NICT Working Group on Research Data Management

White Paper - A View of Swissuniversities' Network of ICT Experts on the Research Data Management Landscape in Switzerland

Final version, all comments integrated - 2019-05-06 (Michael Brüwer)

Contributors (alphabetical):
C. Barillari[1], M. Brüwer[2,9], P-Y. Burgi[3], M. Gay[4,9], N. Lang[5,9,] M. De Lorenzi[6,9], H. Lütcke[1], B. Rinn[1,6], T. Sengstag[7,8]

Affiliations:

[1] Scientific IT Services, IT Services, ETH Zurich
[2] IT-Services, University of Basel
[3] Système d'Information, University of Geneva
[4] Servizio informatico, Università della Svizzera italiana
[5] Corporate IT, FHNW
[6] Swiss National Supercomputing Center (CSCS), Lugano
[7] SIB - Swiss Institute of Bioinformatics, Lausanne
[8] sciCORE, University of Basel
[9] NICT

Preamble

In Summer 2018, the swissuniversities' P5 project approached NICT to obtain a perspective on research data management from the viewpoint of ICT service organizations. This document is the result of a series of workshops held in autumn 2018 by a group of experienced staff from several organizations. Common issues were identified including funding sustainability and infrastructure challenges, and dependencies on other domains of data management (e.g. long-term preservation, governance) and some gaps (e.g. data curation, training) are highlighted. The members discussed these issues with colleagues from libraries and research in order to complete their views. However, this is not a comprehensive overview, and a wider discussion with further stakeholders, in particular with researchers and libraries, should follow.

Summary

Systematic management of research data is fundamental to research excellence in several aspects: it fosters good practice, enables collaboration and facilitates re-use of data. Prerequisites for systematic data management are awareness and sharing of these objectives by the

researchers. Comprehensive research data management (RDM) support services should address researchers' needs and demands during all steps of the data life cycle.

What supporting services should be developed and implemented in order to address gaps in today's RDM service landscape? The working group came up with the following non-exhaustive list:

i) Education and training in the fields of
   a. data governance and
   b. data management plans foster good practice
   c. RDM education/training should become part of the curricula
   d. Researchers should be trained to overcome resistance to changing work habits.

ii) IT-infrastructure should deliver
   a. fast, secure and accessible data storage with clear cost implications;
   b. easy-to-use generic local and cloud services for data handling to support agile and creative work.

iii) Software services:
   a. Electronic lab notebooks (ELN), laboratory information management systems (LIMS) and automated scientific workflows help organize data in the lab and to implement data governance;
   b. Virtual research environments (VRE) foster collaborations and sharing of information across institutional and national borders.
   c. Generic repositories under the FAIR principles provide long-term archiving of research data accompanying publications or facilitate the dissemination of curated data sets for re-use.

Different services at different organizational levels (research group, department, faculty, university or research organization, national, supranational (EU) or international) can address these demands. National services can support researchers and significantly relieve local support and services. National services can complete local technical infrastructures and software services and thereby respect specific national conditions, laws and features. Smaller institutions in particular must rely on emerging national services, or will otherwise not be able to fulfil the new demands of FAIR data management. A number of services emerged from P2 and P5, and it is hoped that more will appear during the next funding period. However, the issue of how these services can be operated sustainably remains, to an extent, unresolved.

In order to improve the overall capability of Swiss universities in the field of research data management, the following topics have to be addressed. Our recommendations are as follows:

- Research data management has to be done from the start of a research project. To enable this, appropriate service offers need to be made easily available to the national community, researchers need to be proactively informed about them, and sufficient funding needs to be made available for researchers to use them as part of the funding for research projects.

- Subsidies and matching funds were essential and appropriate (and will be in the future) for extending and further developing local services to broaden the user base and to bring services to the market. However, this is not a feasible business model for long-term operations of costly services and infrastructures.
- Universities are not prepared to run businesses beyond a certain scale. To improve the sustainability of RDM services, costly long-term national RDM services must rely on legal bodies with clear tasks, governance and long-term commitment. Long-term mandates should be given to these entities, with clear objectives and governance structures. One-off fees and full subsidy are feasible options to finance these organizations. Researchers must be financially enabled to pay these fees and to preserve data beyond the duration of research projects.
- National services cannot fully substitute local or domain specific services. Additional resources and personnel are missing locally in the first place at the research institutions: personnel to support data curation; personnel to run IT-infrastructure and software services to support the early phases in the data life cycle (data creation, processing and analysis); personnel to educate researchers on all career levels on RDM related topics; and personnel to coordinate and integrate national services with local service portfolios, to develop and maintain interfaces and to run the governance processes. RDM services and support must be available at universities that do not have sufficient critical mass or manpower to develop them internally.

Introduction

Academic research produces ever-larger amounts of data that have to be stored, interpreted and shared. Furthermore, funding agencies, journals or academic institutions issue new guidelines and policies regarding data management and sharing regularly. In particular, research data should be published according to the FAIR (Findable, Accessible, Interoperable, Reusable) data principles (see Appendix). As a prerequisite, researchers must properly manage data during all steps of the data lifecycle, from initial data collection to processing and analysis. Data management according to FAIR principles can also benefit individual research groups by increasing efficiency, improving data provenance tracking and reducing the risk of data loss. Professional data management services, support, and software platforms are needed to enable researchers to benefit from the FAIR and open science movements.

