Töwe, M and Barillari, C. 2020. Who Does What? – Research Data Management at ETH Zurich. *Data Science Journal,* 19: 36, pp. 1–6. DOI: https://doi.org/10.5334/dsj-2020-036

PRACTICE PAPER

Who Does What? – Research Data Management at ETH Zurich

Matthias Töwe[1] and Caterina Barillari[2]

[1] ETH Library, ETH Zurich, Zurich, CH

[2] Scientific IT Services, ETH Zurich, Zurich, CH

Corresponding author: Matthias Töwe (matthias.toewe@library.ethz.ch)

We present the approach to Research Data Management (RDM) support for researchers taken at ETH Zurich. Overall requirements are governed by institutional guidelines for Research Integrity, funders' regulations, and legal obligations. The ETH approach is based on the distinction of three phases along the research data life-cycle: 1. Data Management Planning; 2. Active RDM; 3. Data Publication and Preservation. Two ETH units, namely the Scientific IT Services and the ETH Library, provide support for different aspects of these phases, building on their respective competencies. They jointly offer trainings, consulting, information, and materials for the first phase.

The second phase deals with data which is in current use in active research projects. Scientific IT Services provide their own platform, *openBIS*, for keeping track of raw, processed and analysed data, in addition to organising samples, materials, and scientific procedures.

ETH Library operates solutions for the third phase within the infrastructure of ETH Zurich's central IT Services. The *Research Collection* is the institutional repository for research output including Research Data, Open Access publications, and ETH Zurich's bibliography.

Keywords: Research Data Management; Data sharing; University Services; Data Management Tools

Introduction

Like all universities, ETH Zurich needs to address increasing requirements in Research Data Management (RDM). Some of these are driven by continuous development of methods in science and technology. Others arise from the necessity to carry forward well-established standards of Good Scientific Practice into mostly digital workflows. Further requirements are motivated by funders' concern of increasing the availability of Open Data to facilitate their re-use and thus better exploit the potential value of publicly funded data for knowledge generation.

ETH Zurich, as a technical university, has its focus on science and technology. While several services for RDM support for researchers have been in place for years, they were not initially integrated over the full research data life-cycle. For the reasons outlined above and accelerated by changed funders' policies, the need to support the full data life-cycle has become more pressing only in recent years.

In 2019 numbers, ETH Zurich counted more than 22'000 students, of which more than 4'000 were doctoral students. The services described here serve researchers including doctoral students, post-docs, and other, more senior scientific staff. These groups add up to over 6'000 FTE. All of these persons are potential users of the services in their capacity as creators of research data, although their role may vary by responsibility, activity, and concrete needs. The services should also cater for the needs of all scientific disciplines represented in the 16 departments of ETH Zurich.

This paper outlines the current set of services and their integration, as well as the collaboration of the involved organisational units of ETH Zurich. Tools which are employed to provide the services are mentioned. As the current state is not intended to be static or final, its strengths and weaknesses are discussed with a view towards future development.

Discussion

Regulatory framework

The overall formal background for RDM is set by existing regulations on the level of the university and beyond. Guidelines for Research Integrity (ETH Zürich 2011) have long been in place. They outline a few general requirements on RDM and expect scientific communities to know or, where missing, establish standards which are internationally recognised as appropriate for their respective field. They also describe the responsibilities of principal investigators and other staff members.

Further legal obligations, e.g. concerning data protection, animal welfare, ethical aspects, intellectual property etc. must obviously also be observed.

In 2017, the principal public funder of research in Switzerland, the Swiss National Science Foundation, took a further step in pushing for Open Research Data, by requiring researchers to provide a Data Management Plan (DMP) with virtually all grant proposals (SNF 2017).

This regulatory framework governs all activities along the RDM life-cycle, but does not affect all stages of the cycle uniformly. Discipline-specific requirements, as well, result in a range of activities along the cycle which can differ considerably between fields of research. Researchers and support services for RDM in the university are therefore confronted with a wide range of demands. Consequently, a number of contact points and services are available depending on the actual regulatory questions, e.g. the Technology Transfer Office, Office for Research, Legal Services, Ethics Commission, and others.

