ETH Open Research Data Program

Report on ORD-related services and infrastructures in the ETH-Domain

by the Expert Group Services & Infrastructures of the ETH ORD Program[1]

Version 3.0

June 2023

Table of Contents

[1] ***Members:*** *CA Pignedoli (Empa), G Baud-Vittoz (EPFL), H von Waldow (Eawag), H Lütcke (ETHZ, Chair), I Iosifescu Enescu (WSL), I Eula (EPFL), K Malleck (EPFL, Co-chair), L Desimone (EPFL), L Sala (PSI), M Töwe (ETHZ), R Roškar (SDSC), SE Bliven (PSI), U Beyerle (ETHZ), MG Giuffreda (CSCS), N Marzari (EPFL), PGA Pedrioli (ETHZ), R Corvalan (EPFL), C Berner (Eawag), A Gael (EPFL)*

ETH BOARD

1. Summary

The present report on Open Research Data Services and Infrastructures (ORD S&I) in the ETH Domain institutions was prepared by the Expert Group Services and Infrastructures (EG-SI)[2]. It provides the basis for activities in Measure 2 (Coordination of Access to Research Data Management (RDM) Services & Infrastructures) of the ORD-Program of the ETH Domain.

The report summarizes the landscape of existing ORD-related S&I within the ETH Domain institutions along the research data lifecycle. The research data lifecycle provides a simplified model for describing the data-related stages of a typical research process (see Figure 1 below). The report identifies the diverse offers of ORD-related solutions in ETH Domain institutions, ranging from *Data Management Planning* (DMP) consulting offers and tools, S&Is for *Data Acquisition, Storage and Annotation* as well as for *Data Processing and Analysis*, to repositories for *Data Publication and Reuse* as well as issues related to *Data Preservation and Disposal.* Most institutions in the ETH Domain have established solutions and services for assisting researchers in drafting *DMPs*. Due to the fact that individual DMP consulting is labor-intensive and does not easily scale, all ETH Domain institutions make available DMP templates (usually for SNSF grant applications). In the long term, a central DMP repository may help to forecast future infrastructure-related needs and costs, because the requirements described in the DMPs have to be supported by relevant infrastructure and services.

The ETH Domain institutions also offer a variety of *Data Acquisition, Storage, and Annotation* solutions, starting from basic storage services to more specialized solutions. Researchers use various approaches and tools, including conventional file-and-folder based data management and different types of Electronic Lab Notebooks (ELNs). Unfortunately however, the different ELN solutions are not interoperable with each other and this situation should be improved. This report also identifies available S&Is for *Data Processing and Analysis*, focusing on version control, computational notebook offerings as well as platforms for data processing. For version control, Git is currently the most popular tool and GitLab platforms are provided by several institutions in the Domain. Computational notebooks, such as Jupyter, have also become popular in many research disciplines and are widely used for interactive scientific computing. But combining computational platforms with data management systems for reproducible analysis of large datasets is still a challenge. Recently, this challenge started to be addressed with the development of reproducible research platforms (RRPs). These platforms combine and extend tools for version control and interactive computing and aim to make these tools available to a wider audience with less technical skills. The development of such emerging RRPs should be aligned with the researchers' requirements.

Data publication is of central importance for the success of ORD. ORD publication support is already possible through the available landscape of repositories for *Data Publication and Reuse* offering various options for obtaining persistent identifiers for research data publication and citation in the ETH Domain. Nevertheless, the different repository platforms are not interoperable with each other and this situation could be improved. Furthermore, the publishing and sharing of very large datasets (TBs and upwards) remains a significant challenge for the available repositories of the ETH Domain. Finally, regarding the topic of *Data Preservation and Disposal*, the long-term storage and archiving of research data is paramount. Furthermore, the report empasises the importance of *Cross-cutting solutions* such as ORD training and consulting as well as an interoperable identity and access management in the ETH Domain.

In conclusion, the ETH Domain has an extensive portfolio of state-of-the-art ORD solutions that is however not without limitations. Technical challenges, such as the lack of integration and interoperability between existing solutions, should be addressed by the existing S&I providers, and non-technical challenges, such as governance, policies, and adoption should be addressed by the governance structures of the ETH Domain institutions. Consequently, with appropriate governance and resources, these limitations can be overcome, and the ETH ORD program offers unique opportunities for closer integration of ORD S&Is in the domain.

[2] *Ibid.*

2. Overview of existing ORD services and infrastructures used in the ETH Domain

In this section, we provide an overview of services and infrastructures related to ORD that are currently in use in the ETH Domain. Solutions are grouped along the research data life cycle which describes the typical stages that data passes through in a research project: data management planning, acquisition, storage and annotation, data processing and analysis, data publication, preservation and reuse (Fig. 1). In addition, we also include a section on cross-cutting solutions, that is services and infrastructures that relate to more than one stage in the data life cycle. The information contained in this section was obtained through a research process carried out by members of EG-SI in their respective institutions. The experts leveraged their own expertise and experience in the field, in addition to relying on information that is available publically or internally in the institutions. Furthermore, staff responsible for operation of ORD S&Is were consulted, e.g. to obtain usage statistics or operational details. As a consequence of this general approach, the focus in this section is on centrally provided services. Other relevant infrastructures almost certainly exist in departments, institutes or even research groups. However, due to the difficulty of obtaining this information in a reasonable amount of time, it is not covered here. In addition, this section also covers some services and infrastructures which are only partially relevant to ORD. In these cases, we try to specify which fraction of the solution is indeed open.

