ETH zürich

ETH Library

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

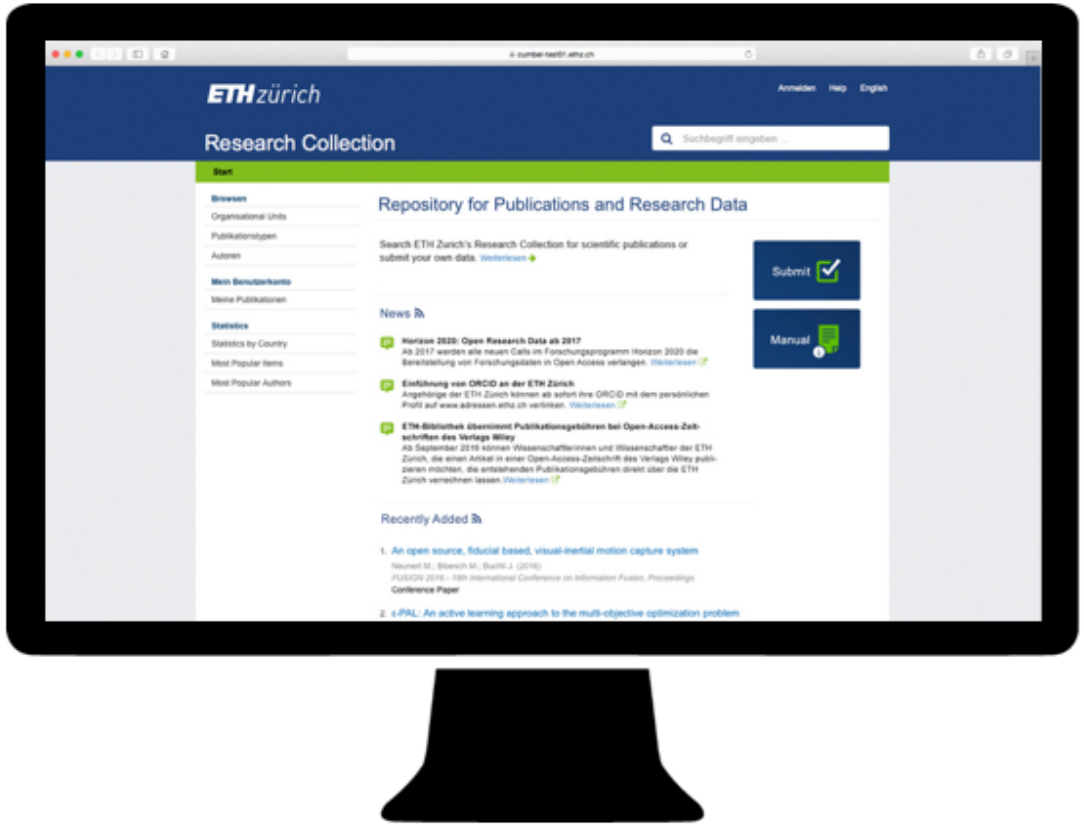
Barbara Hirschmann
Head E-Publishing, ETH Library
22 October 2020, Swiss Research Data Day

Agenda

1. Research Collection: Overview
2. How ETH researchers use the repository
3. Quality assurance and compliance checks
4. New developments