The main actors providing practical data services at ETH Zurich are the Scientific IT Services (SIS) section of its central IT Services and the Research Data Management and Digital Curation Group at ETH Library. Both rely heavily on the IT infrastructure provided by central IT Services. For a simplified view of their activities and for communication purposes, the steps of the data life-cycle are condensed into three phases of the data management process: Data Management Planning, Active RDM, Publication and Preservation (**Figure 1**).

First phase: data management planning

The first phase focuses on Data Management Planning (DMP) and involves all inquiries and planning which need to be done by researchers to formulate a sound research proposal. This includes basic information needed for a DMP, such as which kind of data will be produced, what their volume will be, how they will be used, shared, published, and preserved. For these general demands, ETH Library and Scientific IT Services together offer introductory trainings and workshops on RDM twice a year. In addition, since 2019, the two sections also hold a yearly RDM summer school for PhD students and post-docs. This year, this event was held completely online over one week and proved very successful. Additional events are organised on demand for departments, institutes, or research groups with adaptations to the local situation. These are supplemented by 1 to 1 consulting by staff from both units contributing specific know-how and introducing their respective services. The Library offers this as part of its 'Book a Librarian' service.

In 2019, six half-day workshops on different aspects of RDM were offered and further nine more tailored trainings were held on demand for research groups or institutes. 37 consultations took place, sometimes followed by further requests.

The trainings aim at empowering researchers to ask themselves relevant questions and to assess and choose appropriate solutions suiting their needs. Ideally, individual researchers should also advance the discussion in their respective groups and in their wider community, so that in the longer run, they can achieve a wider agreement on best practices, which today only exists in some fields.

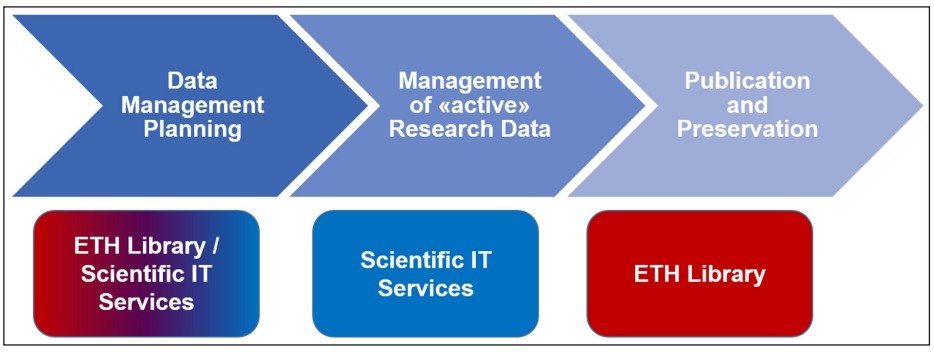

Figure 1: Consecutive phases of RDM and the respective service providers at ETH Zurich. Image by the authors.

Second phase: active research data management

Roughly speaking, the second phase comprises all activities requiring support during the course of a research project. They are summed up as Active RDM, signalling that data in this stage are regularly being worked on and processed. Active RDM is mainly supported by Scientific IT Services. This section of ETH Zurich's central IT Services provides research-related IT services for computing- and data-intensive research activities at ETH Zurich. These include, but are not limited to, the operation of central clusters for High Performance Computing and Big Data applications, support for software developed by research groups, software development on behalf of research groups, specialised processing and analysis pipelines, visualisation services for large data sets. Scientific IT Services also provide trainings in these fields of activity (ETH Zürich 2018a).

With respect to RDM in particular, Scientific IT Services develop and support their own Open Source platform, *openBIS*, for active data management in quantitative research.

openBIS is a web-based application that handles the storage, annotation and backup of data during the lifetime of a project. It was originally used as a solution for managing large amounts of life sciences data, such as the outputs of microscopy, high content screening, proteomics, metabolomics, and sequencing, as well as the derived data from processing and analysis. However, in recent years the development has been more focused on usage as electronic laboratory notebook in mid-size laboratories to document experiments and annotate data, as well as an inventory of lab materials and protocols (Barillari et al. 2016; ETH Zürich 2017) (**Figure 2**). Due to its generic underlying structure, its usage has now expanded to other quantitative research disciplines, such as environmental sciences, materials sciences, physics and chemistry.