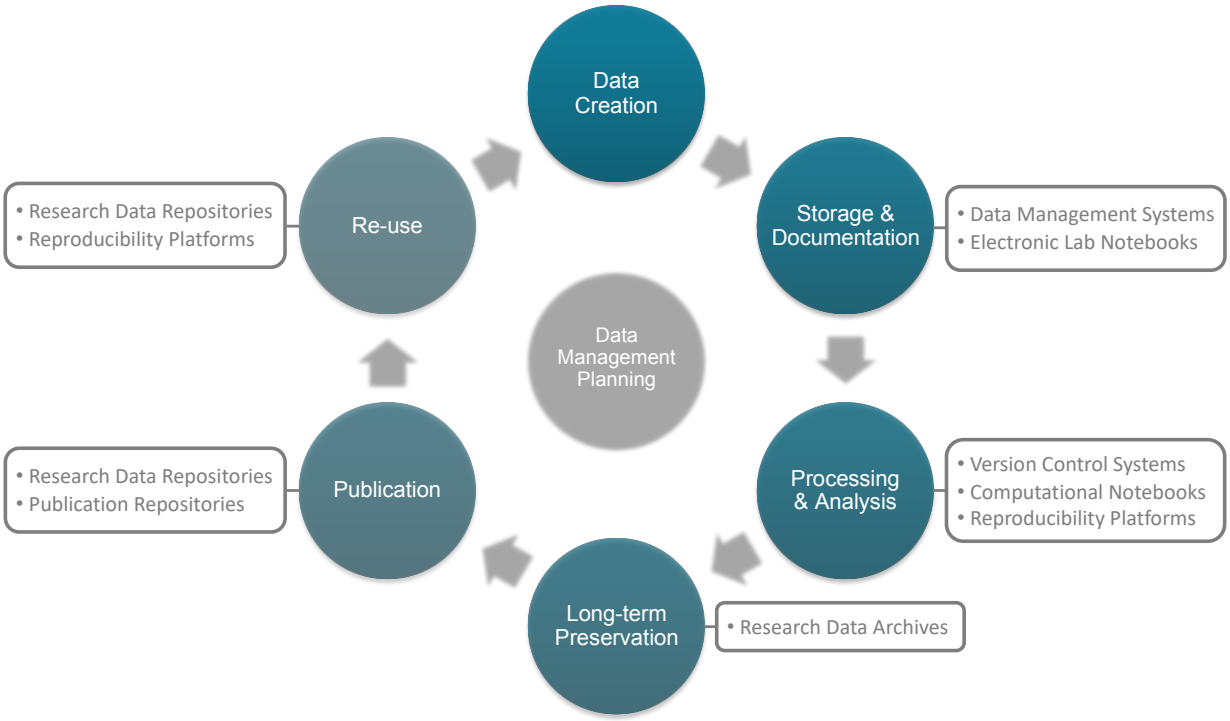

Figure 1. Outline of different stages in the research data life cycle together with some relevant services and infrastructures for each stage. Data management planning adresses all stages of the data life cycle and is therefore represented in the center. After creation, research data has to be properly stored and annotated (documented). Next, data is typically processed and analyzed, thereby creating derived datasets. Some or all of these datasets will eventually be selected for long-term preservation and publication in data repositories. Only a small subset of published datasets will ultimately be re-used, thereby restarting the data cycle for a new research project.

2.1. Data Management Planning

Mandatory Data Management Plans (DMPs) for grant applications were introduced by the Swiss National Science Foundation (SNSF) in 2017. Since then, most funding agencies require DMPs. In response to this,

most institutions in the ETH Domain have established solutions and services for assisting researchers in drafting DMPs. For example, DMP consulting and / or review services are available at EAWAG, EPFL, ETHZ and WSL. These services are typically provided centrally by library and / or (scientific) IT staff, or in WSL's case by EnviDat. Individual DMP consulting is labor-intensive and does not easily scale. It must take into account the institutional framework of infrastructure, services, and rules in each institution. Therefore, DMP consulting is normally not accessible to researchers from other institutions.

An alternative to individual DMP consulting is the creation, curation and distribution of DMP templates. DMP templates (usually for SNSF grant applications) are available at all ETH Domain institutions and are also becoming more common within research groups, institutes or departments. Some RDM infrastructure and service providers in the ETH Domain also provide specific DMP templates for their users (e.g. EnviDat, Materials Cloud / AiiDA, openBIS).

DMPonline is a web-based platform provided by the Digital Curation Centre (DCC) of the Universities of Edinburgh and Glasgow to research institutions in order for institutions to create, review, and share Data Management Plans (DMPs) that meet both institutional and funder requirements. It contains the latest funder templates and best practice guidelines to support users to create high quality DMPs. Moreover, within the tool there is the possibility to add custom guidance and example answers. The growing list of public DMPs published by other users can also be consulted off the tool for inspiration. EPFL is currently conducting a pilot project using this platform. This is a pilot phase so no figures about usage can be given yet. However, DMPOnline is used by a large number of research institutions worldwide. Furthermore, the EPFL Library also has created an online and open source Data Management Cost Calculator, which allows researchers to evaluate the occurring costs for data management, based on EPFL service providers. The tool helps researchers to evaluate the expenses related to managing, storing and publishing data during a research project . Taking into account providers prices, total cost is calculated dynamically based on researchers inputs (data size, project duration, active storage solutions, ELN, Database, Data Repository…). As a result, researchers obtain a detailed exportable presenting their project costs.

Finally, on an individual basis, ETH Domain researchers sometimes use other open-source systems for DMP creation and management, such as the Data Stewardship Wizard (ds-wizard.org), DMPTool (dmptool.org) and RDMO (rdmorganiser.github.io).

In general, with more researchers gaining experience with the process of creating DMPs, we expect more research groups to set up their own data management strategies and / or templates, which will facilitate DMP creation for early career scientists. This could shift demand from general consulting and review to more specific individual questions. The solutions and requirements described in the DMPs have to be supported by relevant infrastructure and services. In the long term, a central DMP repository may help to forecast future needs and costs, because the requirements described in the DMPs have to be supported by relevant infrastructure and services, that must be planned in advance. This resource planning would benefit from more actionable DMPs feeding information directly into the respective repository along with other tools relying on information from DMPs.

2.2. Data Acquisition, Storage and Annotation

The landscape of solutions for data acquisition, storage and annotation currently in use within the ETH Domain is very diverse. As a common denominator, all institutions offer some kind of basic storage service to their researchers. Examples include different tiers of NAS storage, long-term storage (on tapes) and high-performance storage systems (e.g., parallel file systems on HPC clusters). Apart from centrally provided storage solutions, individual institutes or research groups frequently still operate their own storage infrastructures.

