ETH BOARD

Open Research Data
Position of the ETH Domain

Adopted by the ETH Board on 13/14 May 2020

ETH zürich

EPFL

PAUL SCHERRER INSTITUT
PSI

WSL Swiss Federal Institute for Forest, Snow and Landscape Research WSL

Empa
Materials Science and Technology

eawag
aquatic research

Table of Contents

Preamble

Open science practices are quickly becoming a prevalent paradigm in academic research worldwide. Open Research Data (ORD) is at the core of open science, and new technological developments facilitate the sharing and exploiting of data sets, large and small.

The six institutions of the ETH Domain have mandated an internal working group to draft a position paper laying out a vision for ORD in the ETH Domain, including the motivations and challenges related to ORD. The paper was submitted for consultation within the Domain's institutions. The ETH Domain brings together some of the most important actors in the STEM (Science, Technology, Engineering and Mathematics) education, research and innovation landscape in Switzerland. Together, its six institutions represent over 22,000 employees and over 33,000 students (figures for 2019), and contribute significantly to Switzerland's successful research landscape. The ETH Domain has already assumed a leading role in adopting the emerging open science practices that allow scientific research outputs – including publications, data and software – to be disseminated and made accessible.

With this document, the ETH Domain is promoting a pragmatic approach towards open research data in Switzerland. By doing so, it affirms its readiness to play a leading role in the implementation of ORD.[1] The aim of this approach is to balance the interests of data producers and data users. It advocates the wide adoption of ORD for publicly funded research, but also acknowledges that norms vary considerably from one research discipline to another. It is therefore important that data producers should choose which data (sub-)set they wish to publish and when, and which license to apply to the published material. Institutional policies and research support staff guide those choices by providing expert advice, as well as balancing the interests of the various stakeholders, who may sometimes have contradicting objectives. These diverse stakeholder groups include, for example, public funding agencies (e.g. SNSF, EU, NIH etc.), spin-offs and companies that innovate based on ORD.

This document describes the motivation for the promotion of ORD and the challenges that need to be addressed. It outlines the international context of ORD and a vision of a pragmatic ORD approach for the ETH Domain. It is anticipated that some of the initiatives developed by the ETH Domain institutions to support ORD may be extended to all interested parties in the Swiss education, research and innovation landscape.

Why Open Research Data?

Research results are increasingly described in, or based on, digital data which – depending on their volume, complexity and sensitive nature – can be readily disseminated. Open Research Data are typically accessible publicly and can be used, reused and redistributed – provided that the data source is correctly attributed. ORD can therefore be used for further research, analysis or interpretation, as far as legally permissible and ethically justifiable. Two categories of research data can benefit from being shared openly: 1) data associated with

[1] cf. the mandate of the Federal Council to the ETH Domain: "Massnahmen für eine zukunftsorientierte Datenpolitik der Schweiz", 9 May 2018;
https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-70694.html and
https://www.newsd.admin.ch/newsd/message/attachments/52302.pdf

publications (e.g. for further analyses, verification, validation), and 2) additional data sets not directly related to a particular publication (e.g. complete data sets from monitoring programs, of which only parts were used in publications). Both these categories can be of interest to data producers as well as researchers not involved with the data collection, or to other parties outside the academic community (civil society actors, industry, etc.).

There may be various motivations for making research data openly available, including the following:

1. Science often benefits from multiple evaluations of the same data sets by researchers with different backgrounds, perceptions and ideas. ORD allows – and stimulates – new research and discoveries, as researchers can make use of valuable data even though they had not been involved in collecting it.
2. ORD incentivises the publication of well described and citable scientific data sets. It clarifies their authorship and ownership. To facilitate the use of data by others, data sets should be published under suitable licensing arrangements and be readily identifiable by digital object identifiers.
3. ORD makes research more transparent and allows published results to be reproduced, or to be included in meta-analyses.
4. ORD supports the longevity and long-term accessibility of data sets.
5. ORD facilitates collaborations within and beyond established networks, within the research community or with other parties outside the research community (civil society actors, industry, etc.). Researchers not belonging to established networks, such as young researchers, may particularly benefit from public access to research output that is otherwise confined to established networks.
6. ORD can be exploited commercially and may also foster the emergence of new business models for university spin-offs and beyond.
7. ORD can strengthen the impact of individual research projects and the overall impact of science on the economy and society. Furthermore, it helps to increase the visibility of researchers and institutions.

