Three years of publishing data in ETH Zurich's Research Collection: Lessons learned and new developments

Barbara Hirschmann
ETH Library
ETH Zurich
Zurich, Switzerland
http://orcid.org/0000-0003-0289-0345

Abstract – *In June 2020, ETH Zurich's Research Collection celebrated its third anniversary. The Research Collection serves as an institutional repository for ETH Zurich that can host both publications and research data and is operated by the E-Publishing team at the ETH Library. Publishing research data and advising customers on research-data-specific questions in the publishing workflow has emerged as a new field of activity for the team. With over 800 research data items published over the last few years, we have now gained a good understanding of the actual use cases for publishing data in an institutional repository at a large university for science and technology. We regularly talk to researchers about the incentives and requirements for publishing their data and monitor what kind of data they deposit. In this paper, we share and discuss our insights. We present statistics on the types of deposited datasets and explain how "FAIR" they are in terms of accessibility, licences and metadata. We also discuss our workflows for checking datasets for formal quality criteria and compliance with institutional policies and how to bridge publishing and preservation requirements in a research data repository. Finally, we give an overview of two ongoing development projects. The first one aims to enable ETH researchers to deposit datasets directly from the data management tool openBIS, while the second one will deliver a solution for publishing large datasets via the Research Collection.*

***Keywords* – institutional repository, data publishing, quality assurance.**

I. Introduction

This article gives an overview of the functionalities of ETH Zurich's institutional repository Research Collection. It focuses on the repository's features for data publishing and illustrates how ETH researchers have actually used the repository during the last three years. It also explains what type of checks and workflows the ETH Library has set up to assure formal quality and policy compliance of the deposited datasets. Lastly, we report on the status of two ongoing development projects that will expand the repository's capabilities for data publication.

II. The Research Collection

A. *Overview*

The Research Collection is ETH Zurich's repository for publications and research data. It hosts research output produced by academic staff at ETH Zurich, one of the leading universities of science and technology in mainland Europe. The repository is operated by the ETH Library, which serves both as the main library of ETH Zurich and as a Swiss national centre for technical and scientific information.

The Research Collection is a publication platform that offers three main functionalities: it is a directory of all publications produced at ETH Zurich; it is an open-access repository; it is a research data repository. The platform was developed by the ETH Library from 2014 to 2017. The project included a tender process to select a service provider that led the technical implementation of the repository and provides ongoing maintenance support. It also involved the migration of data from two separate legacy systems into the new repository. While the previous systems were both based on the open-source software Fedora, the Research Collection runs on DSpace, an open-source repository tool widely used in academic libraries worldwide.

In order to integrate the Research Collection into the information landscape at ETH Zurich, and to fulfil all the requirements of a platform with the functional scope described above, the ETH Library implemented comprehensive customisations in DSpace. The repository now features various interfaces with internal and external systems. For example, in the ETH Zurich's academic reporting system, in the Annual Academic Achievements, and on researchers' institutional websites, Research Collection data are used to display publication lists (Hirschmann, 2018).

III. FEATURES FOR PUBLISHING RESEARCH DATA

Although the Research Collection hosts open-access publications and research data, certain features were designed and implemented especially with those users in mind that use the repository as a research data repository.

For example, research data can be published as supplementary material with a publication but also as a stand-alone publication. If the dataset is a supplement to a publication, there is a feature to link these two items together. When users upload their research data, they can choose a value from a list of resource types to categorise their data. The types offered are a subset of the resource types defined in the DataCite Metadata Schema (DataCite Metadata Working Group, 2019) and include dataset, image, model, software, sound, video and data collection.

Users define the access rights for their datasets themselves. The possible access rights for research data range from open access to closed access, the latter meaning that only repository staff can access the files. Options for restricting access to a dataset also include embargoes and granting access to all or selected users from ETH Zurich only. It is worth noting that the metadata of an item are always freely accessible, so even closed-access datasets will have a publicly visible landing page. If a dataset has restricted access settings, end users can still request access to the files from the landing page via a request form. Repository staff forward access requests to the submitter or rights holder of a dataset who then decides whether to grant the requester access to their data. For freely accessible datasets, submitters can choose an open content licence that will then be displayed on the landing page of the dataset from where end users download the files.