In this report, we aim to provide an overview of the support for research data management (RDM) that is currently available for researchers in Switzerland. We first illustrate the data lifecycle (Figure 1) to provide overall context and show how data and storage requirements (Figure 2) evolve throughout a typical research project. Next, we show which kind of expertise is needed at what stage of the data lifecycle (Figure 3) and the related support needs (Figure 4). We then present a network diagram of the various actors involved in supporting researchers with their RDM needs (Figure 5). We provide an overview of tools and services that are relevant at different stages of the data lifecycle (Figure 6) and illustrate what fields are candidates for cooperative or national services. Finally, we show the strengths of a different business model for innovation vs. sustainable provisioning of high-quality RDM support services and we provide a discussion of possible future directions to fill the gaps in the currently patchy RDM landscape in Switzerland.

The research data lifecycle

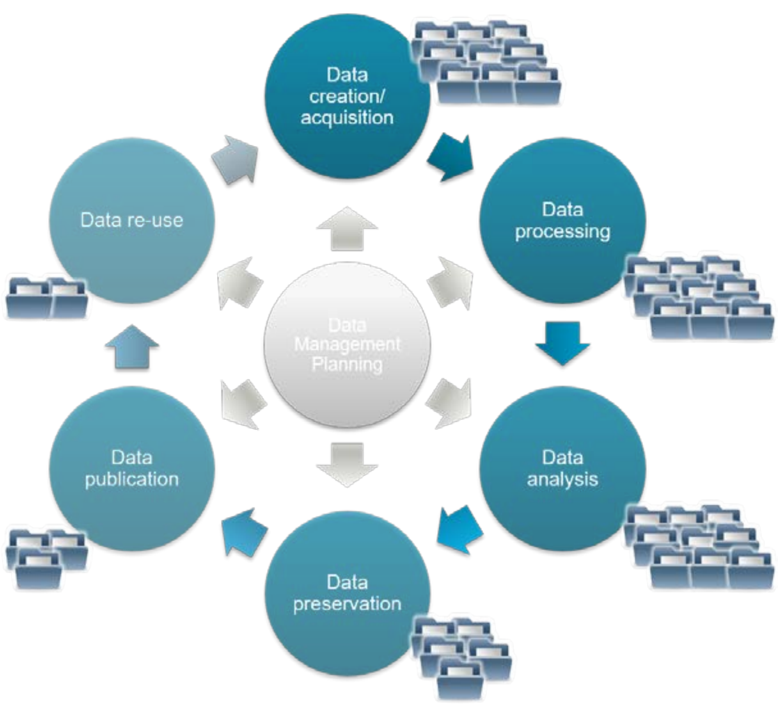

Figure 1: The data lifecycle illustrates how research data evolves through a typical research project. Initially, new data are created or acquired, for example by measurements, observations or simulations. Subsequently, data are processed (e.g. refined, cleaned) and then analyzed. Importantly, the steps of data creation, processing and analysis are typically additive, i.e. each step creates more data and the *total data volume increases* (though the data sizes during later analysis steps is usually smaller than that of the primary data). After data analysis, a small subset of primary and results data is usually selected for long-term preservation in archives and/or publication in data repositories. The selection criteria for publication data is whether it relates to the findings of the journal publication. In an "Open Data" centric research environment, primary data of high quality could also be selected for "data-only" publications. Finally, a smaller fraction of the published data is usually selected for re-use in subsequent research projects. At the heart of the data lifecycle is the data management plan, a formal document that explicitly states how a research project will handle data throughout the steps of the data life cycle. The steps 'Data creation', 'Data processing' and 'Data analysis' are usually summarized as 'Active Research Data Management'.