1. Research Collection: Overview

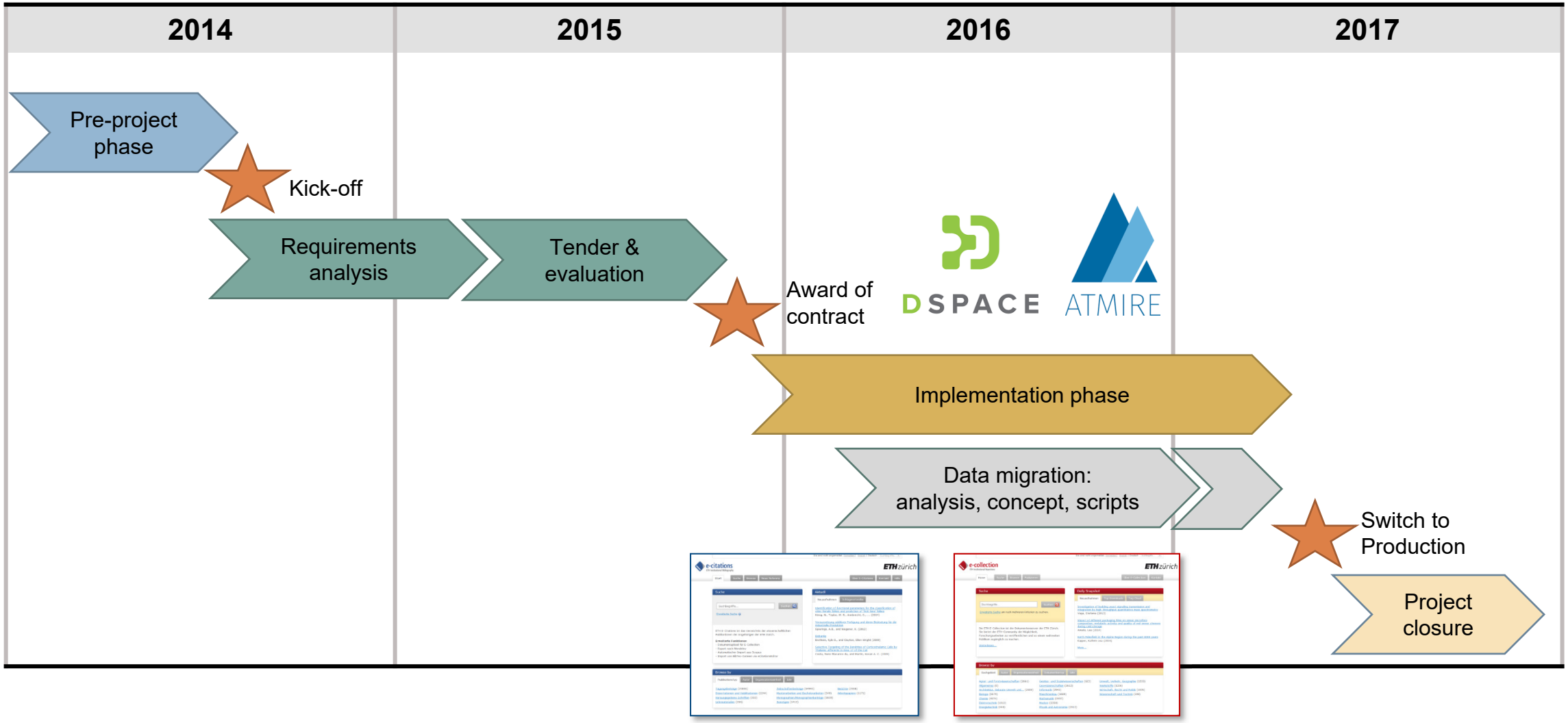
Research Collection: «3 in 1»