For each dataset, a digital object identifier (DOI) is minted. If users need the DOI before actually publishing their dataset – for example to include it in a manuscript – they can reserve a DOI. The Research Collection also displays download statistics for published datasets both at file and item level. For end users, there is a feature to preview the contents of ZIP and TAR containers before actually downloading the files.

In terms of file formats, there is no technical limitation to what users can upload. If a certain file format is known to repository staff and has therefore already been added to the DSpace file format registry, the Research Collection displays the support level for the uploaded files to the submitter. Users can then choose that the library should keep their data for an unlimited period of time or, alternatively, they can indicate a limited retention period of 10 or 15 years, for example if they already know that their file formats will not be usable over the long term. All uploaded files, independent of their retention period, are transferred to the library's preservation system, the ETH Data Archive, which is based on the software Rosetta by Ex Libris (see Töwe & Barillari, 2020).

IV. HOW ETH RESEARCHERS USE THE REPOSITORY

In this chapter, we present some insights into how ETH Zurich's researchers actually use the repository when it comes to publishing their datasets.

Depositing data in the Research Collection is not mandatory for researchers at ETH Zurich. While there is a strict requirement for researchers to report all their publications via the Research Collection so they can be listed in the annual academic reports and there is also an open-access policy (ETH Zurich, 2018) that requires researchers to deposit open-access versions of their papers in the Research Collection, there is no dedicated policy for depositing and publishing research data. The Guidelines for Research Integrity and Good Scientific Practice at the ETH Zurich (ETH Zurich, 2007) do require proper data management and contain a general expectation to share data, but they do not require deposit in the Research Collection.

In 2018, the first full year of operation of the repository, researchers deposited 191 datasets in the Research Collection. The number of published datasets has since grown each year, from 233 items in 2019 to 329 items in 2020. This brings the total number of datasets published in the Research Collection to 865 items at the end of 2020.

As mentioned above, there are various subtypes of research data to select from when uploading data. Around 80% of users categorise their data as either a dataset or data collection, while the more specific types such as image or video are not used as much (Fig. 1).

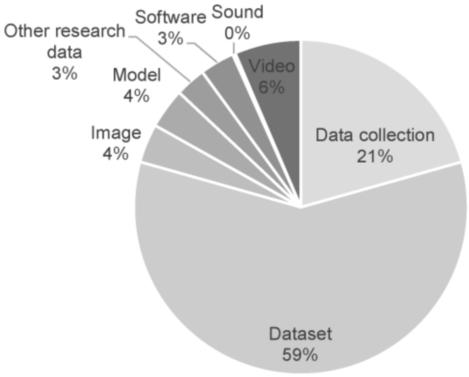

Fig. 1. Types of datasets in the Research Collection

When uploading datasets, users are only required to enter a few mandatory metadata fields in the submission form such as title or creator. However, there is also a range of optional metadata they can provide and this information is particularly important when it comes to making datasets findable in compliance with the FAIR principles (Wilkinson et al., 2016). Looking at the usage of these optional metadata fields, we can see that most users do not provide any information in the methods or software section and only a few datasets are linked to a grant. Less than a third of all research data items contain an abstract or subject keywords. However, around half of all datasets have an entry in the "Related publication" field, so the functionality to link publications and datasets is used quite often and plays an important role in making the data discoverable (Fig. 2).

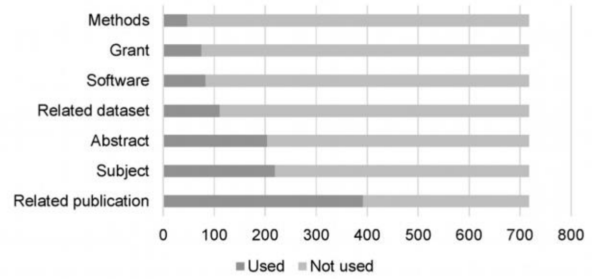

Fig. 2. Share of datasets for which submitters provided certain optional metadata.

Looking at the file formats, there is a large amount of datasets that users provide in ZIP containers. The individual files are often text files or CSV files, but also PDF files or other formats from a long list of other proprietary and non-proprietary file formats (Fig. 3).

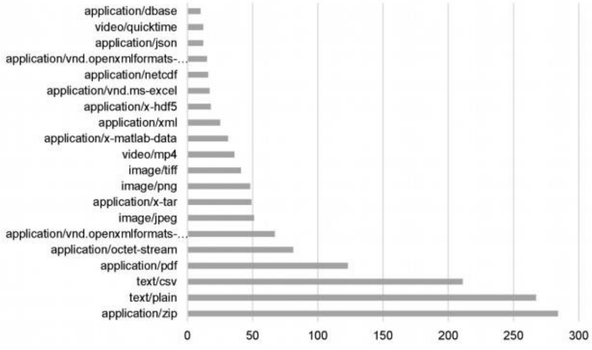

Fig. 3. Number of files of a certain format deposited in dataset items in the Research Collection.

In terms of availability, ETH researchers rarely use the option to restrict access to their datasets that they deposit in the Research Collection. There are datasets that are deposited in an external repository and then only linked from the Research Collection. These are the items described as "Metadata only" in Fig. 4. Apart from these items, almost all datasets are published open access.

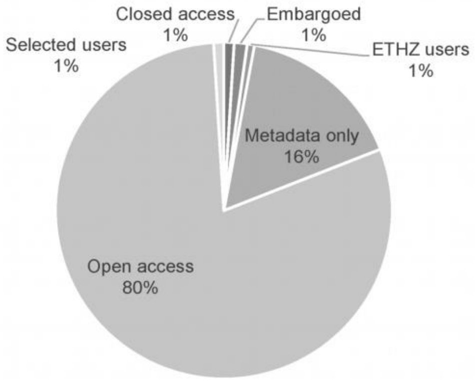

Fig. 4.Access status of datasets in the Research Collection.

When it comes to licensing, around half of the users decided to publish their dataset without an open content licence, instead using the repository's standard copyright statement that allows usage for non-commercial purposes but does not allow redistribution of the content. Among those that chose an open content licence, most users chose the Creative Commons Attribution licence (Fig. 5).

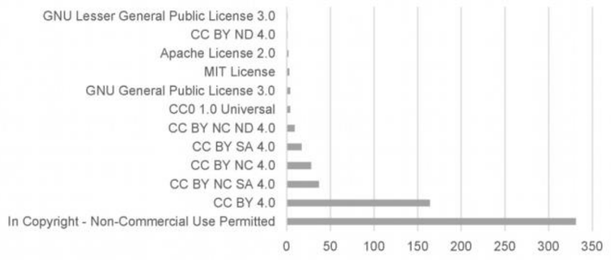

Fig. 5. Rights statements of datasets in the Research Collection.

V. QUALITY assurance AND COMPLIANCE CHECKS

Similar to the experiences of other academic libraries (e.g. Lafferty-Hess et al., 2020), one of the main challenges for the ETH Library when setting up and running a repository for research data has been the definition of the library's responsibility and the scope of its activities when it comes to data curation, quality assurance and compliance monitoring.

Prior to launching the Research Collection, the E-Publishing team at the ETH Library had mainly dealt with quality assurance and copyright compliance of open-access publications and metadata-only records. Working with research data items therefore posed some new challenges for the team, such as how to check the validity of metadata and file formats in research data items and how to deal with research-data-specific risks when it comes to compliance with institutional policies and legal norms. In this chapter, we describe the workflows that are currently in place for dealing with research data items in the repository. However, this a dynamic field and we expect to continuously review and adapt our processes in the coming years as we learn more about user requirements and best practices in data publishing.

A. *Quality assurance*

The repository has a quality assurance (QA) process in place. When there is a new submission, repository staff have a look at the item and perform some basic checks before they release the files to the public or the designated users. These checks involve various aspects regarding the metadata and files provided by the submitter.

As a first step, the repository staff check whether metadata are consistent with repository rules, correct spelling errors and check whether related publications or datasets are correctly linked to the item. The staff also add formal metadata such as MIME type and size of the dataset to the record. Since QA staff are mostly trained librarians and information professionals – not subject experts – they do not add metadata describing the content of the dataset.

The QA process also involves checking the files the submitter has provided. First, the dataset is downloaded in order to detect potentially virus-infected files. Then, QA staff try to check the readability of the files or a sample of files by opening them with a viewer or another tool. They also run the files through DROID – a tool from the UK National Archives – to perform format identification. Once this step is completed, QA staff check whether the detected file formats are compatible with the retention period the submitter has chosen. If there are new formats that are not yet recorded in the repository's file format registry, they determine the support level and add them to the registry. Apart from file-format-related checks, staff also check whether the file names and folder structure are comprehensible.

It is important to note that if QA staff detect any problems with the submitted files during these steps in the QA process, they do not edit or manipulate the submitted files. Instead, they contact the researchers that have submitted the datasets and inform them about the potential problems they have detected. Researchers then have the opportunity to make suggested changes to their files and re-upload them.

B. *Compliance checks*

Another aspect often closely linked to the step of looking into the uploaded files is compliance with policies and legal norms. In principle, compliance with legal norms is the responsibility of the submitter. This is stated explicitly in the Research Collection terms of use. Users also need to confirm in the submission process that they are not violating third-party rights or institutional policies by submitting their dataset.

On the other hand, looking at the uploaded datasets, QA staff regularly detect cases that violate certain policies or legal norms. This is usually not because researchers knowingly decide to violate third-party rights, but rather because they are either not aware that certain norms exist or they simply forgot to delete certain files in their data collection before submitting them.

Observing this discrepancy between what researchers confirm in the submission form and what repository staff see in the submitted datasets, the ETH Library decided that repository staff should inform users if they detect violations of certain norms during the QA process and that they would not release the data to the public before these possible violations were cleared up or resolved. This process is both a service to the researchers but also a risk management measure for ETH Zurich as the institution that hosts the datasets and runs the repository.

Copyright compliance checks involve checking whether a dataset contains third-party copyrighted material and checking files and metadata for licence incompatibilities. When it comes to copyright, what happens on a regular basis is that users include third-party copyrighted material in their data collections without having obtained the copyright owner's permission to publish the material, or that the licence that users choose in the submission form contradicts the licence statement that they have included in their data collection. In order to resolve such potential copyright violations and contradictions, QA staff contact submitters and ask them to delete certain files, obtain permission from copyright holders and/or rethink their licensing choice.

Research data deposited in the Research Collection often contain scripts and software code. As defined in the university's Exploitation Guidelines (ETH Zurich, 2020), all software developed at ETH Zurich and made available to third parties – even under an open-source licence – must be registered with the technology transfer office ETH transfer. Since this policy is not yet well known among ETH researchers, repository staff regularly find software code within the submitted data packages that has not yet been registered with ETH transfer. In these cases, QA staff inform the researchers that they are required to register their code and publish it under an open-source licence.

Disclosure risk is another topic for which repositories must put in place policies and potential mitigation measures. In the Research Collection submission process, researchers have to confirm that they have anonymised all personal data and obtained written consent of the study participants for publication. However, there is still a remaining risk that potentially sensitive, personal data could be released to the public. To mitigate this risk, some specialised, disciplinary repositories have dedicated disclosure risk review workflows in place (see e.g. ICPSR, 2021). However, such a process is not feasible for most institutional repositories such as the Research Collection. The ETH Library has neither the expertise nor trained staff members to perform disclosure risk reviews on datasets. We therefore approach this topic by taking preventative measures. This includes offering training sessions with data protection experts from other ETH units and arranging individual consultation sessions so that researchers working with patient data or other sensitive data can discuss their specific use cases and datasets with a data protection expert before they submit the data to the Research Collection.

C. *How to reconcile publishing and preservation requirements*

One particular topic that has come up in discussions at the ETH Library about the QA process is the question of how to deal with conflicting requirements regarding file formats coming from users on the one hand and from the library's digital preservation experts on the other hand. The Research Collection itself is not a preservation system but a publication platform. There is a data export process that continuously transfers all data from the Research Collection (DSpace) to the ETH Library's preservation system, the ETH Data Archive (Rosetta). The Data Archive is a dark archive that hosts a copy of all Research Collection data and that can – if needed in the future – perform preservation tasks on the deposited files to keep them readable.

One requirement for being able to perform such preservation tasks is that the Research Collection deliver to the Data Archive only files in formats that are suitable for long-term archiving. This, however, is a non-trivial task to achieve in a research data repository because often research data are produced and stored in file formats not suitable for preservation. As described above, Research Collection staff generally do not edit the researchers' files and therefore also do not convert files to other formats. Implementing a file format conversion service would also require considerable additional human resources at the ETH Library. On the other hand, it is also usually not a top priority for researchers to invest time in file format conversions either. When researchers upload their data to the Research Collection, their priority is usually to have it published sooner rather than later.

At the same time, QA staff have noticed that even if researchers submit their data in non-recommended formats, they still often ask the library to keep these data for an unlimited period of time. Actually, three quarters of all datasets are deposited with the user choosing an indefinite retention period, rather than a limited retention period of 10 or 15 years, with a large part of these deposits coming in non-recommended formats.

Looking at this situation, the library has recently decided that it will change its approach to this topic. Originally, we had assumed that every dataset with an indefinite retention period must only contain files in formats suitable for archiving. However, since we have realised that it is not possible to achieve this in practice, we have decided that we will no longer use the retention period as the main indicator for long-term preservation. Instead, we are now working on implementing a separate checkbox in the submission form where we ask users to indicate if they are actually interested in keeping their data readable over the long term. Only if the submitter activates this checkbox will our team provide recommendations and work with the submitter to help them convert their files into suitable formats. In all other cases – independent of the chosen retention period – we will assume that only bitstream preservation is required.

VI. NEW DEVELOPMENTS

The final chapter of this paper discusses two ongoing development projects at the ETH Library that will significantly extend the functionalities of the Research Collection in the coming months.

A. *Integration with openBIS*

The first of these projects is a collaboration project between the ETH Library and ETH Scientific IT Services. It addresses the need of bringing two currently separate tools together: openBIS as a tool for active research data management (Barillari et al., 2016) and the Research Collection as a tool for data publication. By connecting these two systems, we want to provide an integrated solution for ETH researchers that supports their workflows from active data management to publication and preservation.

Form a user perspective, this integration will provide researchers with a seamless workflow for publishing selected data from their openBIS instance via the Research Collection. The user starts in openBIS and selects the files they want to export to the Research Collection. openBIS then creates a ZIP container that includes some basic metadata about the exported data collection and another ZIP that contains the actual bitstreams. openBIS transfers the main ZIP container to the Research Collection via DSpace's SWORD API (Allinson et al., 2008). This API is a standard feature that comes with all DSpace installations. DSpace then creates a workflow item and presents it to the user. The user can add additional metadata, review the imported bitstreams and select access settings and an end-user licence. At this point, the Research Collection sends the permanent Handle URL of the item back to openBIS, in order to display the link in the user's publications collection. In the Research Collection, as with any other item, QA staff will review the submission and, if accepted, publish it in line with the access settings the user has chosen.

B. *Solution for publishing large datasets*

The second ongoing development project aims to provide a solution for publishing large datasets. This project was mainly driven by feedback from users who indicated that they need to publish files in the Research Collection that are much larger than what the repository can currently accommodate. At the moment, the maximum recommended file size is 10 GB per individual file and 50 GB as the maximum size for the total amount of files within one item.

When looking into possible technical solutions for this new requirement, we focused on leveraging existing tools within ETH Zurich, rather than setting up a completely new technical infrastructure. The chosen solution will integrate the Research Collection with a separate storage solution for large files based on ownCloud, which has already been in use at ETH Zurich under the brand name polybox.

The overall concept of this solution is that for large datasets, the Research Collection will provide a metadata record that is also used as a landing page for DOI resolution, and this metadata record then links to a download page on ownCloud. ownCloud will be used for data storage and for uploading and downloading the actual files. From a technical point of view, ETH Zurich’s IT Services have made this possible by extending the ownCloud infrastructure used for the polybox service with an additional server called libdrive (Fig. 6). While polybox is a service provided to individual users at ETH Zurich as a drop-box-like storage solution, libdrive is managed by the ETH Library and will be used for implementing the Research Collection large files workflow but also for other use cases within the library requiring transfer or storage of large files.

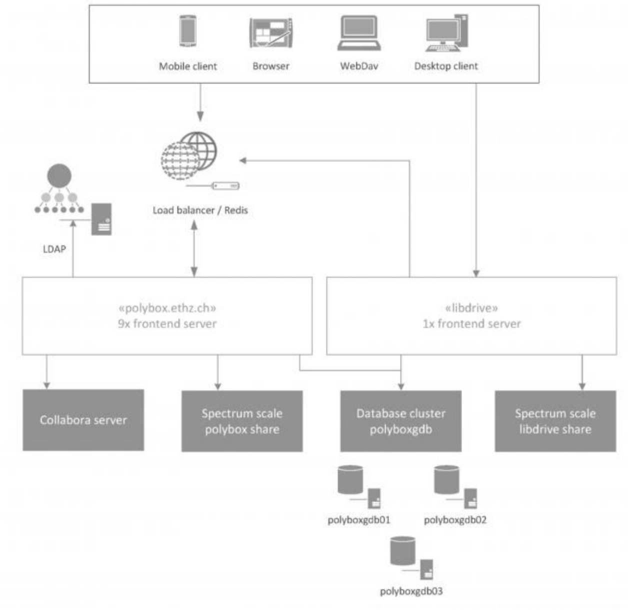

Fig. 6. ownCloud infrastructure including libdrive for publishing large datasets at ETH Zurich.

For all uploads – whether small or large datasets – the users will first go the Research Collection. In the upload form, they can either directly upload their small files to DSpace or request an access link for the upload of large files. Users with large files will then be asked to transfer their files via an ownCloud web client for files between 10 and 20 GB, via a local ownCloud client or WebDav for files approximately between 20 and 200 GB and via an offline USB device if the files are larger than 200 GB. We also expect that there might be some datasets with files that are too large to be downloaded via a browser or client. For these files, we plan to implement an email request form that enables end users to order these files to be sent to them on a USB device (Table 1).

TABLE I. UPLOAD AND DOWNLOAD WORKFLOWS FOR LARGE FILES

ile size	Upload via	Download via
<10 GB	Research Collection submission form	Browser
10–20 GB	ownCloud web client	Browser
20–200 GB (approx.)	ownCloud client or WebDAV	Browser of client
200 GB–1 TB (approx.)	Offline transfer via USB device	Offline transfer via USB device (request access via email form)

VII. WHAT’S NEXT? PLANS FOR THE FUTURE

For 2021, apart from finishing the two projects described in the previous chapter, one of the main goals for the Research Collection is to complete the application process for certification with the CoreTrustSeal (Dillo & de Leeuw, 2018). We believe that the certification process can help us detect gaps and weak spots in our policies and workflows and that the certification will increase the trust of our user community in the repository.

On the technical side, we are planning to improve the metadata fields used for geo-referencing datasets, so that users can more easily execute geolocation-based searches, and we will work on the integration of the Research Collection in Google Dataset Search.

REFERENCES

Allinson, J., François, S., & Lewis, S. (2008). SWORD: Simple Web-service Offering Repository Deposit. *Ariadne*, 54, http://www.ariadne.ac.uk/issue54/allinson-et-al/

Barillari, C., Ottoz, D.S.M, Fuentes-Serna, J.M., Chandrasekhar, R., Rinn, B., & Rudolf, M. (2015). openBIS ELN-LIMS: an open-source database for academic laboratories. *Bioinformatics*, 32(4), 638–640. https://doi.org/10.1093/bioinformatics/btv606

DataCite Metadata Working Group (2019). *DataCite Metadata Schema for the Publication and Citation of Research Data. Version 4.3.* DataCite e.V. https://doi.org/10.14454/f2wp-s162

Dillo, I. & de Leeuw, L. (2018). CoreTrustSeal. *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare*, 71(1), 162-170. https://doi.org/10.31263/voebm.v71i1.1981

ETH Zurich (2007). *Guidelines for Research Integrity and Good Scientific Practice at the ETH Zurich.* https://rechtssammlung.sp.ethz.ch/Dokumente/414en.pdf

ETH Zurich (2018). *ETH Zurich's open-access policy dated 17 January 2018.* https://rechtssammlung.sp.ethz.ch/Dokumente/134en.pdf

ETH Zürich (2020). *Richtlinien für die wirtschaftliche Verwertung von Forschungsergebnissen an der ETH Zürich.* https://rechtssammlung.sp.ethz.ch/Dokumente/440.4.pdf

Hirschmann, B. (2018). Die Research Collection der ETH Zürich. *ABI Technik*, 38(3), 223–233. https://doi.org/10.1515/abitech-2018-3003

ICPSR (2021). *Data Confidentiality.* https://www.icpsr.umich.edu/web/pages/datamanagement/confidentiality/index.html

Lafferty-Hess, S., Rudder, J., Downey, M. Ives, S., Darragh, J., & Kati, R. (2020). Conceptualizing Data Curation Activities within Two Academic Libraries. *Journal of Librarianship and Scholarly Communication*, 8(1), eP2347. https://doi.org/10.7710/2162-3309.2347

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

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. https://doi.org/10.1038/sdata.2016.18