Supporting Scientists with FAIR Research Data Management – Initiatives at ETH Zurich and beyond

Caterina Barillari*, Henry Lütcke**, Ana Sesartic Petrus***, Matthias Töwe[a],
Thomas Wüst**, Bernd Rinn**

Abstract

The FAIR (Findable, Accessible, Interoperable, Reusable) guiding principles for research data management have been established to advance data-driven science and are by now widely adopted among journals, funders and institutions. In order to comply with the FAIR principles, researchers must ensure that data are correctly documented and stored throughout the lifetime of a project. However, dealing with the ever-increasing amount and complexity of data and information generated in research labs requires new data management skills that researchers need to acquire at the beginning of their career. To this end, the commit-

* ETH Zurich, Scientific IT Services, Mattenstrasse 26, 4058 Basel.
Corresponding author

E-mail: *caterina.barillari@id.ethz.ch*

Caterina Barillari, PhD, graduated in Chemistry and Pharmaceutical Technology at the University of Milan in 2001. After a short experience as medicinal chemist in the private sector, she pursued a PhD in Computational Chemistry at the University of Southampton, UK, studying the effects of water molecules in proteins' binding sites. She then held a Marie Curie fellowship in chemoinformatics at the CNRS in Strasbourg, France, and she later worked as chemoinformatician at the Institute of Cancer Research in Sutton, UK. Caterina Barillari joined ETH Zurich in 2011, where she currently holds a position as Manager of Research Data Management Services in the Scientific IT Service of ETH.

** ETH Zurich, Scientific IT Services, Weinbergstr. 11, 8092 Zurich.

E-mail: *hluetcke@ethz.ch*

Henry Lütcke, Dr. rer. nat., is a Scientific Computing and Research Data Management specialist at Scientific IT Services of ETH Zurich. He was trained as a neuroscientist at the Universities of Edinburgh (UK) and Göttingen (Germany). After a postdoc at the University of Zurich, he joined SIS in 2014. Besides supporting researchers in different aspects of data management and scientific computing, he is also project manager of the national openRDM.swiss project.

E-mail: *twuest@ethz.ch*
https://sis.id.ethz.ch/
ORCID: 0000-0001-6901-9277

Thomas Wüst, Dr. phil.-nat., is head of the group Computational & Data Science Support in Scientific IT Services (SIS) of ETH Zurich. He studied physics at ETH Zurich and did a PhD in computational physics at the University of Bern. Before joining SIS in 2014, he held positions as a computational scientist and high-performance computing specialist at the Swiss Federal Institute for Forest, Snow and Landscape Research (WSL) and as a postdoctoral researcher at the Center for Simulational Physics of the University of Georgia, USA.

ment of the academic institutions and funding agencies is of critical importance so that scientists can be provided with tools, education and support for data management. Here we present the ongoing efforts in this area at ETH Zurich, where the Scientific IT Services and the ETH Library support scientists with tools and services in data management. Similar efforts at the national level, with initiatives funded by swissuniversities, are also presented.

1. Introduction

Over the last decade, scientific research in many disciplines has become increasingly data-driven. By implementing novel data science approaches, such as artificial intelligence and machine learning, research processes are expected to become more efficient and effective. To fully leverage the power of data science approaches across different disciplines, research data must be accessible and interpretable, ideally in machine-readable formats. To this end, the FAIR guiding principles for scientific data management and stewardship have been established[1]. The FAIR principles[2] outline key attributes and properties to make digital research outputs more **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable. The principles focus on machine-actionability to enable automated computational exploitation of large and complex datasets. The FAIR principles have been globally endorsed and are widely adopted by journals, funders and institutions.

In order to comply with the FAIR principles, researchers must ensure that data are correctly documented and stored throughout the lifetime of a project. Howver, with the ever-increasing amount and complexity of data and information generated in scientific research labs, most of the traditional approaches to Research Data Management (RDM), based on file systems, spreadsheets and paper notebooks are no longer sustainable. Thus, management of research data is becoming a complex task that requires key skill sets that researchers should now acquire at the beginning of their career.