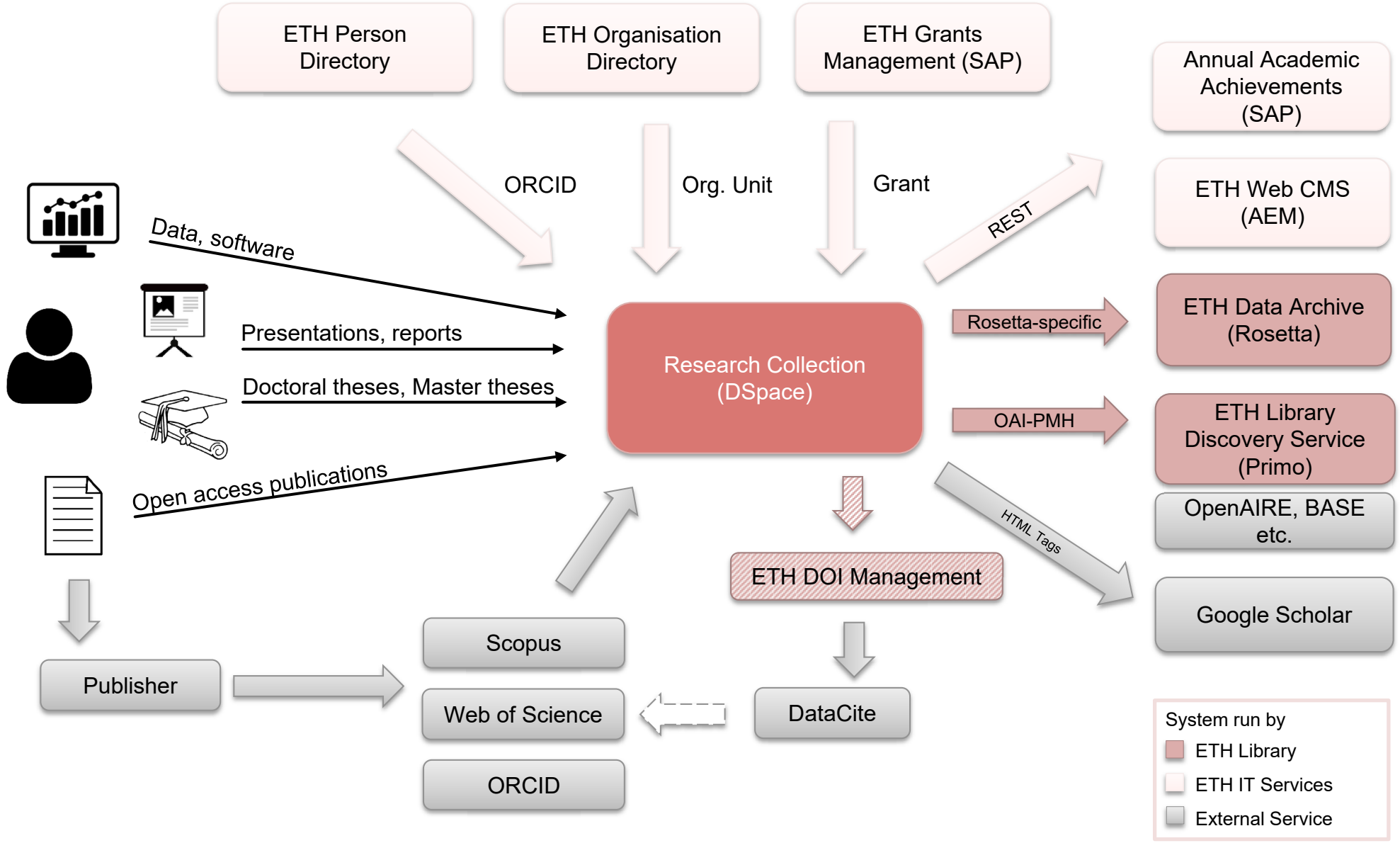
- Publications directory / Institutional bibliography
- Open access repository
- Research data repository