Challenges to address

Sharing research data openly comes with a number of challenges, including the following:

1. Sharing data early may lead to unfair competition.
2. Data may be misinterpreted on purpose or due to lack of detailed contextual knowledge.
3. Reuse of carefully gathered data may not be sufficiently acknowledged by others.
4. Making data publicly available in a useful form requires substantial human resources and infrastructure.

5. Long-term data storage and ensuring the long-term accessibility of data is costly, particularly for very large data repositories (of the order of petabytes), and the cost-benefit ratio for storing a particular data set may be difficult to determine. Substantial technical developments are required to address data curation, data life-cycle management, data archiving, data formats as well as data access via suitable indexes. ORD therefore requires long-term funding solutions for infrastructure and services and careful decision-making with regard to expiry dates for individual data sets.
6. For legal, ethical, privacy or security reasons, some data may not be suitable for sharing, or may only be suitable for sharing after having been carefully validated.
7. Personal data, particularly but not exclusively in medical research, has to be anonymised and protected from re-identification.

These are important points that need to be carefully addressed. The ORD strategy of the ETH Domain acknowledges these concerns and the need to put forward pragmatic solutions which address them.

A governance framework that defines the rights and duties of data producers, data owners and data users is needed. Indeed, a culture that gives proper credit to data producers needs to be promoted, along with the requisite tools; and research communities need to develop best practices for data citation. Such a framework should also define embargo periods during which data owners have an exclusive right to exploit the data and the opportunity to publish the data in a citable way. Embargo periods may be discipline-specific. The reuse of data should be subject to a data licence to ensure that it is used in accordance with generally accepted rules of scientific integrity. While the risk of results being misinterpreted is not confined to ORD, it will be accentuated for ORD, and the data user should be aware of this risk. It is therefore important to acknowledge that there are valid legal, ethical, privacy or security reasons for not sharing data, but any restriction should be justifiable – and, ultimately, justified.

A strategy should be formulated for the development of ORD-related infrastructure and services that responds to future demands and to develop funding schemes in order to support these infrastructures and services in the long-term. Such infrastructure and services, including long-term data storage, must be financially sustainable and should avoid the drawbacks experienced with the commercialisation of scientific publishing in the last century.

ORD: International and Swiss national context

A paradigm shift is taking place in the academic research community towards more openness, and it is important that ORD best practice should not be developed in isolation within the ETH Domain. Indeed, sharing research outputs openly – particularly data supporting results described in manuscripts submitted for publication in peer-reviewed journals – has become commonly accepted or even compulsory in an increasing number of disciplines over

the past two decades. Notable examples include structural biology,[2] genomics, and geosciences. Large research organisations (e.g. NASA[3] and CERN[4]) and large projects in other disciplines (e.g. astronomy[5] and Public-Private Partnerships in Drug Discovery[6]) have established their own data collections.

Most of the earlier movements towards ORD were field-specific efforts. They established ORD and adopted standards with regard to metadata, data formats, etc. However, not all research disciplines have created such ORD frameworks. As a result, there are significant differences with regard to progress to date in implementing ORD.

More recently, funding agencies, such as the Swiss National Science Foundation[7], European Commission[8] and the National Institute of Health[9] in the United States, started to promote – or request – ORD practices for projects funded through their programmes and funding schemes. In Europe, one non-field specific, publicly-funded portal for ORD is Zenodo[10]. In addition, a European Open Science Cloud (EOSC) for research data is in the making.[11]

On the national level, ORD has mainly been promoted through institutional rather than field-specific actions. Swissuniversities – with the involvement of ETH Zurich and EPFL – is currently working on a national Open Research Data Strategy.[12] Within the ETH Domain these actions consisted in establishing institutional ORD repositories[13] [14] [15] and open data policies as well as the promotion of an ORD culture, e.g. through incentives[16] and services to support ORD implementation[17] [18] [19] [20] [21] [22]. The ETH Board has specified the strategic focus area