For managing data stored in these storage infrastructures, researchers make use of different approaches and tools (see Table 1 for a general overview). Conventional file-and-folder based data management is still very common at all ETH Domain institutions. Besides, a number of specialized solutions for data acquisition and

storage are in use. At EPFL, three main ELN systems are offered in a central manner: First, Slims is offered as a life-science oriented Laboratory Information Management System (LIMS) and Electronic Laboratory Notebook (ELN) solution that is locally hosted on EPFL servers. Today there are about 70 laboratories using Slims, mainly in the schools of Life Sciences and Engineering. In addition, a specific ELN solution for chemical sciences is also available. Eln.epfl.ch is an Electronic Laboratory Notebook as well as a repository for spectroscopic data. Third, RSpace is used namely in chemistry. Many laboratories use the free version; thus, it is difficult to have exact usage numbers for this part. The full (paid) version of RSpace is currently being evaluated as an on-premise installation in a lab management system pilot within the School of Engineering in 4 laboratories.

For secure data acquisition, RedCap is used at EPFL as well as other ETH Domain institutions and is currently offered on an on-demand basis. Finally, EPFL also hosts the Open Sample platform that allows scientists to search for antibodies, plasmids, cells or any other biomedical research tool. It allows a search on using the Research Resource Identifier (RRID, a universal identifier) and this gives information as to whether this material has been used at EPFL and the contact details of the associated laboratory.

At ETH Zurich, the Scientific IT Services (SIS) develop and support the openBIS software for storage, annotation and backup of research data. openBIS is a combined data management solution, an ELN as well as a LIMS. openBIS is a web-based client-server application providing a graphical user interface (GUI) and APIs for system integration. The application was originally developed for managing large amounts of life-science data, such as high content screening, proteomics and sequencing data. However, as the underlying data structure is generic, openBIS has more recently also been adopted in other quantitative research disciplines, such as environmental and material sciences. Like other ELNs and RDM systems, openBIS supports FAIR data and ORD by providing researchers with a tool for professional documentation and annotation of research data. In addition, openBIS provides interfaces for publication of research data in data repositories (currently ETH Research Collection and Zenodo).

At ETH Zurich, openBIS is currently used by approximately 70 research groups in 12 different departments. The growth rate is about 5 – 10 new groups per year. The number of active users per openBIS instance is highly variable (Mean: 76 users, Span: 8 – 1262 users) as is the amount of data stored (Mean: 825 GB; Span: 200 MB – 14 TB, excluding archived data). Outside of ETH Zurich, a number of research institutions and groups use openBIS as part of the openRDM.swiss service, offered by SIS. Within the ETH Domain, Empa has started an institute-wide roll-out of openBIS as data management system in all labs. Furthermore, a few research groups at PSI have recently started to use openBIS.

At PSI, a number of solutions are currently used for data acquisition, storage and annotation. This includes different ELNs (openBIS, Biovia ELN, ELOG and SciLog), the LIMS Limsophy as well as Data Catalog for raw and metadata capture at acquisition.

As shown above, the widespread adoption of ELNs in academic and industrial research has led to a huge “zoo” of available ELN software solutions. However, the different ELN solutions are not interoperable with each other and this situation should be improved. The absence of standardized protocols for exchanging data between ELNs is creating a large hurdle for data transfer or migration from one system to another.

2.3. Data Processing and Analysis

Version Control

Version control is a key technique for proper management of software code and text-based data that forms the basis for most data processing and analysis in the ETH Domain. Git is currently the most popular version control system for professional software development. It is widely used in computational research fields where coding of simulation or analytical workflows is an integral part of the research process. In addition to enabling professional code management and reproducible research, Git-based platforms such as GitLab (open-source) or GitHub also support ORD by providing researchers with a platform for publication of code and data. In account of the increasing popularity of Git-based platforms in the research community, GitLab services are provided by several ETH Domain institutions (EPFL, ETHZ, PSI, WSL, SDSC). As an example, the central GitLab instance provided by ETH Zurich IT Services currently has approximately 14'500 users (with 23'700 different projects, of which 1'500 are public) requiring about 6 TB of storage space (yearly doubling for the last 5 years). Besides the central GitLab instances, researchers in the ETH Domain also use the SWITCH GitLab service, commercial platforms (GitHub) as well as locally managed GitLab servers.

Computational Notebooks

Computational notebooks, like Jupyter notebooks or R Markdown, have become extremely popular in many research disciplines and are widely used throughout the ETH Domain for interactive scientific computing. Besides the convenience of notebooks for exploratory scientific data analysis, they also support ORD as they allow easy sharing of a "computational narrative" (code, documentation, results etc.). ETH Domain researchers mostly use computational notebooks locally (i.e., on their own computers) or within a narrow group (i.e., JupyterHub server for a research group). At ETH Zurich, for example, openBIS servers may be integrated with a JupyterHub installation for data analysis. Some researchers also use external notebook services such as RStudio.io or mybinder.org. A number of central notebook platforms have been setup in different institutions of the ETH Domain, mainly based on JupyterHub. Noto is a JupyterHub service operated at EPFL which offers pre-configured compute kernels for a number of different programming languages. Unlike comparable institutional solutions, the Noto platform is also accessible to users outside of EPFL through SwitchAAI. At CSCS, a JupyterHub platform for interactive computing on the supercomputer Piz Daint is available to CSCS users. Besides standard Python kernels, the CSCS JupyterHub service also supports R and Julia as programming languages. A JupyterHub instance with support for Python, R, Octave, and Julia kernels has been deployed centrally on WSL's HPC since 2019. Further JupyterHub servers are currently in use at ETHZ, PSI, WSL and, most likely, numerous research groups within the ETH Domain.

Platforms

Within the ETH domain, there are significant on-going efforts to combine several technologies related to data processing and analysis into larger platforms that try to lower the barrier-to-entry to current best-practices.