Data Lifecycle and Storage Requirements

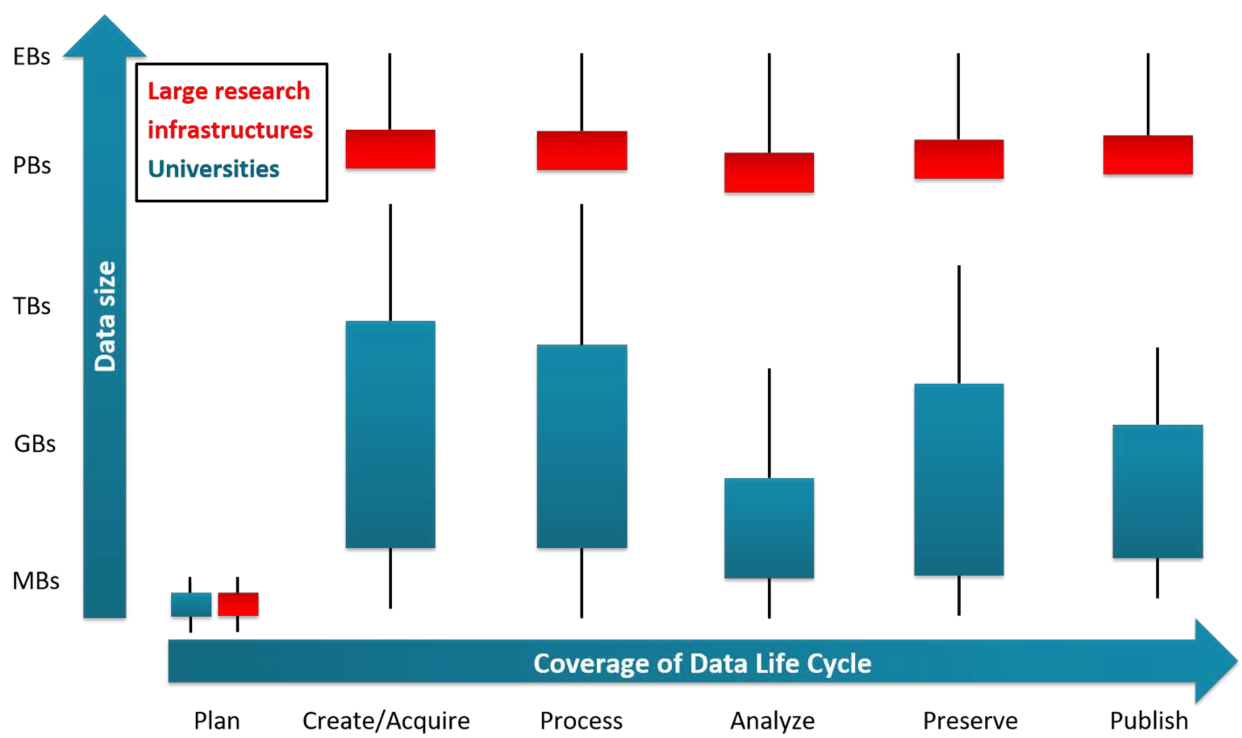

Figure 2: In University research labs, storage requirements tend to differ significantly at different stages of the data life cycle. The largest data amounts are encountered and have to be managed during the 'active research data management' phase (Acquisition, processing &

analysis). The amount of data selected for long-term preservation and publication is typically smaller (sometimes 1 - 2 orders of magnitude, depending on the discipline). Data sizes during the planning phase (Data Management Planning, DMP) are negligible. On the other hand, large-scale research infrastructures (red boxes) produce, process, analyze and publish / preserve data in the Exabyte range. Examples are the SwissFEL (www.psi.ch/swissfel), SLS (www.psi.ch/sls) or complex international projects including the Cherenkov Telescope Array (CTA, www.cta-observatory.org), the Square Kilometer Array telescope (SKA, www.skatelescope.org) and the Large Hadron Collider (LHC, home.cern/science/accelerators/large-hadron-collider). Large research infrastructures such as the LHC typically substantially reduce data volumes immediately after acquisition ("stream processing"). The resulting data are then further processed and have to be preserved for the future without significant reduction in data volumes.

Multiple Expertise Required

Supporting the Data Lifecycle relies on expert knowledge from multiple actors across the research landscape. From IT systems to ethical, legal and social implications (ELSI), four broad categories of expertise are identified:

- **Information technology** expertise: This category of experts is responsible for operating the core IT infrastructure that supports data management: storage and backup systems, operating systems, virtualization and cloud technologies. IT experts do not require specific research domain expertise but must ensure the long-term reliability of information systems, which includes supporting the transition of data across technologies as they evolve.
- **Data management systems** expertise: This category of experts develops and/or maintains the data management software required to capture and/or preserve research data. These experts translate data management requirements of researchers into software requirements, which can then be developed or acquired. They understand the data formats and technologies (e.g. databases) required to manage data efficiently. They also have an understanding of UX (user experience) in the design of human interfaces, and of data exchange protocols (APIs) which allow the implementation of interoperability between data management systems. Expertise in the respective fields of research is required or at least helpful to fulfill these tasks.
- **Data management support and curation** expertise: This category of experts has an in-depth understanding of the research process in a specific domain or interdisciplinary expertise in data management and curation. They can advise researchers on the implementation of the Data Life Cycle in a research project (e.g. DMP support) and on the current best practices in a domain. They provide guidance to researchers in the selection of relevant ontologies for data annotation and support the annotation process. Apart from individual advice, they provide training on RDM for researchers at all career levels. They maintain an overview of the relevant domain-specific repositories (and associated requirements, such as metadata and semantics). They can help researchers select the most appropriate data management system(s) for their research, and as such maintain an overview of available solutions in a domain. In addition, they advise IT staff on how and

when data should be converted to a new format to avoid being locked-in an obsolescent technology that could render data unreadable.
- **Data governance** expertise: This category of experts can provide support on the legal-ethical requirements associated with specific data. They understand the regulations applicable to data management (laws on data protection, copyright, institutional internal regulations) and can provide advice on the suitability of a specific research process (e.g. data analysis process) or of repositories to host specific data (e.g. give institutionally binding opinions on the adequacy of the governance process of a repository given the nature of the data).