www.research-collection.ethz.ch

Project timeline

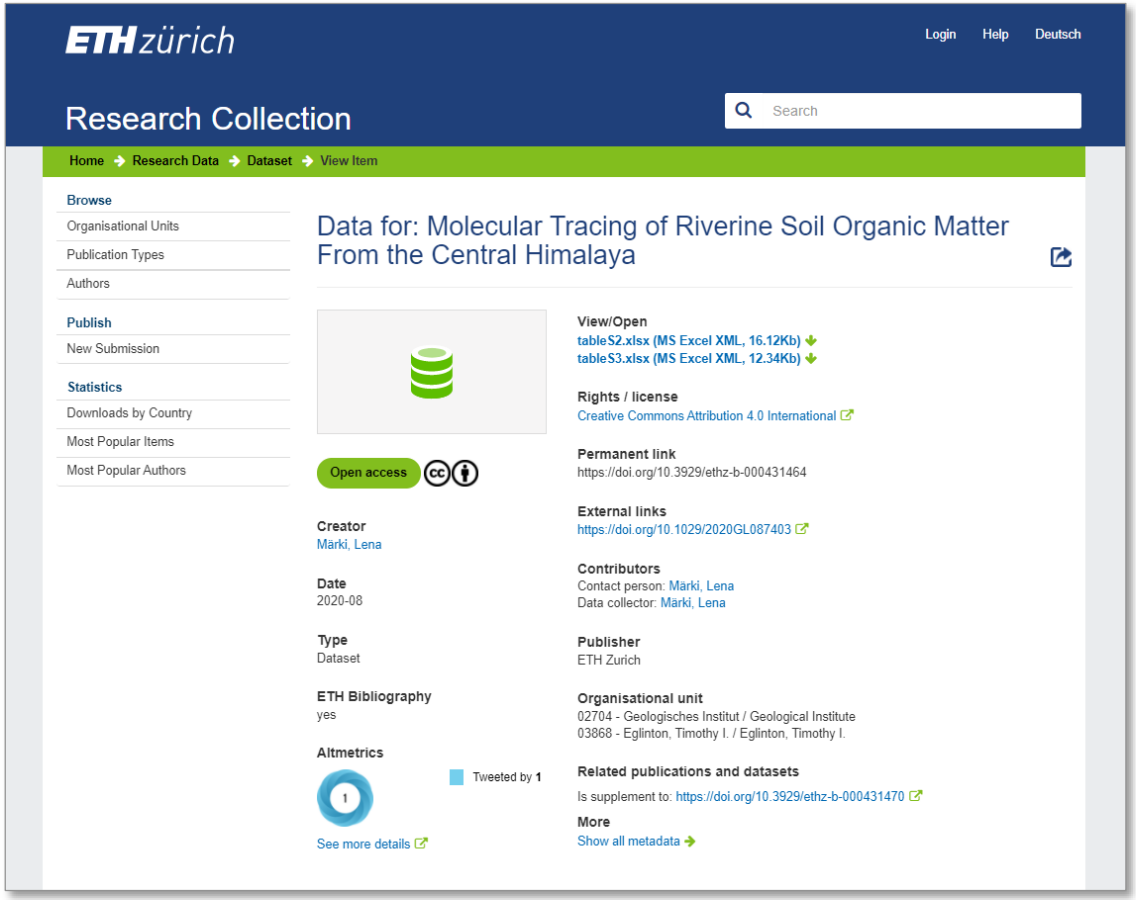


System landscape

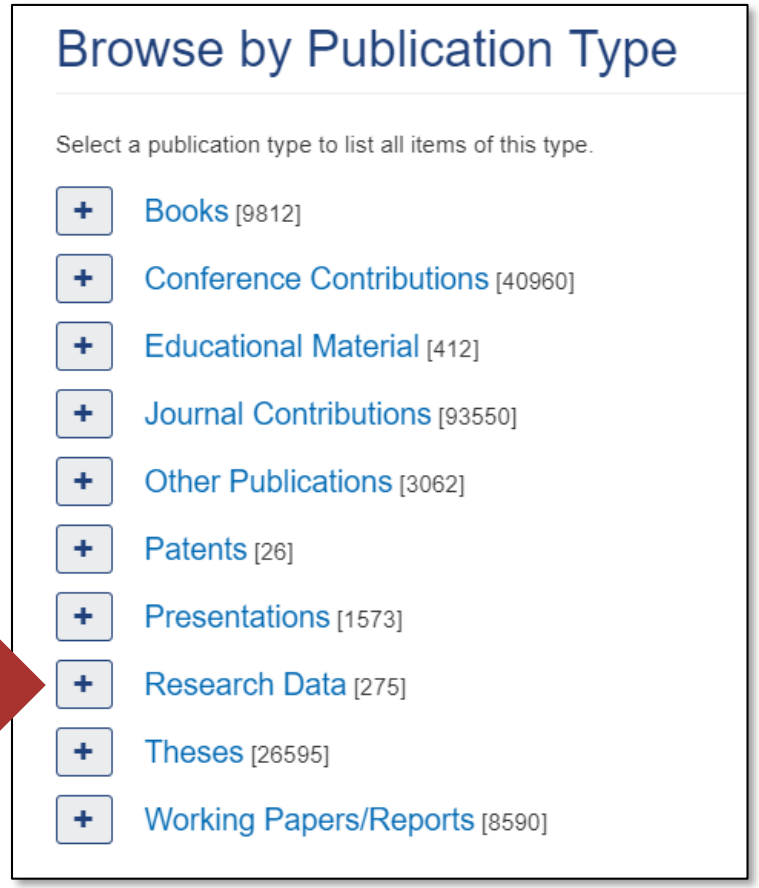


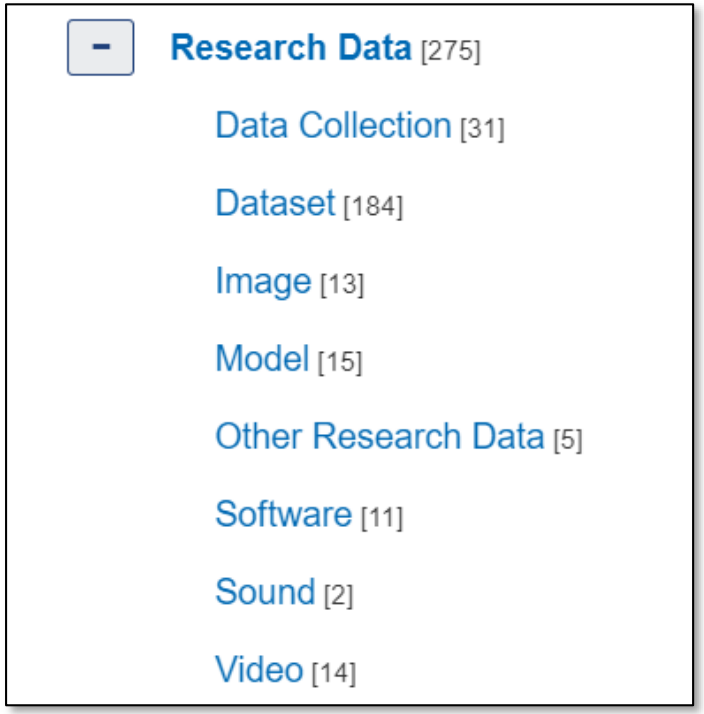
Features for publishing research data

- Deposit research data either as **supplementary material** or as **standalone** publication
- **Linking** between publications and research data
- Flexible **access rights**
- Option for **licensing** data under an open content license
- **DOIs** for research data; option to preview/reserve DOIs
- Usage **statistics** on file and item level
- ZIP- / TAR **preview**
- All **file formats** permitted
- Limited **retention period** possible
- Integration with **preservation system** ETH Data Archive (Rosetta)