Over the last years, SDSC has developed the Renku / Renkulab platform, which combines many of the aforementioned tools in order to create a unified solution for reproducible and collaborative data analysis. Renkulab.io is a free service that can be used by researchers anywhere in the world. For example, the Renku platform itself is fully based on git so everything is versioned by default. The hosted part of Renku is based on Docker containers, but it is not assumed that users are familiar with Docker or containers in general. One of the most common ways of using Renku is through the hosted interactive computational sessions, either using Jupyter notebooks or other front-ends. Interactive sessions are used in scientific collaborations and in teaching. Data scientists and course instructors can set up projects or courses that run in the hosted environment such that the collaborators and course participants can get access to a fully configured interactive session with a single click. Computational notebooks are often used in this context, but a large fraction of courses actually rely on hosted RStudio sessions or virtual desktop environments (VNC) where desktop apps for e.g., imaging applications can be used straight from the browser. In addition to the interactive sessions, an important component of Renku offers users the ability to record the provenance of their research results as they work;

they can use the recorded workflows and pipelines to reproduce or reuse data and methods. The information about data use is preserved in a searchable knowledge graph and can be annotated with custom controlled vocabularies, depending on the application. The flexible nature of Renku's metadata model and its plugin-based expansion system makes it simple to integrate with other sources of data and metadata and serve as a unifying layer between various providers and consumers. Because Renku projects are simply git repositories, users can bring them to any computational resource that fits their needs; the Renku command-line interface enables, for example, users to execute their recorded workflows on HPC clusters. Several deployments of Renku exist, the biggest of which is the public (free) instance at https://renkulab.io. This instance has been live since September 2018 and currently hosts ~4000 users, roughly doubling every year. Of those, there are between 50-200 active interactive sessions at any given time. The instance consumes approximately 1TB of RAM, 350 cpu cores, 20TB of object storage (data and container registry) and 40TB of active storage. Instances of Renku are also deployed at the EPFL School of Life Sciences, at the University of Fribourg and the Lucerne University of Applied Science. Discussions are on-going for a Renku instance at the EPFL Imaging Center and a Renku deployment at ETHZ is currently in the pilot phase.

In many disciplines, simulations and data analysis require the use of supercomputers (such as institutional clusters or the supercomputers at CSCS). For these use cases (particularly common for Materials Science computational research, but widespread to many more disciplines), AiiDA has been developed at EPFL. AiiDA is an open-source Python infrastructure to help researchers with automating, managing, persisting, sharing and reproducing advanced simulation workflows. AiiDA is designed to support high-throughput computations lasting from seconds to weeks. It natively interfaces with remote computation resources, taking care of submitting simulations to job schedulers, monitoring them, retrieving and parsing the results upon their completion, and launching the next workflow steps. Provenance of any simulation managed by AiiDA is stored in a provenance graph and can be queried without knowledge of database languages such as SQL. Provenance graphs can be exported in AiiDA archive files, that can be published and shared in research data repositories (e.g., Materials Cloud Archive, which natively hosts AiiDA archives and also provides interfaces to Renkulab). The AiiDA plugin registry provides a comprehensive set of plugins to support over 110 different simulation codes, and over 130 workflows. In order to facilitate submission of new workflows and analysis of the resulting datasets, AiiDA provides a native Jupyter-based interface, AiiDAlab, that provides researchers with a dedicated and intuitive simulation environment, working directly in the cloud or on remote or local resources. By providing simple GUIs to AiiDA workflows, AiiDAlab enables experimental researchers to get access to advanced computational capabilities via tailored lightweight web applications and provides guided input selection focusing only on the physical inputs, removing the need to set up numerical parameters and simulation details. AiiDAlab is provided both as a hosted service for researchers affiliated with NCCR MARVEL and related European projects (AiiDAlab has ~80 registered users on the server hosted at CSCS and ~110 on the open access Demo server. 18 Users are from Empa with ~2Tb of data on the server. The number of Empa users is expected to double within the next three to five years with two new Empa laboratories adopting the platform). Furthermore, detailed documentation is provided to help re-deploying the same platform in any other infrastructure (bare-metal servers, virtual machines, or scalable Kubernetes clusters).

Combining computational platforms with data management systems for reproducible analysis of large datasets is still a challenge. Recently, this challenge has been addressed with the development of a Reproducible Research Platform (RRP) which aims to support researchers from experimental labs to achieve collaborative and reproducible quantitative analysis of research data stored with openBIS (see above). RRP relies on established open-source tools such as Git for code management, repo2docker for reproducible computing environments, JupyterLab for interactive computational notebooks as well as openBIS for professional data storage and annotation. RRP has been developed by ETH Zurich Scientific IT Services in collaboration with scientists at the Department of Biosystems Science and Engineering and is currently available as an advanced prototype for pilot use cases.

2.4. Data Publication and Reuse

Data publication is of central importance for the success of ORD. However, we believe that data publication is not an end goal in itself, but should serve to foster the reuse of research data. Nevertheless, for data to be widely reusable, a minimum criterion is that they be publicly identifiable and accessible. For small to medium size datasets (e.g., up to 100s of GB), this is typically achieved by publication in a data repository. Currently, the landscape of research data repositories is very diverse (see this SNF study for a Swiss perspective). The repository landscape in the ETH Domain is summarized in Table 3.

Publication Repositories

All institutions within the ETH Domain provide an institutional publication repository to their researchers. These repositories typically make available secondary data including working papers, journal articles or doctoral theses. While research data are normally not included in publication repositories, there may be exceptions, for example in the case of supplementary materials.

The four research institutes of the ETH Domain share the platform DORA 4RI (Digital Object Repository at the Four Research Institutes) as institutional repository and bibliography for all research articles and other publications of their researchers. DORA is based on the open-source software framework Islandora. Currently about 75'600 publications are recorded in DORA, most of which have full text documents available and about half are Open Access (further statistics can be found here).

Infoscience is the EPFL institutional repository, maintained by EPFL Library. By the end of 2022, it contained about 169'000 publications records including 70'000 with full-text (articles, EPFL doctoral and master theses, working papers, conference proceedings, small datasets). Infoscience's current version is based on Invenio technology, provided by TIND. A new version will be released in summer 2023, based on the open-source software Dspace CRIS provided by 4Science. A Digital Object Identifier (DOI) is assigned on demand, through the ETH Zurich DOI Desk (see below: persistent identifier).

Finally, the Research Collection is operated by the ETH Library as institutional repository for publications and research data at ETH Zurich. The Research Collection is based on the open-source software DSpace and by the end of 2022 contained about 241'000 publication items (82'000 with full-text, 71'000 Open Access).

Research Data Repositories

Research data repositories can be classified as general or domain-specific data repositories. General purpose repositories usually accept research outputs of all types without subject-specific focus. At ETH Zurich, the Research Collection is operated as an institutional general-purpose data repository by the ETH Library. All research data in the Research Collection is assigned a Digital Object Identifier (DOI, see below). Researchers can specify data access rights (including automatic embargo handling) as well as retention periods (10 years, 15 years, indefinitely). Other features include content preview for ZIP- and TAR-containers, up-to-date download statistics, connection with ETH Data Archive for secure long-term preservation of most research data (see below) as well as integration with internal and external systems (e.g. openBIS).