At ETH Zurich, SIS provides *openBIS* as a service to research groups. This includes running and maintaining the *openBIS* software on central IT ETH infrastructure, as well as providing user training and support. Over 40 research groups in different research areas and a few facilities at ETH Zurich currently use *openBIS* for managing their research data.

It should be noted, that there is, however, no requirement for research groups at ETH to use an ELN or even *openBIS*, and this is not expected to change in the foreseeable future.

For certain use cases in the phase of Active RDM, the Library can provide the simple editor and viewer *docuteam packer* (Docuteam 2018) for structuring and describing files locally with a view for later submission for archiving. The tool is not a data management solution in itself, but rather meant for locally organising file-based research data with proper context information on the file and folder levels. *docuteam packer* is an Open Source tool that was developed by the private company Docuteam mainly for the archival sector. ETH Library commissioned a number of enhancements for the tool to better cater for research data and to better integrate with existing tools at ETH Library, like the ETH Data Archive or the University Archives' information system. *docuteam packer* is currently used by few research groups with special requirements and by ETH Zurich's University Archives as part of their workflow for describing and submitting ETH Zurich's harvested webpages to the ETH Data Archive.

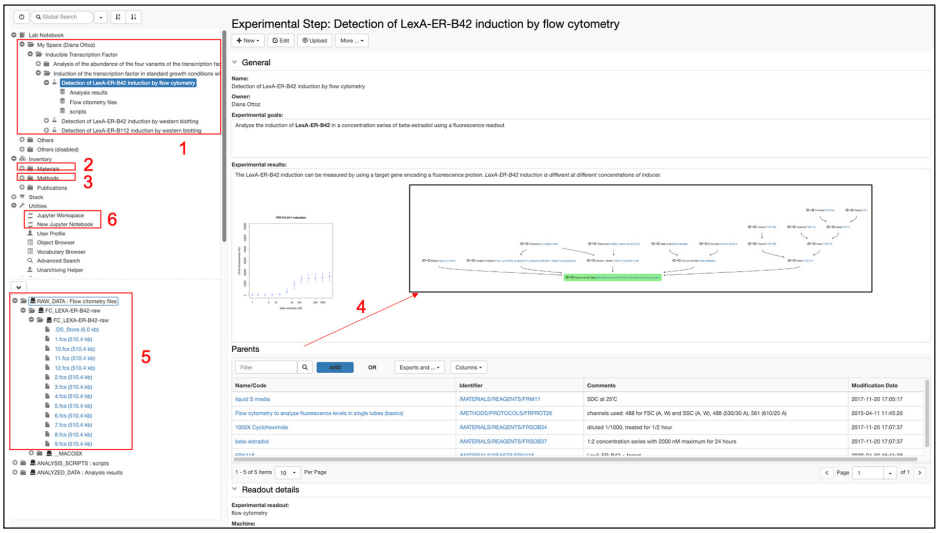

Figure 2: *openBIS* overview. In the Lab Notebook section, each member of a lab has a folder to organize projects and experiments (1). Laboratory samples, materials and methods can be organised in the Inventory (2, 3), from where they can be linked to experimental descriptions (4). Data can be uploaded via the web interface or directly from measuring instruments (5). Data stored in *openBIS* can be analysed with the Jupyter notebooks (6) (Kluyver 2016). Image by the authors.

Third phase: publication and preservation

The third phase comprises the activities of sharing, publishing, and preserving research data and is supported mainly by ETH Library. To publish and preserve research data with relevant context information, such information must be documented from early on, when the data are created. To encourage researchers to document their work and data from the start, the Library engages in training and supporting researchers from the very early stages of the data life-cycle, as discussed above.

On the technical level, in addition to the *docuteam packer* tool mentioned above, research data services by ETH Library today mainly rely on two applications: *Research Collection* and *ETH Data Archive*. These are operated by the Library on infrastructure of ETH Zurich's central IT Services.

The *Research Collection* (ETH Library 2017) was launched in mid-2017 as the Repository for ETH Zurich's research output using the Open Source repository software *DSpace* (Duraspace 2018) in a tailored implementation. It comprises the formerly separate services of ETH Zurich's Open Access Repository and its University Bibliography, which feeds into the academic reporting and at the same time powers publication lists on researchers' websites. Its new third role is the one of a Research Data Repository.