Research data as independent publication type





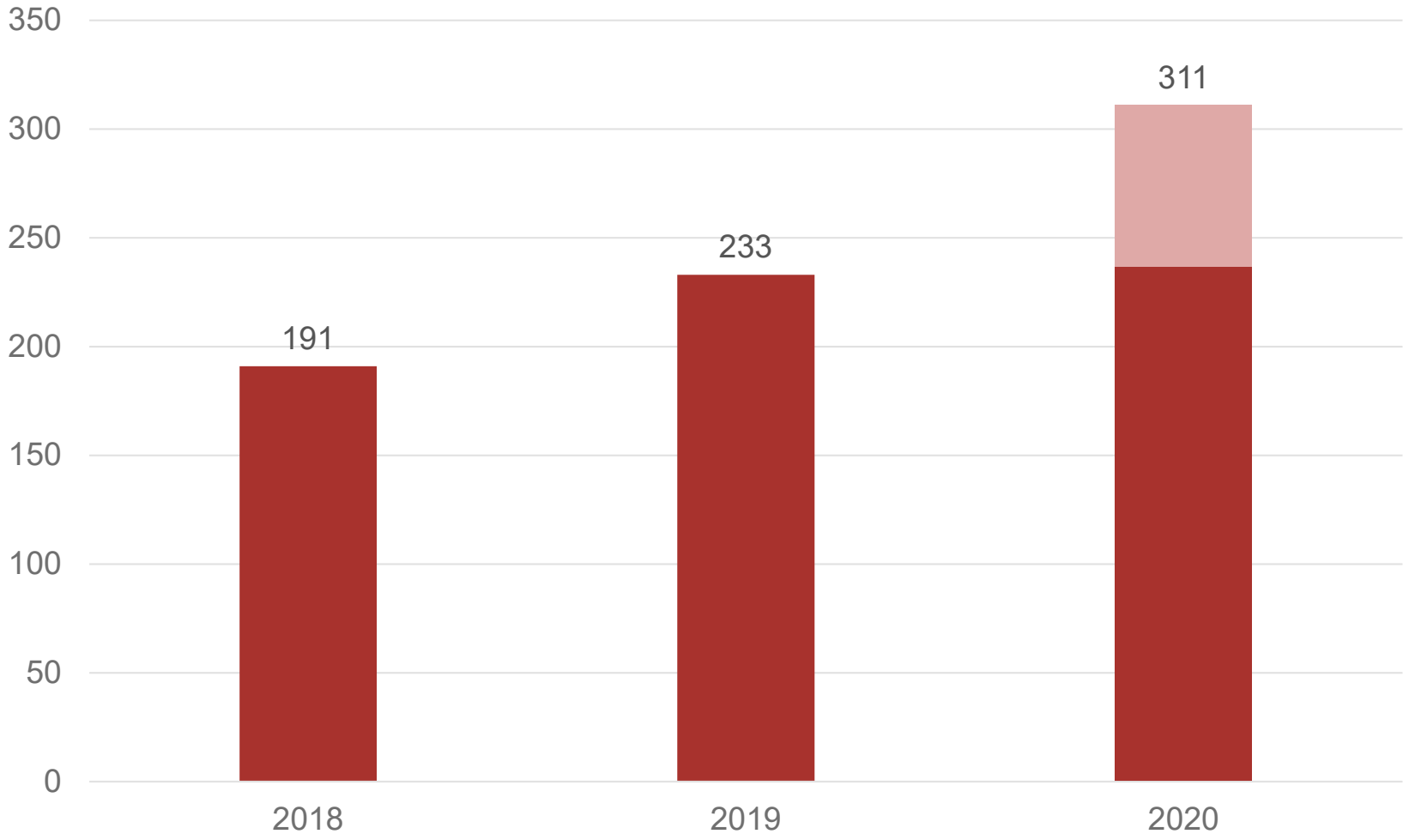
Access rights

	Open access	Embar-goed	ETHZ users	Selected users	Closed access
Publications	✓	✓			
Research data	✓	✓	✓	✓	✓