The Research Collection by the end of 2022 contained about 1'700 research data items (about 1'400 are Open Access). There are about 5'500 active users publishing ca. 42 TB of data in total. It can safely be assumed that usage of the Research Collection for research data will grow considerably over the next years as more researchers get used to making their data available. Eliminating restrictions on file and upload sizes will broaden the scope of use cases. For example, a solution for the deposit of large files up to the TB-range in a local cloud with only metadata being held in DSpace was launched as a service in 2022 (Libdrive, included in the numbers for the Research Collection). The Research Collection plays a strategic role in ETH Zurich's mission to further increase the share of FAIR data in its research output. The Research Collection is, however, not meant to compete with subject-specific repositories. Researchers are free to choose the most suitable repository as long as it can be considered FAIR. A major advantage of the Research Collection is its integration with numerous systems inside and outside of ETH Zurich. This level of integration prevents a shared use of

the system together with other institutions. To achieve the same level of integration, institutions would need their own installations.

At EPFL, in 2021 a project was mandated by the Vice Presidency for Academic Affairs to build a data dissemination tool, with the objectives to provide a centralized access point to EPFL academic outputs (publications, data, code and metadata); to promote, disseminate and share EPFL academic outputs while enforcing FAIR principles through interoperability and connectivity with other tools. Upon analysis of the Request for Information (RFI) responses, the project Steering Committee decided to abandon the acquisition of a commercial platform as it would not meet EPFL researchers' expectations. It was also decided that the "new version" of Infoscience (see above) would incorporate small datasets (<5TB) and data set description (metadata and bibliographic references).

Currently, Zenodo is the generic research data repository most widely used by EPFL researchers. This data repository has been built and operated by CERN and OpenAire, with data stored in the CERN Data Centre. EPFL researchers can contribute their published datasets within the principal Zenodo EPFL Community, which currently contains 296 publications (95% in Open Access). In addition, 20 EPFL Zenodo communities also exist. In addition, the use of Zenodo for long-term retention and findability of code is suggested by Github, which is the code sharing platform most adopted by EPFL researchers. A Zenodo Curation policy has been added to the EPFL Community in order to increase ORD and interoperability.

Outside of EPFL, Zenodo and other public research data repositories are probably widely used by researchers in the ETH Domain for data publication. Similar to other Git-based platforms (see above), SDSC's Renkulab platform may be used for data publication via *Renku datasets*. Renku datasets are fully searchable and can be made easily reusable across different projects. Currently, Renku datasets do not receive a DOI for identification but they can be imported and exported from / to external repositories (current integrations include Zenodo, OLOS and Dataverse).

The research institutes of the ETH Domain operate a number of research data repositories. The Eawag Research Data Institutional Collection (ERIC) is a repository specifically for archiving and disseminating research data produced by Eawag scientists. ERIC is comprised of two distinct parts: an internal repository - ERIC-internal, and an open data portal - ERIC/open. The separation between ERIC-internal and ERIC/open enables metadata-rich archiving of sensitive personal data in accordance with Eawag's data privacy guidelines. ERIC-internal is a repository in which all data produced by Eawag scientists are preserved for at least the time-span required by scientific integrity requirements, and in some cases the data are planned to be held in perpetuity, for example environmental time-series. ERIC-internal has ~400 users and currently contains ~600 data packages totaling ~19Tb. ERIC/open is the Open Research Data publishing platform of Eawag. All data packages in ERIC/open are registered in the DOI system with a DOI obtained from DataCite, of which Eawag is a Direct Member. ERIC/open currently contains ca. 190 data packages from 87 research groups. Eawag currently does not track downloads from ERIC/open. Both, ERIC-internal and ERIC/open are based on the open-source system CKAN (ckan.org).

Another prominent data portal in the ETH Domain is the Environmental Data portal (EnviDat) by WSL. EnviDat is an institutional repository specialized on environmental research data, similar to Eawag's ERIC. It hosts and publishes with DOIs environmental research data from Switzerland and all over the world. The data is being provided by researchers of the many research units of WSL, as well as by cooperation partners from other institutions, including from the ETH domain (e.g., EPFL, ETHZ, Eawag, PSI – limited to environmental data). Researchers have the possibility to restrict the access to datasets in EnviDat, due to different reasons, such as embargo periods or manual data usage tracking. EnviDat has not implemented detailed tracking of its users, downloads and views. The repository is based on CKAN and currently contains more than 500 datasets from 80 research groups, including WSL research partner organizations, such as institutes from EPFL and ETH Zurich. About half of the datasets qualify for being included in opendata.swiss as fully open governmental data (OGD). EnviDat has currently more than 650 registered users and more than 2'000 additional unregistered data consumers per month. EnviDat has published around 20 TB of data, with the existing repository limited currently to a maximum of 30TB of available storage. Requests for more than 50 TB of additional unpublished

data have been already registered by EnviDat support. Consequently, it can safely be assumed that the size of the published research data will grow considerably over the next years and plans have been started for increasing the available EnviDat repository size until the end of 2024.

At PSI, the Data Catalog is a central data catalog for storing research data. Data and metadata collected by both internal and external users at the Photon Science facilities SLS and SwissFEL are automatically deposited in the catalog at a rate of 3 – 4 PB/year. At its core it uses SciCat, an open-source software developed in collaboration with the European Spallation Source and the Swedish MAXIV synchrotron. Each dataset is uniquely identified with a globally unique persistent identifier, with published datasets assigned a DOI. The data are tagged with searchable metadata and can be archived in a Petabyte Archive System. The Petabyte Archive system is responsible for packaging, archiving and retrieving the datasets within a tape-based long-term storage system located at CSCS.

Finally, several domain-specific data repositories are provided by research groups at EPFL. For example, Materials Cloud is a platform for seamless sharing and dissemination of resources in computational materials science. It provides curated high-quality data in the Materials Cloud Discover section, that is accessible via the standard OPTIMADE REST API agreed upon by the Materials Science community. Furthermore, it also offers a raw-data repository, Materials Cloud Archive, that hosts research datasets from the field of materials science, is open to the world and assigns DOIs to each of them. Materials Cloud Archive is currently recommended as a discipline-specific repository (for Materials Science) by SNSF, the EU Commission via the Open Research Europe journal and Nature's journal Scientific Data. Materials Cloud currently hosts over 22 million crystal structures, 7.5 million of which have associated density-functional-theory simulations, and with over 650'000 fully-reproducible simulations managed with AiiDA. Furthermore, Materials Cloud Archive is integrated with SDSC's RenkuLab: for any reproducible dataset produced with AiiDA (see before) and hosted on Materials Cloud Archive, a link is automatically displayed to open and inspect the data directly inside RenkuLab, for which a template tailored for AiiDA datasets has been developed.

EPFL Living Archives is an ongoing project developed by EPFL Architecture in collaboration with ENAC-IT4Research, harvesting Infoscience, to promote information and knowledge sharing via an online tool. This catalog aims to facilitate access to the outputs from the Architecture department. As such, Living Archives organize research and students works structured according to an open tag structure.

In summary, the availability of data repositories is increasing in the ETH Domain. However, the different solutions are not interoperable with each other and this situation should be improved. The absence of interoperable metadata schemas and APIs is creating a large hurdle for integrating the repositories with other systems and infrastructures that operate at different stages of the research data lifecycle.

Persistent Identifiers

Persistent identifiers (PIDs) are long-lasting references to digital resources. In the context of ORD, they are essential to reliably identify datasets, for example in repositories. PIDs are a way to identify a data artifact (file or data collection, for example) in a way independent from the physical place where it is located. In other words, a PID adds an indirection pointer between the data requester and the data physical object. Resolving a PID to provide the method to access the data object is done by an extremely reliable infrastructure that guarantees that the identifier is persistent and always translatable to the data object's current location. Most commonly, Digital Object Identifiers (DOIs) are currently used for this purpose.

The ETH Library runs the ETH Zurich DOI Desk. The DOI Desk is an official DOI registration service in the DataCite association. ETH Zurich's DOI Desk registers DOIs for primary data (research data) and for secondary data such as working papers, articles or doctoral theses. This service is available to all organizational units from Swiss higher education and research institutions but not to individuals. The basis for registration is an agreement with the DOI Desk that documents all technical and organizational matters. The DOI Desk is run by the ETH Library and integrated with DataCite's global infrastructure. ETH Zurich's IT Services provide technical infrastructure and support. DataCite is an association and official DOI registration agency within the International DOI Foundation (IDF). DOI registration is free for all ETH Zurich organizational

units. Other academic institutions and non-profit organizations must pay the fees defined by DataCite. The number of total DOIs registered until December 2022 is about 3'200'000 (ca. 300'000 new registrations in 2022). The DOI Desk currently provides the registration service to 77 institutional customers, of which 50 are not part of ETHZ (for example for WSL's EnviDat, Materials Cloud, or EPFL InfoScience).

DOIs may not be appropriate as identifiers for all kinds of research data. For this reason, alternative PID systems have been created, such as the eResearch Persistent Identifier Consortium (ePIC) consortium. Since 2018, CSCS provides and resolves ePIC PIDs for academic institutions in Switzerland. The ePIC consortium offers a service to create, manage, and resolve persistent identifiers. The increasing amount of research data, the variety of the usage profiles and the international exchange within different infrastructures demand to uniquely assign the data with a PID with a high degree of flexibility and robustness and ePIC offers a reliable mechanism to guarantee these features of persistent identifiers.

Sharing Solutions

Besides dedicated data repositories, researchers in the ETH Domain use a range of other services for sharing of research data. Cloud-based storage services are of course highly relevant in this respect. Most institutions in the ETH Domain have access to SWITCHdrive, a cloud storage service provided by SWITCH. An exception is ETH Zurich, which offers its own cloud storage service, Polybox, to its members. Polybox is operated by the central IT services and is very widely used by ETHZ members (ca. 41'500 users storing about 234 TB of data).

Commercial cloud storage services, such as DropBox, Google Drive or Microsoft OneDrive, are also widely used by ETH Domain researchers. EPFL for instance has a specific contract with Google as it is much used in research. ETH Zurich offers its researchers both Google and Microsoft cloud subscriptions (incl. storage). Regardless of the actual cloud storage provider, researchers typically use these services for convenience, with data being either kept private or restricted to a small group of colleagues / collaborators (privately shared). Public sharing of research data is relatively rare and does not conform to ORD and FAIR standards, as the data is not discoverable.

Finally, publishing and sharing of very large datasets (TBs to PBs and upwards) remains a significant challenge in several research disciplines of the ETH Domain (e.g., climate science, microscopy). Several institutions use the Globus fast data transfer system, managed by the University of Chicago, for efficient transfer of very large datasets. Within the ETH Domain, Globus endpoints have been established for example at CSCS, PSI and ETHZ. For the data repositories of the ETH Domain, however, publishing and sharing of very large datasets remains a significant challenge

2.5. Data Preservation and Disposal

Long-term preservation ensures that valuable research data, and by extension scientific results, remain interpretable and reusable for years to come. In general, one can distinguish curated and non-curated solutions. Curated solutions are often referred to as archives and are typically in the library domain. They are important for ORD as they are a part of professional data repositories to ensure the long-term accessibility and interpretability of digital objects. Non-curated long-term storage (e.g., tape libraries) typically does not ensure the interpretability of stored data. File formats may no longer be readable, data may be poorly described or not at all and even knowledge about the existence of data may fade with time due to a lack of accessible metadata. Non-curated long-term storage solutions, sometimes described as 'data graveyards', are therefore of limited relevance for ORD.

Curated long-term preservation solutions in the ETH Domain are typically attached to the data repositories (see above). For example, EnviDat (WSL) and Data Catalog (PSI) implement long-term preservation of published data. The ETH Data Archive is the preservation solution for research data at ETHZ, as it is also the preservation layer behind the Research Collection for items which should remain usable for more than ten years. Usually, data will be accessed in dedicated online platforms which are optimized for access and delivery, while data in the ETH Data Archive are not routinely accessed by end users ('dark archive'). Exceptions with

online access are several legacy items from before the Research Collection was established and open-source code packages which are archived on behalf of ETHZ technology transfer office. The latter are kept for reference purposes while regular re-use is expected to happen via GitLab or GitHub. The ETH Data Archive is operated by the ETH Library within the infrastructure of ETH Zurich's IT Services. It is based on the commercial digital preservation solution Rosetta (Ex Libris). By the end of 2022, the ETH Data Archive held close to 1'200 Research Data items, more than half of which were open-source code packages or other Open Access Research Data items which are not available from the Research Collection.

At EPFL, Academic Output Archive (ACOUA) is a new institutional archive for the long-term preservation of research data produced by EPFL researchers. ACOUA was launched in the first quarter of 2021 and is curated by the EPFL Library. While ACOUA currently hosts a modest number of datasets, the tool has the potential to connect with other infrastructures to preserve their data. Through strengthened collaborations between the Library and SCITAS and/or "EPFL Centres", it is envisioned that ACOUA's use will increase. The main features of the service are:

- Storage repository with long term retention capabilities (10 years)
- Compatible with the OAIS standard
- Data is securely archived with metadata
- Provide persistent identifiers
- Capability to export datasets to Zenodo
- Analytic reports and associated functionalities available to researchers
- Scalable service that could cope with larger volumes of data
- Data protection measures against hardware/software failures
- Clear licensing of the data

At CSCS, the Long Term Storage (LTS) service enables CSCS users to preserve their scientific data and ensures that it can be publicly accessed through a persistent identifier. The current implementation of the LTS service addresses the first two principles of the FAIR quadrant: findable and accessible. The main features of the service are:

- Storage repository with long term retention capabilities (10 years)
- Provide persistent identifiers
- Ability to set public access to data when needed
- Data stored in LTS easily accessible from a web browser (HTTP protocol)
- RESTful API to integrate with third party applications/portals
- Scalable service that can cope with large volumes of data
- Resiliency due to data protection measures against hardware/software failures
- Clear licensing of the data

2.6. Cross-cutting solutions

A number of data services and infrastructures are relevant across all or multiple stages of the research data life cycle. We refer to these as "cross-cutting solutions". Examples include general storage and computing infrastructure, identity management as well as training and consulting services. In this section we focus on RDM / ORD training and services currently available in the ETH Domain as well as a short discussion of identity management approaches.

Training and Consulting

Availability of high-quality RDM trainings and consulting for researchers is as important as providing relevant services and infrastructures. This is recognized by the ETH ORD program, which dedicates a separate measure to the improvement and alignment of existing RDM trainings in the ETH Domain (Measure 3). At both ETHZ and EPFL, general RDM trainings covering the whole data life cycle are organized by the libraries. For

example, at ETHZ a series of RDM-related workshops is organized twice per year by. In these workshops, the basic concepts of FAIR and ORD are introduced and approaches for implementing them are discussed. The main partners are ETHZ Scientific IT Services (SIS) and groups from ETH Library. Apart from regular, general workshops, RDM courses are organized on demand and adapted to the needs of the respective research group. The general RDM workshops at ETHZ are not advertised for external participants, but they are usually open. As they refer to internal services and infrastructures, ETHZ researchers usually benefit the most and are given priority. In 2022, regular workshops had between 7 and 23 participants, including a small number from Empa, PSI and WSL.

Since 2019, the ETH RDM Summer School has been a credited one-week course aiming at early career scientists (doctoral students, post-docs). ORD is a relevant part of the week, but the overall scope is broader and mainly discusses RDM, which is considered a necessary condition for ORD. The Summer School is organized by the ETH Library with lecturers from numerous services at ETH Zurich. The RDM Summer School is open to participants from the ETH Domain. In 2022, 25 participants joined the on-site edition of the Summer School, of which 8 had their main affiliation with EPFL, 2 with Eawag, 2 with Empa, 1 with PSI and 1 with WSL. General RDM training events are complemented by offers on more specific topics, e.g., methods in computing and programming from Scientific IT Services or on licensing issues by ETH Library.

At EPFL, the library offers a comprehensive suite of on-demand trainings and workshops, and regularly scheduled trainings, which encompass the various steps of the research lifecycle. These programs primarily target doctoral students and the EPFL research community at large. The scope of the trainings goes beyond the ORD paradigm. For example: RDM crash course, Data and Ethics, Data and Software Carpentry, RDM credited courses for PHD, RDM and DMP, How to publish my data, Good coding practices, Metadata in the research activities. In addition, in collaboration with the OS Unit and faculties, the library periodically participates in Summer schools on Open science.

At the research institutes of the ETH Domain, RDM and ORD trainings are currently offered sporadically or focused on specific tools. At Empa, trainings on version control, programming and reproducible analysis are offered. WSL organizes EnviDat trainings and information days on demand. The EnviDat team is also in regular contact with the research units and groups, for offering individualized training and support for individual scientists that use EnviDat. For other RDM trainings, WSL is recommending its employees to participate in the courses organized by ETHZ - both ETH Library and C2SM (Center for Climate System Modelling) at ETHZ. At Eawag, the IT department has been organizing various courses for basic data handling and programming skills, such as Git, Bash and Python courses. Finally, SDSC offers trainings on ORD best-practices and related concerns, such as version control, containerization and continuous integration.

Identity and access management

Access management and control is an essential part of ORD. All the above-mentioned platforms and tools are based on their identity management systems. Data repositories, for example, typically allow download of datasets without restrictions. Nevertheless, data deposition is only allowed for a restricted group of users, for example members of the institution which operates the repository. Similar restrictions are in place for most of the above-mentioned solutions, with some notable exceptions (e.g. Renkulab platform). Providing access to ORD solutions for members of other institutions and universities requires use of federated authentication and authorization mechanisms. For example, at EPFL most of the tools use a central authentication mechanism which can be combined with Switch AAI federated authentication and authorization mechanisms to allow other Swiss schools and universities to use such tools, if required. However, this excludes users which are not affiliated to Switch AAI and especially non-Swiss organizations. Ultimately, enabling widespread accessibility to ORD solutions across the ETH Domain entails more than just resolving technical hurdles such as the wider implementation of federated authentication and authorization systems. It necessitates a comprehensive evaluation of administrative and financial factors, as the provision of services or infrastructure by one institution to another demands careful consideration.

3. Conclusion

This report provides, to our knowledge, the first comprehensive overview of services and infrastructures related to ORD and RDM in the ETH Domain. It uses the “lens” of the research data life cycle to identify existing ORD solutions at different stages of the research process as well as to highlight apparent gaps and limitations, while keeping the researcher at the center. Even though the data lifecycle admittedly represents a simplification of the true complexities of data management in a research project, the authors believe that it nevertheless provides a useful metaphor for grouping together similar ORD solutions and identifying gaps.