Data sets can be submitted either as Supplementary Material to formal publications or they can be deposited separately as self-contained entities. This makes the *Research Collection* a one-stop shop for a range of tasks related to publishing.

Researchers can publish data immediately under Creative Commons Licenses, with an Embargo Period or with more restricted access rights. This latter option is mainly intended for data that are considered unsuitable for broad distribution, but are shared selectively or should only be safely archived for answering possible challenges to published data or articles.

The *Research Collection* and *openBIS* have recently been integrated. Members of a research group using *openBIS* for their Active RDM can easily select in the *openBIS* interface data to send directly to the *Research Collection*. The handling of large files still poses some challenges and the need to move around such files should be minimised. From a user perspective, it is not useful to offer large files of several hundreds of GB online for immediate download via the browser, so alternative paths of access must be established. By the end of 2019, the *Research Collection* held 544 research data objects.

All the content of the *Research Collection* is exported to the *ETH Data Archive* for long-term preservation for defined retention periods of 10 or 15 years or for permanent retention. The *ETH Data Archive* uses the commercial software *Rosetta* (Ex Libris 2018) and serves as the backbone for digital preservation at ETH Zurich.

Currently, data submitted via *docuteam packer* is also sent directly to the *ETH Data Archive*. In the future, this workflow will be reconsidered and might be switched to submitting data to the *Research Collection* instead. This would have the advantage of reducing the number of different processes for both users and service providers.

The vast majority of the *ETH Data Archive's* content so far consists of master files from digitisation projects of ETH Library. This is due to the fact that there is no obligation for researchers to deposit their data in the *Research Collection* which, in addition, has only been available since mid-2017. Furthermore, digitisation data are continuously processed and ingested to the *ETH Data Archive* automatically and in large quantities, whereas research data so far had to be uploaded manually. To avoid this, integration between *openBIS* and the *Research Collection* is advancing as described above.

Beyond the technical implementation, ETH Library also provides consulting on questions of Open Access publishing, RDM, DMPs, and preservation which might come up when researchers use the *Research Collection*. Depending on researchers' concrete needs, they may also be redirected to Scientific IT Services or other specialist units such as ETH transfer, the technology transfer office. These services are of course also available to members of ETH Zurich who do not intend to use the repository.

Communication

Distinguishing the three phases described above has proven so far a powerful approach also from the point of view of communication, as it illustrates that RDM consists of several different tasks. However, given the number of services and tools available and the different units involved at different stages of the process, researchers might find it challenging to find their way through available offers for their specific needs. To facilitate orientation, a common website was established in 2017 by ETH Scientific IT Services and ETH Library (ETH Zürich 2018b). This does not provide exhaustive guidance, but rather contains basic information and guides researchers to relevant contacts and services within ETH Zurich. Currently, a common mailbox is also in operation.

Conclusion

Despite the lack of long-term experience with the overall set-up of RDM services at ETH Zurich, some weaknesses and desirable developments are already obvious. On the level of policies, ETH Zurich is still in the process of issuing a dedicated Research Data Policy which is not too prescriptive in narrowing down the choice of scientific methods and is equally applicable to all disciplines. This clearly affirms ETH Zurich's commitment and its expectations towards its members. A commitment of the university itself is of course necessary to provide the required resources to support the existing and future services. Moreover, the mere cost of storage has come back into focus, because cost per Terabyte is no longer decreasing as fast as data volumes increase in certain fields. There is no sustainable solution to address this discrepancy and eventually, communities might be forced to agree on more restrictive practices of what data to keep and for how long. RDM can support this by distinguishing data of only temporary relevance from essential data according to such accepted community criteria and thus providing a reliable basis for informed decisions on data deletion, temporary retention, or preservation.

The current set-up does not serve all disciplines equally well and it would be very optimistic to expect this in the near future. The existing data management platform *openBIS* has its focus on quantitative scientific disciplines. In any case, research groups at ETH Zurich can freely decide to use the tools they consider as most useful for them. It is of course advisable that groups make sure that such tools are supported either by central IT Services or by services on the departmental level. It should not be expected that the use of any, or even of one particular system, will be enforced in the future.