- Metadata are always freely accessible (via UI search and interfaces)
- Access to restricted datasets can be requested via a form

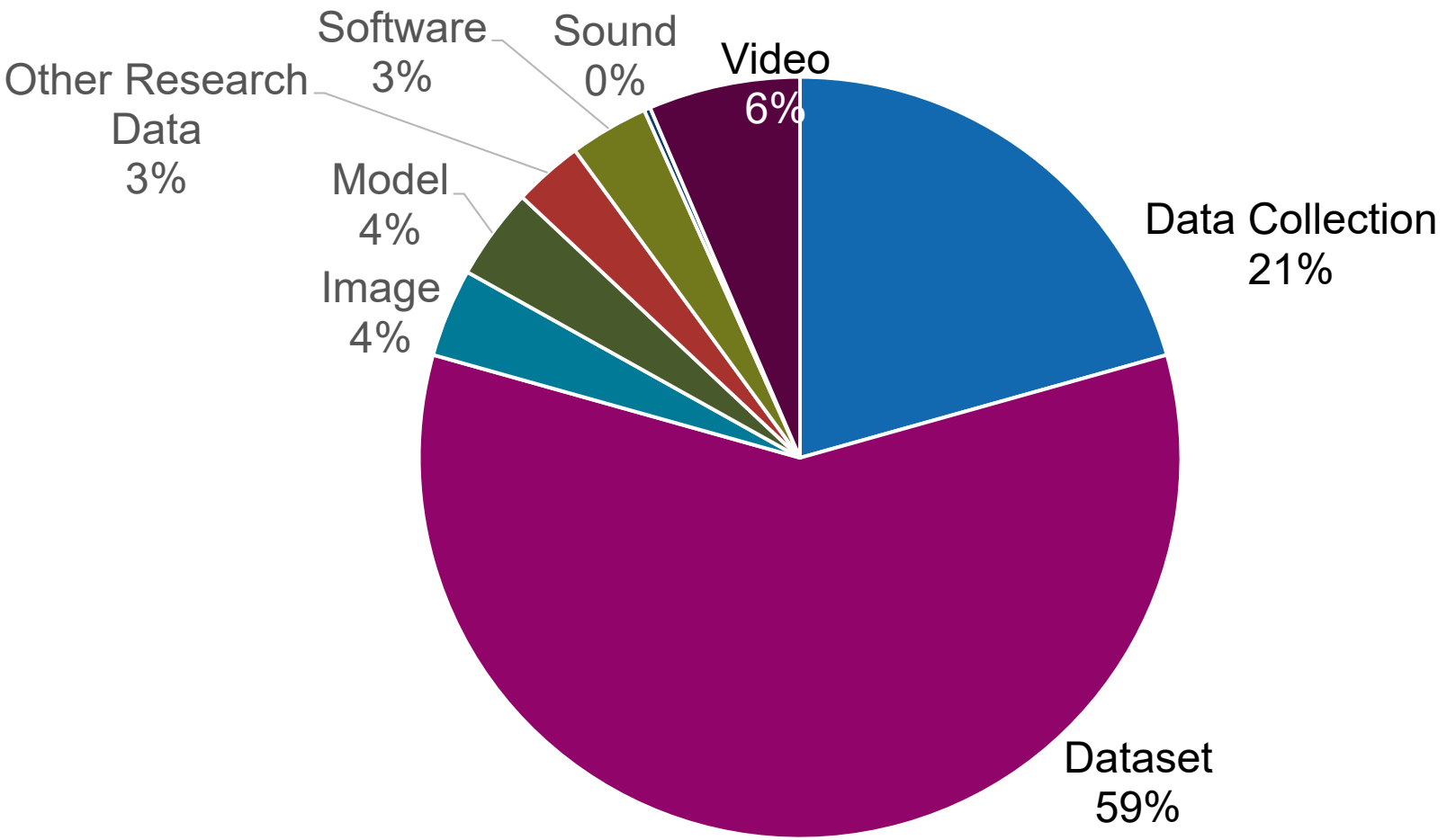
2. How ETH researchers use the repository