Overall, the report identifies an impressive portfolio of state-of-the-art services and infrastructures related to ORD in the ETH Domain. In general, researchers in the ETH Domain should have access to services and infrastructures that are very competitive, both on the national and international level. Moreover, several dedicated ORD solutions that have been developed in the ETH Domain are now in operation and recognized both on a national and international level, highlighting the innovative potential of the ETH Domain with respect to ORD. Nevertheless, the report also highlights a landscape of services and infrastructures that is rather fragmented. This fragmentation reflects organizational and budgetary realities, but is undesirable from the perspective of the ETH Domain as a whole and of its research communities. Some services are not well known to potential users and related services sometimes lack easy interconnections. Possible synergies between institutions are not exploited consequently and there is a risk of duplicating efforts. The overall, rich landscape can benefit from an even better alignment with the requirements and expectations of the research community.

In this respect, the report identifies a number of technical as well as non-technical challenges when it comes to ORD solutions.

Regarding technical challenges, one element that has been coming up regularly is the ability to closer integration of existing ORD solutions, for example the lack of open and documented APIs to facilitate interoperability. Even for services and infrastructures with APIs, interoperability remains challenging for average users, for example due to different data models or access controls.

It has to be emphasized that multiple non-technical challenges are as important as the technical challenges in the context of the wider research ecosystem and are therefore addressed by other Measures in the ETH ORD Program. Examples include proper governance and policies, adoption / buy-in by researchers (Measure 1), legal aspects (Measure 4) and the availability of suitable career paths (Measure 5). Furthermore, a plethora of different financial aspects as well as the established institutional boundaries and “ways of thinking” may impede the adoption of ORD practices. While some of these challenges are tackled in other measures of the ETH ORD program, it has to be kept in mind that these aspects are in fact codependent and cannot be solved adequately in isolation.

The group is convinced that given appropriate governance and resources, these technical and non-technical limitations can be overcome. The ETH ORD program and in particular its Measure 2 offer unique opportunities in this respect. The members of the EG SI are committed to leverage this opportunity in a responsible and transparent manner to achieve a closer integration of S&I in the ETH Domain and consequently foster ORD in the domain. While the members of the EG SI are committed to contribute in different areas as required, ultimate success will also depend on other factors outside the group’s immediate control.

4. Overview tables

Solution name	Main institution	Other institutions	Type	Domain	Selected statistics
Files & folders	All	N/A	Data management	All	This is the 'default' option in most research groups.
openBIS	ETHZ	Empa, PSI, Eawag	Data management, ELN, LIMS	Quantitative Sciences	≈ 70 labs at ETHZ 13 labs at Empa Pilots at PSI & Eawag
Slims	EPFL	Unknown	ELN, LIMS	Life Sciences	≈ 70 labs at EPFL
Eln.epfl.ch	EPFL	None	ELN	Chemical Sciences	N/A
RSpace	EPFL	Unknown	ELN	Chemical Sciences	Users of free version unknown On-premise pilot in 4 labs
RedCap	EPFL	ETHZ	Secure data acquisition system	Clinical Sciences	N/A
Biovia ELN	PSI	None	ELN	Life Sciences	PSI BIO division
ELOG	PSI	Empa	Electronic Logbook		N/A
SciLog	PSI	None	ELN		N/A
Limsophy	PSI	None	LIMS	Scientific laboratories	N/A

Table 1: Overview of commonly used and supported solutions for data acquisition, storage and annotation in the ETH Domain.

Solution name	Institutions	Functionality	Users	Storage
GitLab	EPFL, ETHZ, PSI, WSL	Version control	Dependent on installation (e.g. 14'500 for ETHZ GitLab)	Dependent on installation (e.g. 6 TB for ETHZ GitLab)
JupyterHub	EPFL, ETHZ, CSCS, WSL, PSI	Interactive comp. notebooks	Dependent on installation	Dependent on installation
Renku / RenkuLab	SDSC, EPFL	Version control Interactive comp. notebooks Workflow management Provenance tracking Reproducibility	≈4000 for public instance	≈60 TB for public instance
AiiDA / AiiDAlab	EPFL, Empa, PSI	Interactive comp. notebooks Workflow management Provenance tracking Reproducibility	≈ 190	≈20 TB / year

Table 2: Overview of commonly used and supported solutions for data processing and analysis in the ETH Domain.

Repository name (with link)	Hosting institution	Other data-providing institutions[3]	Repository type	Software	PID type	Selected statistics[4]	Remarks
Digital Object Repository at the Four Research Institutes (DORA)	Lib4RI	Eawag, Empa, PSI, WSL	Publication	Islandora	DOI	≈ 75'600 publications	Further statistics can be found here
Infoscience	EPFL	None	Publication	Invenio	DOI	≈ 162'000 publications	Next release to include datasets
Eawag Research Data Collection (ERIC)	Eawag	None	General data	CKAN	DOI	≈ 150 open datasets ≈ 500 internal datasets	
ETH Research Collection	ETHZ	None	Publication General data	DSpace	DOI	≈ 241'000 publications ≈1'700 datasets (42 TB total volume)	
Data Catalog	PSI	Facility users, CSCS (see remarks)	General data	SciCat	DOI, PID	> 400'000 datasets, 9 PB total volume, >1'600 groups of users	Active proposal to provide broader access to ETH institutions
Zenodo	CERN	Open	General data	Invenio	DOI		Some institutional customization, e.g. via "EPFL groups"
EnviDat	WSL	Collaborations approved by WSL (see remarks)	Domain-specific (Environmental Sciences)	CKAN	DOI	≈ 540 datasets (20 TB total volume)	Currently hosting **environmental datasets** from WSL and other collaborating institutions
Materials Cloud	EPFL	PSI,Empa	Domain-specific data (Materials Sciences)	Invenio (customized)	DOI	≈ 22M crystal structures ≈ 7.5M simulations	
Living Archives	EPFL	None	Domain-specific data (Architecture)	In-house	PID	≈ 11'000 items	

Table 3: Overview of commonly used repositories in the ETH Domain for publication of research data and outputs.

[3] ETH Domain only

[4] Numbers are not directly comparable. Publication repositories, for example, contain various scientific outputs such as reports, theses and paper publications. Some publication repositories (e.g. DORA) do not contain any research datasets and are therefore of limited relevance to ORD.