This categorization is illustrated in Figure 3.

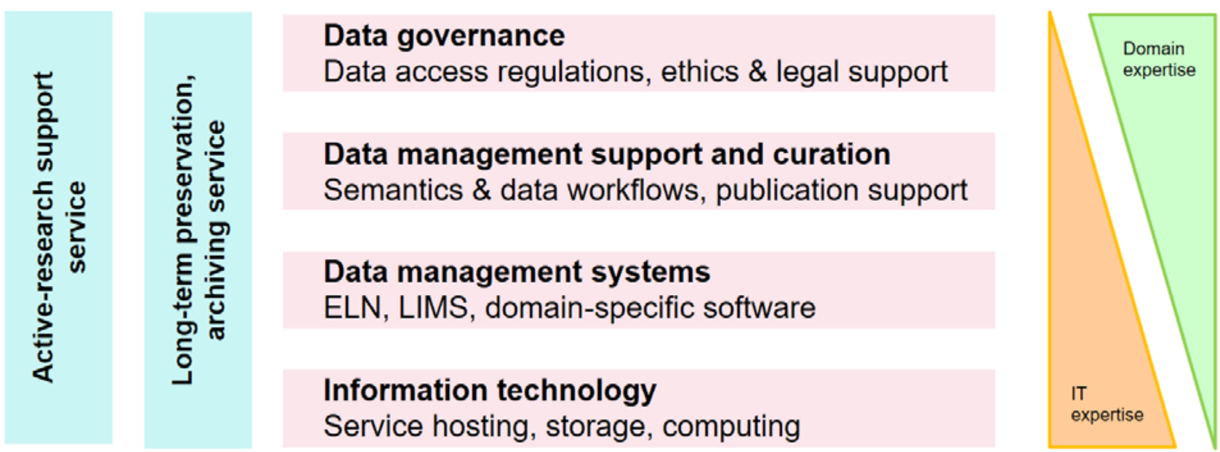

Figure 3: Four categories of experts involved in Data Lifecycle support. Various levels of IT and research-domain knowledge is required in each category. Experts in each category should collaborate closely with experts in the neighboring categories to ensure seamless data management support. (With contribution from S. Bellanger University of Basel.)

Recognizing the complexity of research data management and the interdependence of services in supporting the Data Lifecycle, several universities have initiated an internal coordination effort. For instance, at the University of Basel, the ongoing FDM (Forschungsdatenmanagement) project is establishing the synergies between IT services, the library, the legal services and the research support services of the faculties. An essential component of this project is the governance structure, which has its steering at the Rectorate level and involves researchers as key stakeholders in the definition of the project priorities. The first major outcome of the project was the recognition that the largest needs of the local research community are at the level of data management support and curation (as described above) and DMP support. Figure 4 illustrates the most critical support gaps as identified in the project.

At ETHZ, the ETH Research Data Management support has two key players: the ETH Library and ID Scientific IT Services (SIS). Together, the two units support ETHZ researchers during the whole data lifecycle, with consulting, tools and services. (Further details can be found on the Research Data Website (http://www.ethz.ch/researchdata)). More specifically, both units provide a joint consulting service for DMP writing. SIS provides tools and services that cover the active data management phases (from data acquisition to data analysis). The ETH Library provides tools and services for long-term data preservation and publication. Strong emphasis is also given to research data management training, and several courses are run jointly in an effort to cover the complete data life cycle. Each unit also provides independent courses on topics specific to their area of expertise.