In most academic institutions, principal investigators are responsible for the data produced in their groups,

---

1 Wilkinson, M. D. *et al.*: The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018 (2016).
2 *https://www.go-fair.org/fair-principles/*

but often there are no precise guidelines in place for managing the data. Providing tools, education and support for RDM at the institutional level is thus becoming a necessity. At ETH Zurich (ETH), the executive board has recognized this need and has invested in initiatives for providing RDM services to ETH scientists, with the aim to facilitate the publication of data that adhere to the FAIR data principles. To cover the complete data life cycle, i.e. the different stages of data generation and re-use, this work is primarily led by two units: Scientific IT Services (ITS SIS[3]) and ETH Library.[4] RDM services include adequate planning of which data will be generated and how they will be handled, data acquisition, data analysis, long-term data preservation and data sharing. While ITS SIS help and support scientists with RDM activities during an ongoing research project, ETH Library supports the later stages of data publication, long-term preservation and sharing. Both units also provide joint consulting on RDM planning as well as training.

The need for improving RDM expertise is recognized also on a national level in Switzerland. From 2015–2018, swissuniversities, the umbrella organisation of Swiss Higher Education Institutions, established the national project DLCM (Data Life Cycle Management),[5] to organize and promote the development of tools and services for FAIR RDM in Swiss academic institutions. ETH Zurich, with ITS SIS and ETH Library, has been a partner in this project. Based on the success and the recognized importance of the DLCM project, two successor projects have been approved in 2018, one on active RDM (openRDM.swiss, led by ETH ITS SIS) and one on a national data repository (DLCM Phase 2, led by University of Geneva IT).

In this article we give an overview of services at ETH Zurich for RDM support along the stages of the research data life cycle with some conclusions for these functions in universities in general.

2. Overview of RDM services at ETH Zurich

At ETH Zurich, ITS SIS and ETH Library provide tools and services to help researchers with RDM during the different stages of the data life cycle (see Fig 1). Both units provide joint training and consulting in the early stages of a research project, with respect to data management planning. ITS SIS provides tools, services and trainings in active RDM, i.e. the daily management of data generated throughout a research project, from initial data acquisition to data processing and analysis. ETH Library provides tools, services and trainings

for the later stages of data publication and data sharing. Moreover, both units run joint trainings in RDM aimed at students and research staff.

3. Data management planning

In order to successfully manage research data, adequate planning of methodologies and resources is essential. Data management plans (DMPs) are now

E-mail: *brinn@ethz.ch*
https://sis.id.ethz.ch/

Bernd Rinn, Dr. rer. nat., is head of Scientific IT Services of ETH Zurich and a group leader at the Swiss Institute of Bioinformatics. He received his PhD in theoretical physical from the University of Konstanz. After finishing his PhD, he joined Genedata AG in Basel. End of 2006, he joined the Center for Information Sciences and Databases (C-ISD) at the Department of Biosystems Science and Engineering of ETH Zurich. There, he focused his work on designing and building the research data management system "openBIS", which is widely used as a FAIR data management solution by research labs and projects today. From 2009 to 2013, Bernd Rinn led C-ISD. In 2013, he joined the Executive Board of ETH IT Services to establish the new section Scientific IT Services (SIS). SIS focuses on enabling compute- and data-intensive research at ETH Zurich and covers a wide range of scientific infrastructures and services, including high-performance computing infrastructure, scientific software engineering services, data science services for research projects, FAIR research data management, workflow solutions, as well as consulting and training on these topics.