Requirements in RDM, research practice, and tools keep evolving rapidly and services at ETH Zurich must adapt to new developments and needs over time. The current approach of rather loosely coupled services might have advantages in flexibly addressing those from different angles, but time must tell if there are major disadvantages outweighing this.

Noticeably, the challenges cannot be addressed by institutional services alone. The units involved in RDM at ETH Zurich also participated in a major collaborative effort on the national level, in a project on Research Data Life-Cycle Management (DLCM) (DLCM 2017), funded by swissuniversities. This has facilitated the exchange between Higher Education Institutions in Switzerland and both existing and new solutions have made considerable progress towards sustainable service provision. In 2018, two follow up projects were funded by swissuniversities: openRDM.swiss, aimed at establishing a national RDM service for academic institutions based on *openBIS*; DLCM2, aimed at providing a Swiss national data repository and establishing a national training platform. SIS has also recently submitted a project proposal to the EGI (European Grid Infrastructure) Foundation, with the aim to expand its RDM services to the European level.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

Both authors worked on the text as a whole. Originally, C. Barillari contributed the majority of the content on phase two (Active RDM) and M. Töwe contributed the majority of the content on phase three (publication and preservation).

References

Barillari, C, et al. 2016. openBIS ELN-LIMS: an open-source database for academic laboratories. *Bioinformatics*, 32(4): 638–640. DOI: https://doi.org/10.1093/bioinformatics/btv606

Data Life-Cycle Management. 2017. The DLCM Project. Geneva, Switzerland: DLCM. Available at https://www.dlcm.ch/about-us/dlcm-project [Last accessed 7 August 2020].

Docuteam. 2018. Software – Our tools for digital archives. Baden-Dättwil, Switzerland: Docuteam GmbH. Available at https://www.docuteam.ch/en/products/it-for-archives/software/ [Last accessed 7 August 2020].

Duraspace. 2018. About DSpace. Beaverton (OR), United States of America: Duraspace. Available at https://duraspace.org/dspace/about/ [Last accessed 7 August 2020].

Eidgenössische Technische Hochschule Zürich. 2011. *Richtlinien für Integrität in der Forschung – Guidelines for Research Integrity*. Zurich, Switzerland: ETH Zürich. DOI: https://doi.org/10.3929/ethz-b-000179298 [Last accessed 7 August 2020].

Eidgenössische Technische Hochschule Zürich. 2017. openBIS ELN-LIMS – Features. Available at https://openbis.ch/ [Last accessed 7 August 2020].

Eidgenössische Technische Hochschule Zürich. 2018a. IT in Research – Your IT support for research. Zurich, Switzerland: ETH Zürich. Available at https://ethz.ch/services/en/it-services/it-in-research.html [Last accessed 7 August 2020].

Eidgenössische Technische Hochschule Zürich. 2018b. Research Data at ETH Zurich. Zurich, Switzerland: ETH Zürich. Available at https://www.ethz.ch/researchdata [Last accessed 7 August 2020].

ETH Library, Eidgenössische Technische Hochschule Zürich. 2017. Research Collection. Zurich, Switzerland: ETH Library. Available at https://www.research-collection.ethz.ch/ [Last accessed 7 August 2020].

Ex Libris Ltd. 2018. Rosetta – Preserve your digital assets for the future. Jerusalem, Israel: Ex Libris. Available at https://www.exlibrisgroup.com/products/rosetta-digital-asset-management-and-preservation/ [Last accessed 7 August 2020].

Kluyver, T, et al. 2016. Jupyter Notebooks – a publishing format for reproducible computational workflows. In: Loizides, F and Schmidt, B (eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, 87–90. Amsterdam, Netherlands: IOS Press. DOI: http://doi.org/10.3233/978-1-61499-649-1-87

Swiss National Science Foundation. 2017. Open Research Data. Bern, Switzerland: SNF. Available at http://www.snf.ch/en/theSNSF/research-policies/open_research_data/ [Last accessed 7 August 2020].

How to cite this article: Töwe, M and Barillari, C. 2020. Who Does What? – Research Data Management at ETH Zurich. *Data Science Journal,* 19: 36, pp. 1–6. DOI: https://doi.org/10.5334/dsj-2020-036

Submitted: 25 June 2018 **Accepted:** 01 September 2020 **Published:** 22 September 2020

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS