" [5] [6] The rationale is that, in order for others to verify and build upon their work, researchers have to share more than just a wordily description of their methods. The documentation required to guarantee an acceptable level of reproducibility in chemistry and other natural sciences was not sufficient in an era when scientists increasingly used computer models and large data sets to make discoveries and prove hypotheses. While errors are a crucial component of the scientific enterprise, only transparency allows others to correct them.

After over 350 years of going virtually unchanged, it may be time for the 'paper' format to be rethought. Its limitations, exacerbated by the commodification of knowledge that big publishers have put in place through paywalls as they consolidated their catalogue of journals, start to transpire in every field of research [7]. The foundation of open access were laid over two decades ago, but today, open science supporters promote public access to the data, the computer code, algorithms, and other digital products that support the results described in research articles. They seek to solve the problem of under-reporting or misinterpretation of results that led to a perceived reproducibility crisis [8]. Or they want to tackle insufficient skills and infrastructures that would allow coping with the ever-increasing importance of computers and digital data in research (see for example the teaching programs by the community initiative *The Carpentries* [9]). As a consequence, new practices emerge in the research community, together with new expectations from journals and funding agencies, but also from the public.

By going back to Tim Berners Lee's original vision, open science initiatives are bringing back digital innovation to academia and changing the way researchers make and share discoveries. They use the Web to its full potential and explore new strategies allowing 21st century science to thrive.

The Promises (and Challenges) of Open Science

A number of influential scientific institutions have released reports that identify the potential impact of open science practices. As early as 2012, the Royal Society has laid out the vision, and the necessary steps to enable its realisation [10]. More recently, the American National Academies of Sciences, Engineering and Medicine systematically mapped the benefits and motivations, as well as the barriers and limitations of an implementation of open science principles [11]. The authors of the report put the researcher at the centre of the concept of open science by design, expecting them to both contribute and take advantage of the open science practices. Their assumption is that research conducted openly and transparently leads to better science. Indeed, open science is a vow to invest into best practice that will eventually lead to the following outcomes:

Visibility = larger impact
Scrutiny = better quality
Reuse = higher efficacy
Public access = fair opportunity

Physics has been at the forefront of open science in many ways. One salient example is the widespread adoption of digital preprints as soon as Paul Ginsberg launched the online repository *arXiv.org* in 1991. In the spirit of the Web original purpose, sharing a digital version of research articles before their formal publication has rapidly become the norm in this discipline. What may come as a surprise to contemporary researchers is that correspondence between scientists to collect feedback on one's work before formal publication has always existed. In fact, following the steps of the physics community, biologists also attempted to adopt preprints mailing lists in the 1960s although in did not succeed at the time [12]. Their recent growing popularity, in particular in biology [13] and chemistry [14], may be in part related to an increase in the number of interdisciplinary projects over the past decades. Researchers who have been trained as physicists or mathematicians started contributing to research in genetics and genomics, systems biology, biophysics, molecular simulations and other subfields. In this important case of cross-pollination between disciplines, biology preprints were initially submitted to a sub-section of arXiv – called *q-bio* (*arxiv.org/archive/q-bio*) for "quantitative biology" – before the creation of *www.bioRxiv.org*.

Beyond publications, the potential of data sharing has recently gained the attention of the field of particle physics. In a key move, the Compact Muon Solenoid (CMS) experiment, one of two large detectors built on the Large Hadron Collider (LHC) at CERN, made three data releases in November 2014, in April 2016 and in December 2017. Anyone can now explore the data with their own point-of-view, bringing diversity in interests and skills. This allowed MIT researchers to make new discoveries, independently from the team based in Geneva [15].

It is important to realize that allowing others to reuse information is only one of the two sides to the reproducibility coin. This main motivation to produce reproducible results rests on the idea of "Nanos gigantum humeris insidentes" (dwarfs standing on the shoulders of giants) to use a popular metaphor attributed to 12th century French philosopher Bernard of Chartres. But the ability to hand over a methodology from one person to another in a timely and efficient manner enables project integrity as personnel changes over time. In a very pragmatic move, CERN needs to make sure the data it produces will not become obsolete. As pointed out, "*data and the knowledge needed to interpret them are more likely to survive in the long term if many people outside an experiment are constantly trying to make sense of them.*" [16] But open is not enough [17]. Significant efforts must be invested in careful documentation, the development of standards as well as classification methods. In fields that lack such best practice, a substantial investment and a cultural change are required, demanding efforts from every stakeholders of the research ecosystem (see Figure 1).

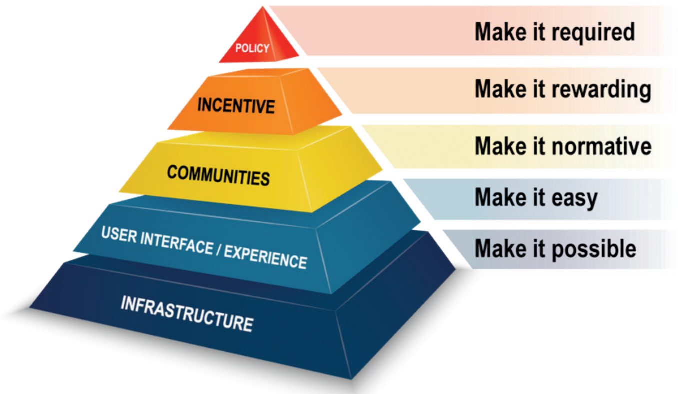

Fig 1. Support necessary for the cultural change leading to open science (adapted from original work by Brian Nosek, Center for Open Science [18])

But even within the physics community, there are different cultures. On the one hand, researchers using facilities – the Hubble telescope or a synchrotron beam line - compete for instrument time and collect data they often can keep under embargo. On the other hand, participants to an experiment – the Sloan Digital Sky Survey [19] or the Gaia space observatory [20] for example – participate to the construction of an infrastructure that will release data openly to all. In this case, it is the originality of their analyses that will be advantageous. For theorists or experimentalists having their own infrastructure, there is no particular incentive to participate in a collective endeavor. In each of these groups, the relationship between the researchers and the data is therefore different and collaboration between these cultures can lead to friction. We can ask ourselves why astrophysics and particle physics have been amongst the most open research disciplines. One possible answer is that there is only one sky, and there is also only one Large Hadron Collider. In

"Big Science", researchers have to mutualize resources and share outputs amongst consortia. As more and more disciplines adopt a team science strategy, openness becomes increasingly justified.

Open Science at EPFL

Because one of the missions of EPFL is to perform world-class research and disseminate the resulting knowledge and technologies as broadly as possible, there is a strong interest in exploring the potential of open science. In a world where science is increasingly international, and where interdisciplinarity and collaboration have demonstrated significant impact, our researchers have a strong incentive to share information in order to maximize their impact on the research of others, and on society at large. To some, open science may seem like a nebulous term, with philosophical, rather than practical, implications. While some initiatives explore practices that have not yet gathered consensus, others have now demonstrated value in supporting high quality research results. As research communities adopt new standards, we need to support our researchers when adopting these practices so that they stay competitive.

Since October 2017, a strategic committee is exploring the topic and how to support researchers best. It is composed of faculty members representative of the many disciplines present on EPFL campus, as well as central services providing support with issues related to open science. This committee has now made a series of recommendations on how to best leverage existing central services and resources to foster change, including a need for communication, training and new types of infrastructures.

Based on these recommendations, the EPFL Presidency has decided to dedicate resources in the form of an Open Science Fund [21]. This competitive funding will support projects that explicitly aim at increasing the use of tools and best practice in research management, as well as lowering the technical, social, and cultural barriers that make it hard for researchers to openly share some of the research outputs with their colleagues, and ultimately with anyone outside of the academic community. The fund will support projects leading to the development and bottom-up adoption of innovative ways of making research robust, accessible and reusable at any stage of the research life cycle. They should facilitate the curation and dissemination of valuable and original research outputs in all possible forms: scholarly publications, data files of all kinds, the associated metadata, software and hardware, experimental setups, methods and instruments, etc. with the hope that researchers will earn recognition for these outputs, in complement to their publication record.