*** No postal address available.

E-mail: *ana.petrus@alumni.ethz.ch*
https://www.linkedin.com/in/anapetrus/
ORCID: 0000-0002-0928-8894

Ana Sesartic Petrus, Dr. sc. ETH, works as a freelance Sustainability Consultant. She studied Environmental Sciences at ETH Zurich and performed her doctoral studies at the Institute for Atmospheric and Climate Science until 2011. Postdoc positions in Berne (Institute of Geography, 2011/2012) and at the Center for Climate Systems Modeling at ETH Zurich (C2SM, 2012/2013) brought her to the Terrestrial Systems Ecology Group at ETH in 2013. From 2015 to 2020 she was strongly engaged in the creation of the services of the Research Data Management and Digital Curation group at ETH Library, where she acted as Data Management Consultant.

[a] ETH Zurich, ETH Library, Rämistrasse 101, 8092 Zurich.

E-mail: *matthias.toewe@library.ethz.ch*
https://www.library.ethz.ch/FDM
ORCID: 0000-0001-9671-083X

Matthias Töwe, Dr. phil. nat., is head of the group Research Data Management and Digital Curation at ETH Library. After his studies in chemistry in Hamburg and his doctoral studies at the University of Basel until 2001, he was trained as a Scientific Librarian at the University Library Basel. Since 2003 he has been working for ETH Library, first for the Consortium of Swiss Academic Libraries in its project E-Archiving on digital long-term preservation, later as coordinator for the programme Swiss Electronic Library – e-lib.ch. Since 2010 he has been engaged in the implementation of the ETH Data Archive and in the creation of RDM services.

---

[3] *https://sis.id.ethz.ch/*
[4] *http://www.library.ethz.ch/en/*
[5] *https://dlcm.ch/*

DM Planning   Active DM   Publishing & Preservation

ETH Library
Scientific IT Services
• Training & Consulting

Scientific IT Services
• Training & Consulting
• Soft- & Hardware
• Services

ETH Library
• Training & Consulting
• Publications services
• Preservation services
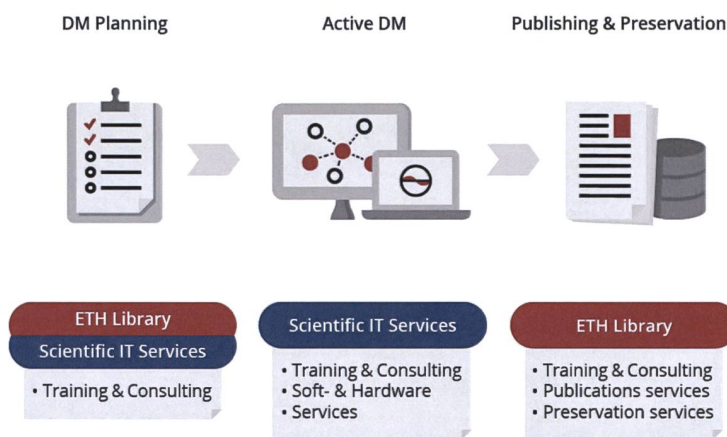
Figure 1. Stages of data life cycle support at ETH Zurich. Data management (DM) planning is supported by both library and IT services, the active DM phase is covered by the Scientific IT services and, finally, the publishing and preservation are supported by ETH Library.

increasingly requested by most funding agencies. In Switzerland, the Swiss National Science Foundation (SNSF) made DMPs mandatory for grant applications from October 2017. Horizon 2020 has required a DMP since January 2017. In the context of the DLCM project, the libraries of ETH and EPFL have elaborated a DMP template document with explanations and example answers to all questions of the SNSF DMP document, available to all Swiss scientists.[6] ETH Library subsequently adapted this template to ETH specifications, for use by ETH scientists. In addition, ITS SIS provide a partially pre-filled version of the DMP for users of the ETH active RDM service to cover the non-project specific parts of the template. This has been greatly welcomed by researchers who had to fill in DMPs for their grant applications. Furthermore, ETH Library and ITS SIS regularly provide consultancy to researchers on DMP writing (see Fig 2), a service that is highly valued by scientists.