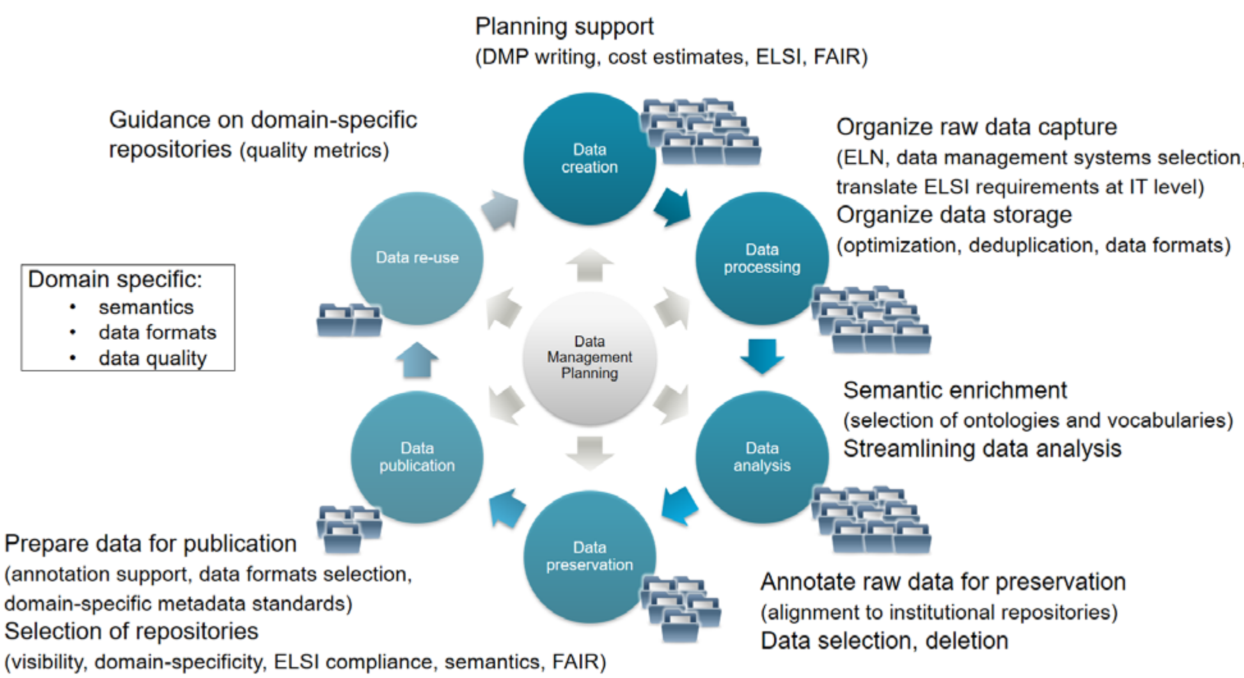

Figure 4: Data management support gaps as identified in the FDM project (Uni Basel) Domains of support with the greatest needs (processes and active support). These gaps are also understood to exist in other Swiss academic institutions (Data Stewardship gap). (With contribution from S. Bellanger University of Basel.)

Aspects of Active Research Data Management

The relevant aspects of managing active research data differ between research fields. For quantitative research fields however, there are general aspects that tend to be relevant in most cases. In experimental labs, the process of Active Research Data Management starts with the management of an inventory of lab assets, which are usually managed by a Laboratory Information Management System (LIMS). We find that most laboratories nowadays have a digital tool to manage their lab inventory, but that the sophistication and maturity of these tools varies greatly between labs.

A second aspect is the management of primary research data (measured or computed from models), together with an appropriate set of metadata to describe them to researchers not part of the original research team. In our experience, all labs have a filing system for their original research data, usually with appropriate backups. The metadata describing them, however, are often spread between different locations and systems, and it is not uncommon to find metadata to only be given in implicit form that is hard for other people apart from the author to understand.
A third aspect is the management of standard (experimental or computational) protocols. As with the metadata, such protocols are often in heterogeneous formats, locations and systems.
A fourth aspect covers data processing workflows, data analysis source code and notebooks, processed (intermediary) data and publishable results. Today, many tools exist to structure and manage these assets; however, we find that the chance of these tools being used is usually decreasing the further a digital asset is remote to a final published result.
A fifth aspect concerns the experimental descriptions and notes for individual experiments. For these assets, it is still common to use paper notebooks that try to give an indication on where the connected binary artifacts of the experiment can be found. Electronic Laboratory Notebooks (ELNs) that are suitable for academic research labs and capable of linking together all information and digital assets belonging to an experiment are already available, but are currently only used by a minority of researchers, we assume, because many researchers are not well informed about available tools. However, from our experience, the main reason is that introducing any ELN (or even more basic IT-Infrastructure such as managed PC and remote filesystems) into the work process of a researcher requires a substantial change to their work habits, which many researchers perceive as being disruptive to their research productivity, and thus tend to defer this to the future. In order to overcome the required ‘activation energy’, a researcher usually needs to fully understand the rationale, be provided with an easy-to-use tool for this purpose, receive sufficient training and support to use the tool, and additionally obtain strong backing from their lab head to go through the process. At ETH Zurich, for example, a unified offering for ARDM with a low entrance barrier has been identified as offering high value to improve RDM practices. To this end, ETH SIS has developed the “ETH Research Data Hub” service, which is available to all researchers at ETH Zurich.
In addition to LIMS and ELNs that are mainly used in “hard” science, Virtual Research Environments (VRE) are proposed to promote the proper use of computer techniques in basic research in the arts and humanities. Such environments allow the management of digitized materials, from their digital capture, character recognition, RDF representation, syntactic and semantic analysis, etc., to their publishing on web pages. One instance of such a project is the Bodmerlab (bodmerlab.unige.ch), a research and digitization project resulting from a partnership between the University of Geneva and the Martin Bodmer Foundation, which makes a significant proportion of rare and fragile documents available to the research community and the wider public. Each collection has its own particularity: two-thousand-year-old papyri require a different technical and bibliographical treatment than 19th century editions of Faust. Before aggregating hundreds of thousands of digitized pages under the umbrella of "big data", it is necessary to mark the significant features in the coding procedures. When these data are also accompanied by comments that highlight their cultural significance, we enter the field of "smart data" that VREs help organize.

RDM Support in Switzerland - Key Actors

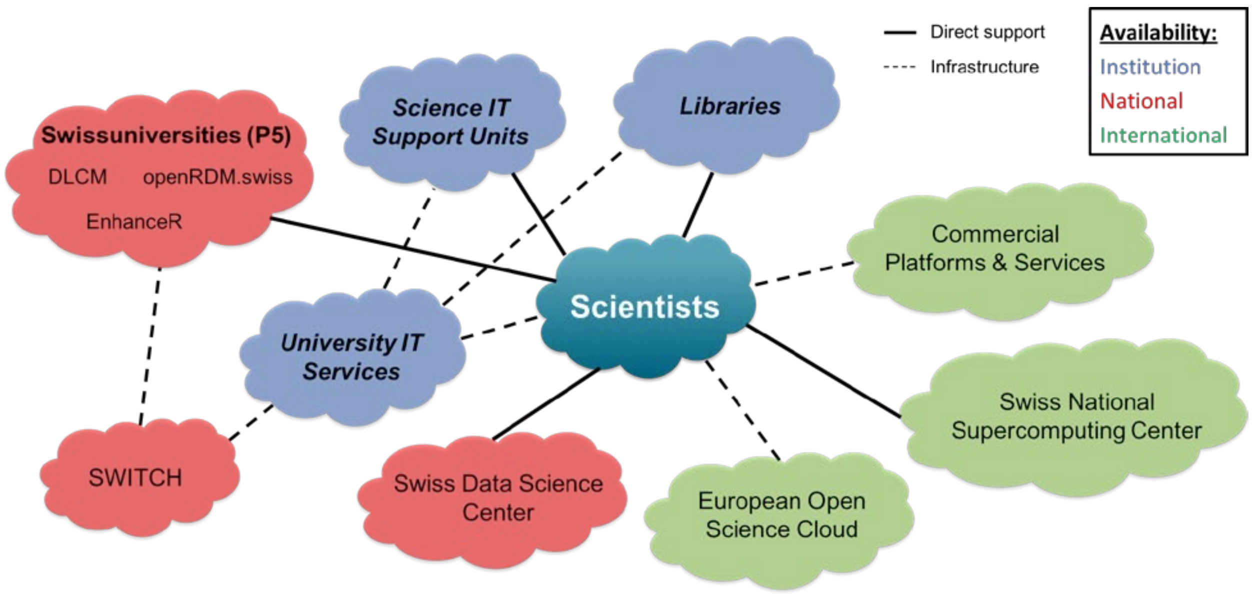

Figure 5: Network diagram of key actors supporting Swiss scientists with RDM. Core RDM support is provided by the Science IT Support units and libraries at the different institutions. Science IT units support the data creation, processing and analysis steps of the research data lifecycle ('active research data management'), while libraries support long-term preservation and data publication. Via national projects funded by Swissuniversities within the P5 program (e.g. EnhanceR, DLCM, openRDM.swiss), IT units are able to provide support to scientists at different institutions. In parallel, projects such as DLCM facilitated the dissemination of knowledge and competencies on the level of training and supporting materials. Two national domain specific repositories have been funded by P5 and became productive: Forscenter.ch and Dasch.ch.
IT infrastructure services relevant to RDM (e.g. storage, archiving, networking, computing) are provided by IT services, sometimes leveraging national infrastructure provided by SWITCH. Furthermore, scientists frequently rely on commercial ('cloud') services and platforms for their RDM needs (e.g. storage, lab notebooks, Wikis).
The Swiss National Supercomputing Center (CSCS) provides support for computationally demanding research projects. In addition, the recently established Swiss Data Science Center (SDSC) provides a platform and support for data-intensive science. Note that the key actors are also research infrastructures such as CERN, SKA or SwissFEL, and international consortia and collaborations, e.g. European Open Science Cloud, that need federated storage solutions because they operate in an international context.

Tools and Services to Cover the Data Life Cycle

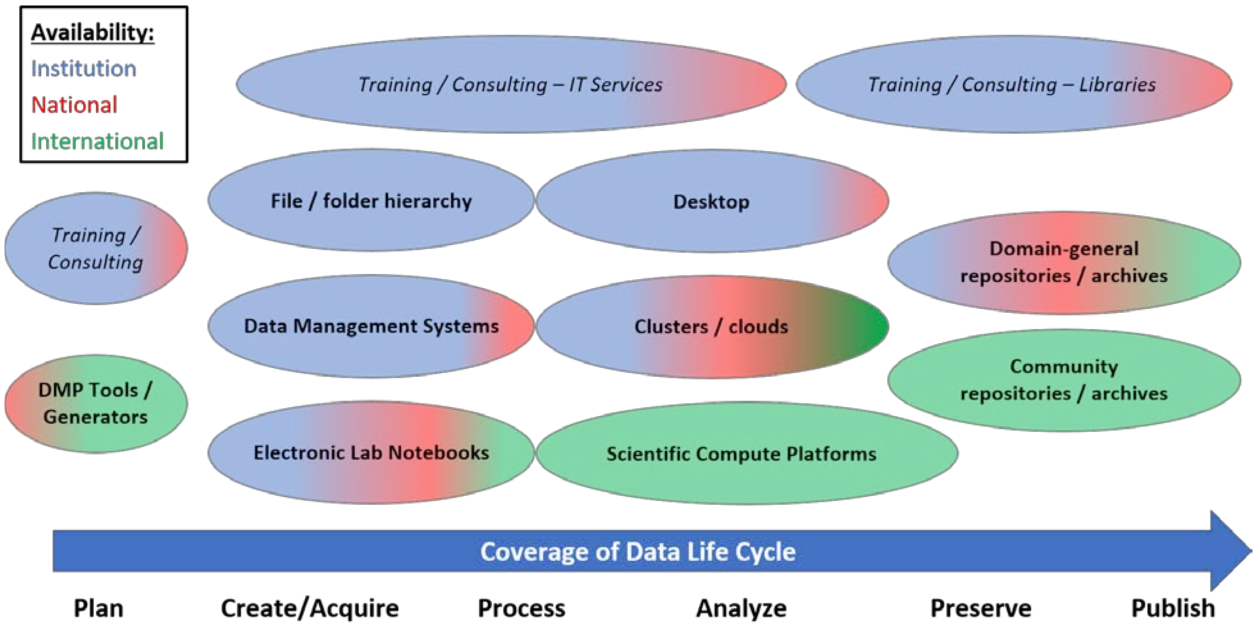

Figure 6: A diversity of RDM tools and services are available locally at institutions, nationally or internationally. Currently, training and consulting services are typically provided by specialized units in the IT services or libraries of individual institutions. In future, these services could be accessed by others via a national coordination desk. Usually, IT services provide the infrastructure for file or folder based data management and desktop computing. At some institutions, RDM systems or electronic lab notebooks are also available. For advanced data processing and analysis, institutional compute clusters and clouds are available, complemented by national offerings (CSCS, SWITCHengines). Most recently, international scientific compute platforms such as CodeOcean or mybinder.org have become popular in parts of the scientific community. For data preservation and publication, numerous subject-specific repositories are available (typically international). In addition, a number of domain-general repositories exist for institutions (e.g. ETH Research Collection), nationally (e.g. DLCM longterm preservation, available early 2019) and internationally (e.g. zenodo.org). Because of the size of the managed data, especially for research infrastructure, it makes sense to develop specific tools (and in part services) to cover one or more steps in the data lifecycle. An example is SwissFEL where CSCS developed, together with PSI, an infrastructure to cover the steps related to processing, analyzing and, in part, publication of their data.

Funding and business models for cooperative RDM support and services

There is a variety of options for funding research data repositories and related data management services. We have grouped these into three categories: i) structural funding, ii) cooperative funding and iii) fees-based service-networks[1].

i) Structural funding of research data management services is guaranteed long term, when the service is delivered by a **research institution** (e.g. university) or by an organization with a national mandate (e.g. CSCS and SDSC). Funding agencies can provide sustainable funding as well (e.g. Zenodo, partially financed by the EU commission). Large research projects with huge data volumes (e.g. the European Spallation Source-ESS or the Square Kilometer Array-SKA projects) have to deal with data management and preservation during the development of the research infrastructure. The nature of structural funding may impose constraints on the use of repositories and data management services. For example, institutional repositories may only be available to researchers collaborating in projects at the funding institution. On the other hand, funding from the home institution of a researcher may be the only option to sustain large-volume research data beyond the end of a research project.

ii) Cooperative sustainable funding occurs in a variety of forms and **legal bodies**: Foundations (e.g. SWITCH), joint-stock companies (e.g. Swiss Library Service Platform - SLSP), and associations (e.g. Speicherbibliothek, which is also a stock company). Here, the costs and risks for development and operations of the services are shared among partners. The sharing of costs usually takes into account usage fees and/or the size of the partners. Most repositories and research data management services in this category are based on hybrid tariff models, where usage fees are supplemented with subsidies from institutions or funding agencies. The "freemium" model adopted by some services foresees a free (and therefore subsidized) service for small volumes of data and the payment of fees for larger volumes. This seems particularly appropriate for consulting or information services which are offered on a basic level free of charge, with more customized or time-intensive services being billed. In any case, the partners show a long-term commitment to guarantee that the services are delivered and further developed by these institutions. Thus, this model is especially suitable for long-term services and has been adopted in the past with very good success.

iii) Within a **network of collaborating institutions,** one partner delivers services to others. Some institutions that have built a good reputation by operating services with higher maturity than others are in a position to open up these services to third parties. Service providers (which will probably be larger organizations) can further improve the services with the income from projects and from service fees, while the customers (probably smaller organizations) can benefit from the knowledge and maturity of the providing partner. The volatility and innovation, propelled by seed money, is higher in this business model than in the others, which might put

[1] An in-depth discussion of the business models can be found in: "Business models for sustainable research data repositories", *OECD Science, Technology and Industry Policy Papers*, No. 47, OECD Publishing, Paris, https://doi.org/10.1787/302b12bb-en.

some restrictions on the usage of this model. Service providers must plan to provide the service mid- to long-term and must ensure that a denounced service has an appropriately long ramp-down phase and a migration path for existing customers. Researchers must ensure that data is preserved beyond the duration of the funding of individual research projects. Contracts and service level agreements are needed. Options for tariff models within the "network business model" include free usage, one-off fees, annual subscriptions paid by a university (e.g. Olat), and fees based on consumption rates. Recent discussions envision that services under this business model will be supported by SWITCH by means of a service catalogue, and contracts and SLAs and will be governed by swissuniversities..