[2] The Protein Data Bank (PDB) began as a grassroots effort in 1971, and by the early 1990s "virtually all journals that publish crystal structures of biomacromolecules made deposition into the PDB archive a requirement for publication". Berman HM, Kleywegt GJ, Nakamura H, Markley JL. The Protein Data Bank archive as an open data resource. *J Comput Aided Mol Des*. 2014;28(10):1009–1014. doi:10.1007/s10822-014-9770-y

[3] One example is the NASA EarthData portal: https://earthdata.nasa.gov

[4] CERN Open Data Portal: http://opendata.cern.ch

[5] The Sloan Digital Sky Survey, for example, represents "the most detailed three-dimensional maps of the Universe ever made, with deep multi-color images of one third of the sky, and spectra for more than three million astronomical objects": https://www.sdss.org

[6] Open Targets: https://www.opentargets.org and the Pistoia Alliance: https://www.pistoiaalliance.org

[7] ORD policy by SNSF: http://www.snf.ch/en/theSNSF/research-policies/open_research_data/Pages/default.aspx

[8] EU Science Hub of the Joint Research Centre (JRC): https://ec.europa.eu/jrc/en

[9] Several data collections are curated and maintained by the National Center for Biotechnology Information: https://www.ncbi.nlm.nih.gov

[10] https://zenodo.org/

[11] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

[12] https://www.swissuniversities.ch/themen/digitalisierung/open-science

[13] https://opendata.eawag.ch

[14] https://www.cscs.ch/publications/press-releases/2018/589/

[15] https://www.research-collection.ethz.ch/

[16] Open Science Fund of EPFL https://www.epfl.ch/research/open-science/in-practice/open-science-fund/

[17] Data Life-Cycle Management https://www.dlcm.ch/

[18] Data Champion Program of the EPFL Library https://www.epfl.ch/campus/library/services/services-researchers/rdm-contacts-communities/epfl-data-champions/

[19] https://opendata.eawag.ch/eawagrdm/

[20] https://www.envidat.ch/

[21] https://www.empa.ch/web/s909

[22] https://openrdm.swiss

"Data Science" as a strategic priority for the ETH Domain. Together, EPFL and ETH Zurich have established the Swiss Data Science Center (SDSC).[23]

An ORD vision for the ETH Domain

With ORD – and open science in general – it will hopefully become more straightforward to build on research work by other colleagues in the same field and beyond. This should make it easier to test new scientific hypotheses using data and data workflows developed in earlier work. ORD ensures that background information on data sets ("metadata") is available that allows reanalysis of the data independently of the data producer and possibly even in a different context from the one originally intended. In addition, ORD can be used in new fields and scientific approaches, particularly data science.

The culture and software environment of ORD and open science in general are emerging only now. The definition of the FAIR (Findable, Accessible, Interoperable and Reusable[24]) principles for ORD enabled the objectives of ORD to be made specific for the first time. However, actually translating the FAIR principles into concrete forms of ORD is an ongoing process. It is expected that Switzerland can lead the development and implementation of innovative technical solutions. As past experience has shown, the practical application of ORD that builds on these technological solutions will differ between scientific fields and follow field-specific standards. Solutions will therefore need to be customised depending on the field. Discussions on field-specific standards and technical solutions will need to take place in particular in those research communities, in which international standards do not yet exist; these discussions should be led in cross-institutional working groups in the different disciplines, potentially with input from other national and international actors.

As world-leading research organisations, the institutions of the ETH Domain believe that ORD has the potential to further increase the visibility and the impact of their research within the scientific community, the economy and society as a whole. The vision of the ETH Domain is to foster a research environment that

- supports ORD in research practice and creates an environment that values ORD, particularly in the context of evaluation committees on all levels;
- promotes ORD practices for research projects that are publicly funded;
- makes ORD tangible and visible as a catalyst for cutting-edge research and technology transfer;
- creates new opportunities for researchers, companies and other entities by developing innovative infrastructure and services that make ORD searchable and allow external data analysis tools to exploit the data via APIs. Ready-made tools for visualising ORD can increase the reuse of data, especially those in complex data sets;

[23] https://datascience.ch/; https://ethrat.ch/en/eth-domain/sfa/data-science

[24] Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; et al. (15 March 2016). "The FAIR Guiding Principles for scientific data management and stewardship". Scientific Data. 3: 160018. doi:10.1038/sdata.2016.18.

- creates new opportunities for researchers in Switzerland and worldwide that are inclusive and transparent by providing ORD and associated tools.

The ETH Domain envisions an ORD environment with the following characteristics, among others:

- Accessibility of ORD: access to ORD will be open to the general public through a choice of corresponding licenses.
- Searchability of ORD: information on where to find research data, on how to access it and on its reusability is publicly accessible. Particular attention should be given to establishing indexes in research fields that do not yet have established solutions. Data sets will be findable through search engines and information on the reusability of the research data will be complemented with quantitative indicators. These indicators on the use of the data can support decisions as to which data should be preserved for the long term, thus avoiding wasteful data storage.
- Infrastructure that combines data repositories with tools for data visualisation and exploitation: pure data repositories will be replaced by or integrated with infrastructure that provides standard interfaces between data storage and computational facilities; the latter can be deployed as user-specific computational tools for data visualisation and exploitation. While the infrastructure can be distributed, it is recommended that the ETH Domain should aim in the long term to create an over-arching access portal and unified standards with regard to metadata, licensing and searchability.

The quality and speed of implementation of the ORD culture into the research environment within the ETH Domain will depend not only on the commitment of the researchers but also on the availability of a suitable infrastructure, supporting services and an environment that values ORD as an important research output. The ORD culture has to be brought to life, and researchers – particularly new arrivals – must be made acquainted with it. A successful transition to an ORD research environment also requires new skills. The digitalisation of science means that all students and researchers from all fields are using computational methods to produce, organise, analyse or share data at various levels of complexity. Therefore, all students and researchers in the ETH Domain – from Bachelor to postgraduate level – should have access to training on research data management, data science, statistics and computational sciences.

Sources

Research Data Management Policy (2017) DLCM, CH. Web. PDF.

Higher Education Funding Council for England, (2016) Concordat on Open Research Data. Project Report. Higher Education Funding Council for England, UK. Web. PDF.

von der Heyde, Markus. (2019, May 22). Open Research Data: Landscape and cost analysis of data repositories currently used by the Swiss research community, and requirements for the future (Version 1.0.0). Zenodo. http://doi.org/10.5281/zenodo.2643460

Swedish Research Council. (2019) An outlook for the national roadmap for e-infrastructures for research. ISBN 978-91-88943-07-1. Web. PDF

Verheul, Ingeborg, Imming, Melanie, Ringerma, Jacquelijn, Mordant, Annemie, Ploeg, Jan-Lucas van der, & Pronk, Martine. (2019, May 6). Data Stewardship on the map: A study of tasks and roles in Dutch research institutes. Zenodo. http://doi.org/10.5281/zenodo.2669150

Dunning, Alastair. (2018, June 26). TU Delft Research Data Framework Policy. Zenodo. http://doi.org/10.5281/zenodo.2573160

Ashley, Kevin (2016) Developing Skills for Managing Research Data and Software in Open Research. Wellcome Trust. https://dx.doi.org/10.6084/m9.figshare.4133916

Directorate-General for Research and Innovation (European Commission) O'Carroll, Conor; Hyllseth, Berit; Berg, Rinske van den; Kohl, Ulrike; Kamerlin, Caroline Lynn; Brennan, Niamh; O'Neill, Gareth (2017) Providing researchers with the skills and competencies they need to practise Open Science - Publications Office of the EU. https://doi.org/10.2777/121253

Universities UK Open Research Data Task Force (July 2018) Realising the potential: final report of the Open Research Data Task Force. Web. PDF.

Universities UK Open Research Data Task Force (June 2017) Research data infrastructures in the UK. Web. PDF.

ETH Board
Zurich and Bern
www.ethboard.ch
3 June 2020