4. Active Research Data Management
In order to publish research data in accordance with the FAIR data principles, data should be treated as FAIR from the start of a research project. This implies that data need to be annotated, organized and backed up from initial acquisition through to final results. Moreover, all steps and procedures to obtain the data must be well documented. Typically, researchers rely on data organization using a defined folder structure and file naming convention, possibly adding README files with data annotations if needed. Experimental and computational procedures recorded on paper or in a digital notebook are linked to the data stored on the file system. All other relevant information should be appropriately stored and referenced, as well. While

this approach may work well for individual researchers managing modest amounts of data, it does not scale well to very large and complex datasets as well as big research consortia. For these cases, RDM platforms have been developed that combine a digital notebook with a sample and protocol inventory. The main advantage of such platforms over traditional file-based approaches is that all data and related information is stored (or at least catalogued) in one central location with a management layer for tracking and retrieval on top.

Since 2007, ITS SIS have developed an open-source software solution for active RDM: openBIS.[7] The solution covers storage, annotation and backup of data during the lifetime of a project. openBIS is a client-server application with a strong focus on providing an easy user interface, and well-designed APIs for system integration. The standard user client is web-based. The application was originally developed for managing large amounts of life-science data, such as microscopy, high content screening, proteomics, metabolomics, sequencing, as well as their derived processed and analysed data. In recent years the development has been focused on adding electronic laboratory notebook capabilities to document experiments and annotate data of labs of any size.

Furthermore, openBIS has been integrated with Jupyter notebooks and the JupyterHub server to allow direct analysis of data managed by the system.[8] Analysis notebooks are connected to these data when they are stored in the database, thereby ensuring provenance tracking of the analysis results. In addition, openBIS can be used as an inventory of lab materials and protocols. Due to its generic underlying data structures, the usage of openBIS is now expanding to other quantitative research disciplines, such as environmental and material sciences. ITS SIS are working closely with these scientific communities to understand their needs and, when needed, adapt the software to accommodate them.

In accordance with the FAIR principles, all items in openBIS get a unique identifier, and all entities are indexed and searchable via the web-interface and the API. For use in the life sciences, a default metadata model is provided that has been developed with the user commu-

6 See *https://www.dlcm.ch/resources/dlcm-dmp*

7 Bauch, A. *et al.*: openBIS: a flexible framework for managing and analyzing complex data in biology research. BMC Bioinformatics 12, 468 (2011). Barillari, C. *et al.*: openBIS ELN-LIMS: an open-source database for academic laboratories. Bioinformatics 32, 638–640 (2016). *https://openbis.ch/*.

8 Kluyver, T. *et al.*: J. D. T. Jupyter Notebooks – a publishing format for reproducible computational workflows. Position. Power Acad. Publ. Play. Agents Agendas 87–90 (2016). doi:10.3233/978-1-61499-649-1-87
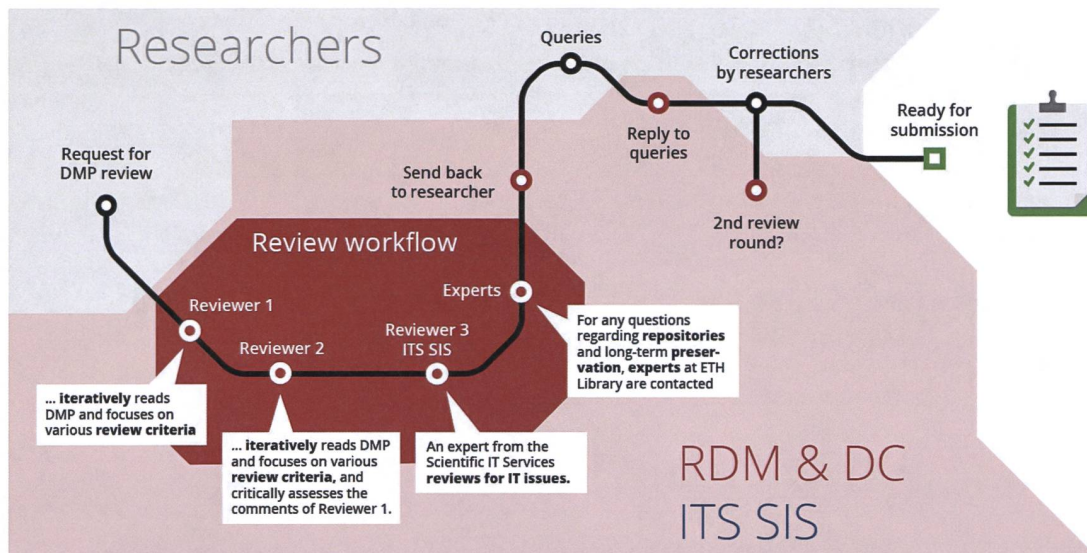
Figure 2. DMP review workflow at ETH Zurich, covered by the Scientific IT services (ITS SIS) and the RDM and Digital Curation (RDM & DC) team of ETH Library.

nity. Users can modify the defaults and customize the metadata information according to their needs. ITS SIS are collaborating with other scientific communities at ETH Zurich for implementing default metadata fields in openBIS for non-life-sciences disciplines too.

openBIS requires authentication and every operation performed by any user is registered in the database and is thus traceable. All data and metadata stored in openBIS can be accessed by authorized users with sufficient rights via HTTPS and/or SFTP via built-in servers. It is possible to delete bulk data without deleting the corresponding metadata. Moreover, even when metadata are deleted in the user interface, they are still stored in the database in 'history tables', and can be retrieved from there if need be. Data and data models can be annotated using native openBIS mechanisms and semantic annotations and they can be retrieved via the JSON-RPC API.

Adoption of a proper RDM solution in a research lab is time-consuming, implies a steep learning curve and requires a change in established procedures and work routines. To ease the process and increase adoption, training and support from technical experts with scientific knowledge and understanding are fundamental. For this reason, ITS SIS provide openBIS as a service to ETH research groups in two ways: The *ETH Research Data Hub* is available to all ETH research groups and provides a low entrance barrier (e.g. no extra cost). As an alternative for groups who need a more customized solution, a research group can also have their group-private openBIS installation, fully maintained by ITS SIS. In both cases, initial training and support for customization is provided to users and continuous support is offered afterwards.

5. Data publication and long-term preservation

One of the central roles of ETH Library is the support of the processes for publication and preservation of scientific data. RDM is a key prerequisite for effective data sharing and publication, which, in turn, can increase the visibility of scholarly work and its citation rates.[9] Managing research data throughout their lifecycle is furthermore key to efficient long-term preservation, because the latter must rely on technical, administrative and rights metadata, as well as sufficient context information being available to make sure that data remain usable and understandable in the long run.[10]

ETH Library provides several tools for the publication and long-term preservation of data. The Research Collection is a FAIR repository for publications and data acknowledged by funding agencies, where ETH researchers can share and publish their work and/or their data free of charge.[11]

In accordance with the FAIR principles, every item in the Research Collection receives a DOI. The rich metadata used ensure that data are findable, interoperable and reusable.[12] Data and metadata are accessible via the OAI-PMH protocol.[13] When data are deleted, the metadata are still accessible. It is possible to determine access rights for datasets and the desired Creative Commons license can be selected.[14]

9 Piwowar, H. A., Day, R. S. & Fridsma, D. B.: Sharing detailed research data is associated with increased citation rate. PLoS One 2 (2007). Piwowar, H. A. & Vision, T. J.: Data reuse and the open data citation advantage. PerrJ Comput. Sci. 25 (2013). doi:10.7717/peerj.175

10 Sesartic, A. & Töwe, M.: Research Data Services at ETH-Bibliothek. IFLA J. 42, 284–291 (2016).

11 *https://www.research-collection.ethz.ch/*

12 *https://documentation.library.ethz.ch/display/RC/Metadata+Schema*

13 *https://documentation.library.ethz.ch/display/RC/OAI-PMH+interface*

14 *https://documentation.library.ethz.ch/display/RC/Access+rights)*

The ETH Data Archive on the other hand, is ETH Zurich's long-term preservation solution for digital information, such as research data, documents, software, images.[15] All data that are uploaded to the Research Collection are automatically preserved in the Data Archive and additional tools exist for archiving data from other sources.

IT SIS and ETH Library also provide a close integration between openBIS and the Research Collection. This allows researchers to use openBIS in their daily work for annotation and processing data and finally select what should be published at the end of a research project. Selected objects can be automatically transferred to the Research Collection for publication. This will allow scientists at ETH to ensure that their data are Findable, Accessible, Interoperable and Re-usable.

6. Overview of Swiss initiatives towards FAIR data

Between 2015 and 2018, the Swiss Conference of Higher Education Institutions ("swissuniversities") funded the national DLCM (Data Life Cycle Management) project, to organize and promote the development of tools and services for FAIR RDM in Swiss academic institutions. Several partners from academic institutions across Switzerland were involved, including ETH Zurich, with ITS SIS and ETH Library. The project had tracks covering the different aspects of the data life cycle. Each track saw collaborative work between some of the different partner institutions.

One track focused primarily on guidelines and policies and this working group elaborated the DMP template document described above. Another track focused on active RDM, and part of the work involved a market analysis of existing commercial and open-source solutions for active RDM.[16] This working group also prepared a guideline document on how to introduce an Electronic Laboratory Notebook (ELN) and Laboratory Information Management System (LIMS) in a laboratory, available from the same webpage. A third track focused on the development of a Swiss national data repository platform that adheres to the FAIR data principles and will be available to all scientists working in Swiss higher academic institutions. Finally, a fourth track was dedicated to education and training. Several trainings in RDM were carried out at the different partner institutions during the project and both materials and expertise were collected and exchanged. As an example, the RDM – Educational Material Collection by ETH Zurich showcases presentations developed at ETH.[17]

The main aim of the DLCM project was to set the basis for establishing national services in RDM that can be self-sustainable, which is often challenging for academic services. As follow up of the DLCM project, two more projects aimed at establishing national services were funded in the 2018 – 2020 funding round by swissuniversities: openRDM.swiss and DLCM Phase 2.

openRDM.swiss is led by ITS SIS and aims to establish a service for providing the active RDM platform openBIS to the Swiss academic community.[18] Similar to the service provided at ETH, ITS SIS provide openBIS installations on the cloud infrastructure of the Swiss academic network provider SWITCH, and also provides user trainings and support. For institutions with more IT infrastructure and expertise, the project provides support and training to institutional IT administrators for setting up an on-premise installation of openBIS.

DLCM Phase 2 continues the development of the national data repository started in DLCM Phase 1 which is now being made available under the name OLOS.[19] Moreover, it also established a national platform for RDM training.

7. Conclusion

Several factors are key to successful RDM. First, general education in RDM is of fundamental importance, in order to raise awareness of the topic. It is essential to target trainings at different organizational levels, i.e. management, professors, staff and students. ITS SIS and ETH Library run joint workshops on RDM, that cover various topics such as ETH regulations, RDM planning, active RDM, data sharing, data publication and long-term preservation.[20] Events are scheduled throughout the year, but often workshops are organized on demand for specific research groups, institutes or departments. Since summer 2019, an ETH Summer School in RDM is also jointly run by the two units. In addition, both units organize separate workshops and trainings more specific to their areas of competence. Second, adopting a software solution or a new workflow for the daily management of data generated in a research laboratory often requires a change in working mode and is associated with a steep learning curve. Key to this is a strong commitment from the principal investigators, acknowledging and accepting that the process requires a time investment. This time investment is still often perceived as time subtracted from research, and as such not valuable. In fact, time spent on setting up a professional RDM solution should be considered as an investment for the future, since hav-

15 http://www.library.ethz.ch/en/ms/Research-Data-Management-and-Digital-Curation/ETH-Data-Archive
16 A list is available at https://www.dlcm.ch/services/dlcm-eln
17 https://www.research-collection.ethz.ch/handle/20.500.11850/296468
18 http://openrdm.swiss/
19 https://www.dlcm.ch/olos/about
20 https://www.ethz.ch/services/en/service/a-to-z/research-data/support-and-training.html

ing well annotated and organised data, easily searchable and retrievable, can save time in the longer run, i.e. when retrieving data for a publication or thesis.

Current efforts at ETH Zurich are aiming at overcoming this misconception. In our opinion, one of the main obstacles to the adoption of RDM solutions is that the benefits of using them are often not immediate, but become only evident with time. For this reason, we believe that having training and support from system experts can lower the barrier of adoption of such solutions, and reduce the time investment from the research labs, thus encouraging more people to use them. To this end, ITS SIS provide openBIS as a service to ETH research groups. This includes operating and maintaining the openBIS software on ETH infrastructure, as well as providing user training and support. Through the openRDM.swiss project, a similar service has now been established at the national level.

In addition, researchers should be supported with data publication and sharing, e.g. when making informed decisions on which data they can and should make available as FAIR data. ETH library has set up an extensive service portfolio in the field of open access and open data in order to support researchers in the open sharing of their articles and data.[21]

RDM is part of the quality assurance of research. Ideally, researchers should acquire the necessary skills during their training to empower them to implement appropriate RDM for their research projects. In the long run, this would be more sustainable than communicating RDM as an issue which is seemingly separate from the core scientific education. To this end, ETH Library, in collaboration with ITS SIS, will intensify its efforts to anchor the topic of RDM as early as possible in scientific careers and will also try to sensitize students to the topic. This includes offering a Summer School since 2019 that is rewarded with ECTS credits.[22] During the Summer School, young research-

ers from the ETH Domain get to know the principles of RDM along the data life cycle and deepen their knowledge in practical examples and exercises. ETH Library and ITS SIS also strive to develop advisory services that are specially tailored to the needs of project leaders. Based on experience and exchanges with colleagues, we know that RDM can only work if it is supported by the management level. According to the Guidelines for Integrity in Research of ETH Zurich23, it is the responsibility of research project leaders to ensure RDM within teams and that all team members are informed of what is expected from them.

Professional RDM is a means to an end and not a goal in itself. The aim must be to make it easier for researchers to access the methods and tools that can help them to carry out their tasks, as they are ultimately the experts on their own data. The main interest of researchers is to underpin excellent research with adequate RDM – and not vice versa. It is therefore central to mediate between the RDM requirements and the concrete needs of the researchers. This balancing act will challenge all those who support researchers in managing their data in the coming years.[24]

The advent of interdisciplinary, data-driven research powered by novel data science methods promises important new scientific insights. To fully leverage the potential of this approach, however, will require the challenging implementation of RDM practices according to the FAIR data principles. ETH Library and Scientific IT Services are committed to supporting researchers at ETH Zurich and beyond with this challenge.

Acknowledgements

21 *http://www.library.ethz.ch/en/ms/Open-access-at-ETH-Zurich*
22 *https://www.library.ethz.ch/en/ms/Conferences/ETH-Research-Data-Management-Summer-School-2020*

23 ETH Zürich. Guidelines for Research Integrity. 31 (2011).
24 Sesartic Petrus, A. & Töwe, M.: Forschungsdatenmanagement an der ETH Zürich: Ansätze und Wirkung. BIBLIOTHEK – Forschung und Praxis (2018). doi:10.18452/19633