In their 1992 article, Jon Claerbout and Martin Karrenbach wrote – with a certain optimism – that "*we are nearing a time when it will simply be the author's choice whether to keep detailed means to results confidential with the use of traditional publication or to communicate fully by using reproducible documents.*" [2] This choice is still that of the researchers, but as the technological barriers have constantly been lowered since the invention of the Web thirty years ago, it is has become difficult to justify not choosing the open and reproducible route.

References

[1] Tim Berners-Lee (1989): Research Management: A Proposal, CERN (*http://info.cern.ch/Proposal.html*)
[2] Mick Watson (2015): When will 'open science' become simply 'science'?, Genome Biology (*https://doi.org/10.1186/s13059-015-0669-2*)
[3] Tony Hey, Stewart Tansley, Kristin Tolle (2009): The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research (*https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/*)
[4] Jon Claerbout and Martin Karrenbach (1992): Electronic documents give reproducible research a new meaning, in Proc. 62nd Ann. Int. Meeting of the Soc. of Exploration Geophysics, pp. 601–604 (*http://sepwww.stanford.edu/doku.php?id=sep:research:reproducible:seg92*)
[5] J. B. Buckheit, D. L. Donoho (1995): WaveLab and Reproducible Research. In: Antoniadis A., Oppenheim G. (eds) Wavelets and Statistics. Lecture Notes in Statistics, vol 103. Springer, New York, NY (*https://doi.org/10.1007/978-1-4612-2544-7_5*)
[6] David L. Donoho (2010): An invitation to reproducible computational research, Biostatistics, Volume 11, Issue 3, Pages 385–388 (*https://doi.org/10.1093/biostatistics/kxq028*)
[7] Richard Van Noorden (2013): Open access: The true cost of science publishing, Nature (*https://doi.org/10.1038/495426a*)
[8] Monya Baker (2016): 1,500 scientists lift the lid on reproducibility, Nature (*https://doi.org/10.1038/533452a*)
[9] *carpentries.org*
[10] Royal Society (2012): Science as an open enterprise (*https://royalsociety.org/topics-policy/projects/science-public-enterprise/report/*)
[11] National Academies of Sciences, Engineering and Medicine (2018): Open Science By Design: Realizing a Vision for 21st Century Research (*http://sites.nationalacademies.org/pga/brdi/open_science_enterprise/*)
[12] Matthew Cobb (2017): The prehistory of biology preprints: A forgotten experiment from the 1960s, PLOS Biology (*https://doi.org/10.1371/journal.pbio.2003995*)
[13] Kaiser (2017): Are preprints the future of biology? A survival guide for scientists, Science (*https://doi.org/10.1126/science.aaq0747*)
[14] Demma Carà, Ciriminna, and Pagliaro (2017): Has the Time Come for Preprints in Chemistry?, ACS Omega (*https://doi.org/10.1021/acsomega.7b01190*)
[15] Larkoski et al. (2017): Exposing the QCD Splitting Function with CMS Open Data, Phys. Rev. Lett. 119, 132003 (*https://doi.org/10.1103/PhysRevLett.119.132003*)
[16] Elizabeth Gibney (2013): LHC plans for open data future, Nature (*https://doi.org/10.1038/503447a*)
[17] Chen et al. (2018): Open is not enough, Nature Physics (*https://doi.org/10.1038/s41567-018-0342-2*)
[18] Brian Nosek (2018): Societies' role in improving openness and reproducibility of research (*https://osf.io/r9v3p/*)
[19] Sloan Digital Sky Survey, Alfred P. Sloan Foundation Science Programs (*https://sloan.org/programs/science/sloan-digital-sky-survey*)
[20] Thomas Sumner (2018): Gaia Data Releases Spark Open-Access Success, Simons Foundation (*https://www.simonsfoundation.org/2018/09/24/gaia-data-release/*)
[21] EPFL Open Science Fund (*https://research-office.epfl.ch/epflopensciencefund/*)