To summarize: Funding agencies require long-term data management plans while the project funding only covers a limited number of years. This was also one of the outcomes of a recent workshop[2] and researchers demand a solution to this problem.

Future directions

The current landscape of RDM services and support in Switzerland is segmented. On one hand, a number of institutions have established excellent RDM support services and tools for their researchers (e.g. ETH Zürich, EPF Lausanne, UNIGE). On the other hand, many (smaller) institutions so far lack the resources to set up dedicated RDM support units. To present a comprehensive catalog of existing RDM solutions available for the different domains was out of scope for the working group, but would greatly improve the visibility of available solutions.

At the national level, some projects have addressed these issues and already offer RDM support and services to all research groups in Switzerland (e.g. DLCM, openRDM.swiss), or will do in the near future. In these projects, one or several institutions offer expert RDM services to the entire academic community in Switzerland. Currently, these projects operate with a hybrid business model (see above) where they receive funding from funding agencies in addition to collecting user fees. The long-term sustainability of these projects (without subsidy funding) remains to be proven. If Switzerland wants to maintain its strong research position, adequate financial resources must be dedicated to RDM. Furthermore, it is highly desirable that Swiss researchers will be able to benefit from future RDM services established in the context of European Open Science Cloud (EOSC).

Switzerland is currently dotted with a number of high-quality services and infrastructures providing excellent RDM support. Nevertheless, these services are currently not available to all researchers to the same extent. Institutions differ in profile, priorities and maturity. Existing services still partly lack visibility, governance and sustainability. The challenge and the vision of the P2/P5 programs that all researchers in Switzerland have access to the same high-level RDM support still remains valid and attractive.

---

[2] https://naturwissenschaften.ch/service/events/105080-open-data-and-data-management---issues-and-challenges

For the next funding period and future “open science” projects, we believe that a national network for scientific computing and data analytics would be highly valuable. At the moment, commercial computing and reproducibility platforms, such as Code Ocean (codeocean.com), are becoming increasingly popular in certain scientific disciplines and are endorsed by publishers to ensure reproducibility of published results. However, as computing demands and the prices charged by commercial platforms are likely to increase in the future, the associated costs may become prohibitive. Moreover, the Swiss academic community already has access to world-class scientific computing clusters. We therefore believe that, setting up a national scientific computing and analytics network that makes existing clusters available to all Swiss researchers will be more cost-efficient, in the long run, than relying on commercial solutions. A useful amendment could be to have the network offer reproducibility services like https://mybinder.org for all of the Swiss research community.

Data curation is an essential step in the data life cycle. Data curation is crucial to enable re-use. Another open issue in the context of reproducibility of scientific results is the archival of hard- and software, that has been used to transform or analyze data. These are major gaps in today’s RDM landscape.

Appendix: The FAIR Guiding Principles

The FAIR Guiding Principles are guidelines to make data discoverable and processable by *both* humans and machines. They were first published in the article by Wilkinson et al.: *The FAIR Guiding Principles for scientific data management and stewardship*, Scientific Data **3**, 2016, DOI: 10.1038/sdata.2016.18

The FAIR Guiding Principles are agnostic about the standards used to annotate data; domain-specific consensus is expected to emerge over time. A data resource is FAIR if it complies with the criteria enunciated in the article (and reproduced below). A formal framework to evaluate "FAIRness" of a resource is provided by the FAIR metrics: http://fairmetrics.org/

FAIR criteria

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource
To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
    A1.1 the protocol is open, free, and universally implementable
    A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available
To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data
To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
    R1.1. (meta)data are released with a clear and accessible data usage license
    R1.2. (meta)data are associated with detailed provenance
    R1.3. (meta)data meet domain-relevant community standards

Appendix: Handles and Persistent Identifiers

To reap the benefits of data use, reuse and recycle inherent to the FAIR principles, data should be unambiguously and clearly identified by something that depends on data content and not location. When these identifiers are trusted, they become the basis for trusting the data and for authorship assignment.

Historically there were plenty of persistent identifier proposals: for example, ISBN for books, IETF RFC and others. Today there are two schemes for digital artifacts; DOI for publications, and PID for scientific data.

DOI and PID work on similar principles. Both are handle systems that associate an opaque string to a URL pointing to the digital artifact. Both send this string to a resolver (https://dx.doi.org/ for

DOI and https://hdl.handle.net for PID) that returns the corresponding URL to the requestor. If the data or publication moves, the owner changes the associated URL in the resolver database. The two systems differ in the structure of the managing consortium and mission, and in the metadata associated with the identifier.

Persistent Identifiers (PID) are managed by the ePIC consortium (https://www.pidconsortium.eu/), which CSCS is member of, that manages and resolves them for Switzerland. DOI is managed by ETHZ for Switzerland.