Number of published datasets per year

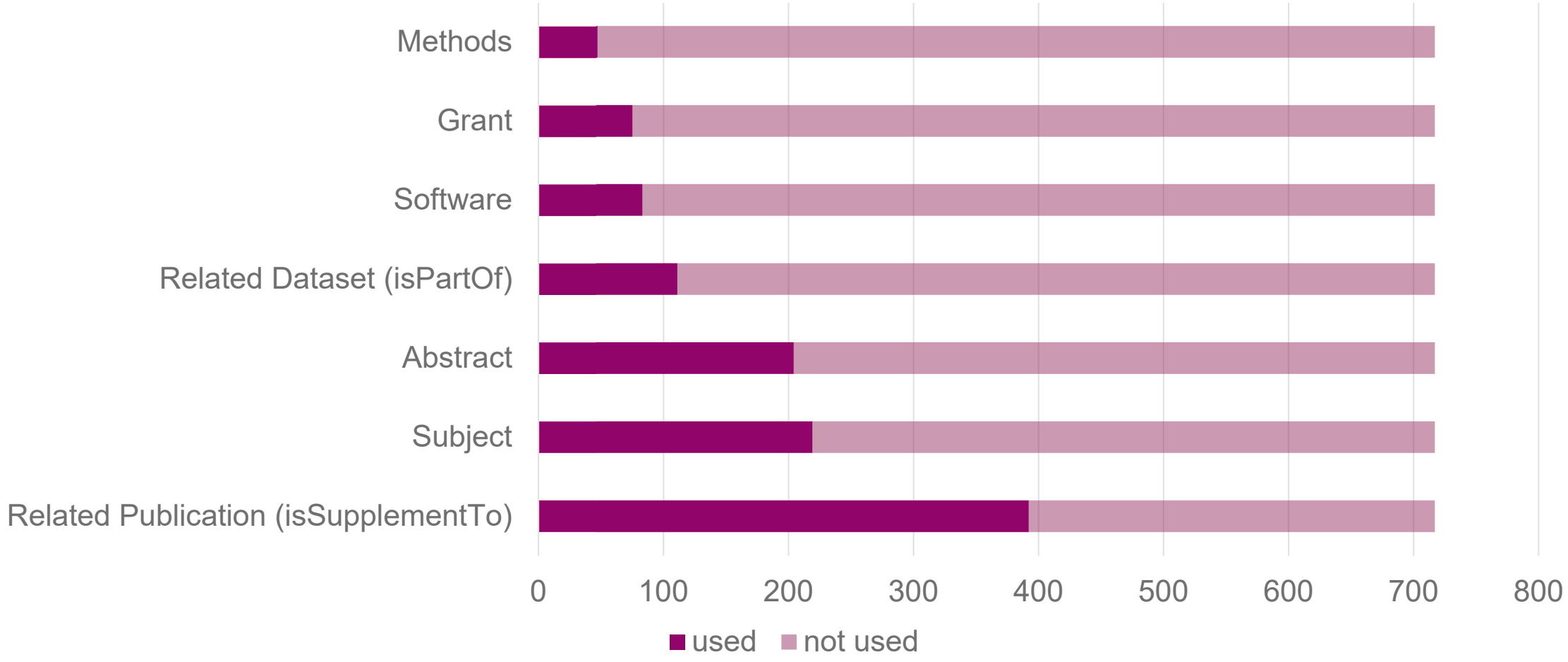


Total:
778 Datasets

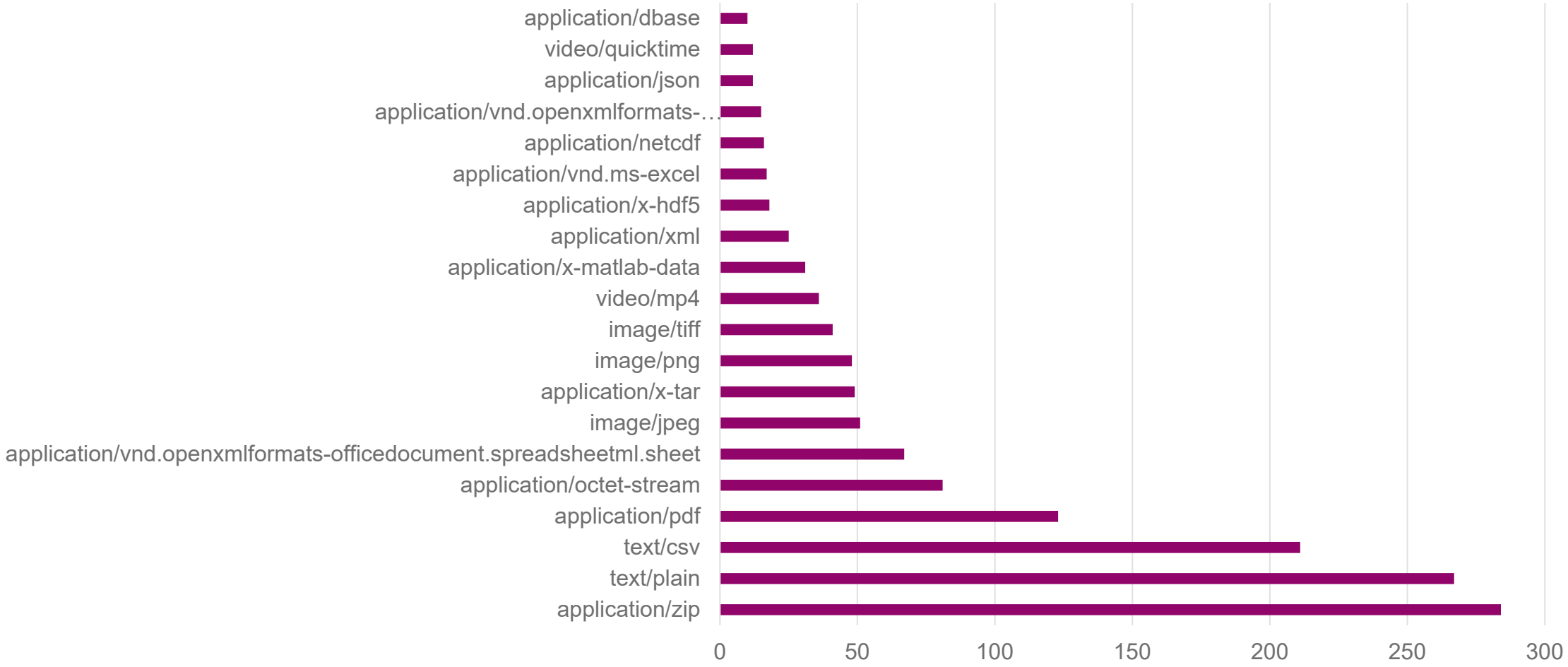
Dataset types



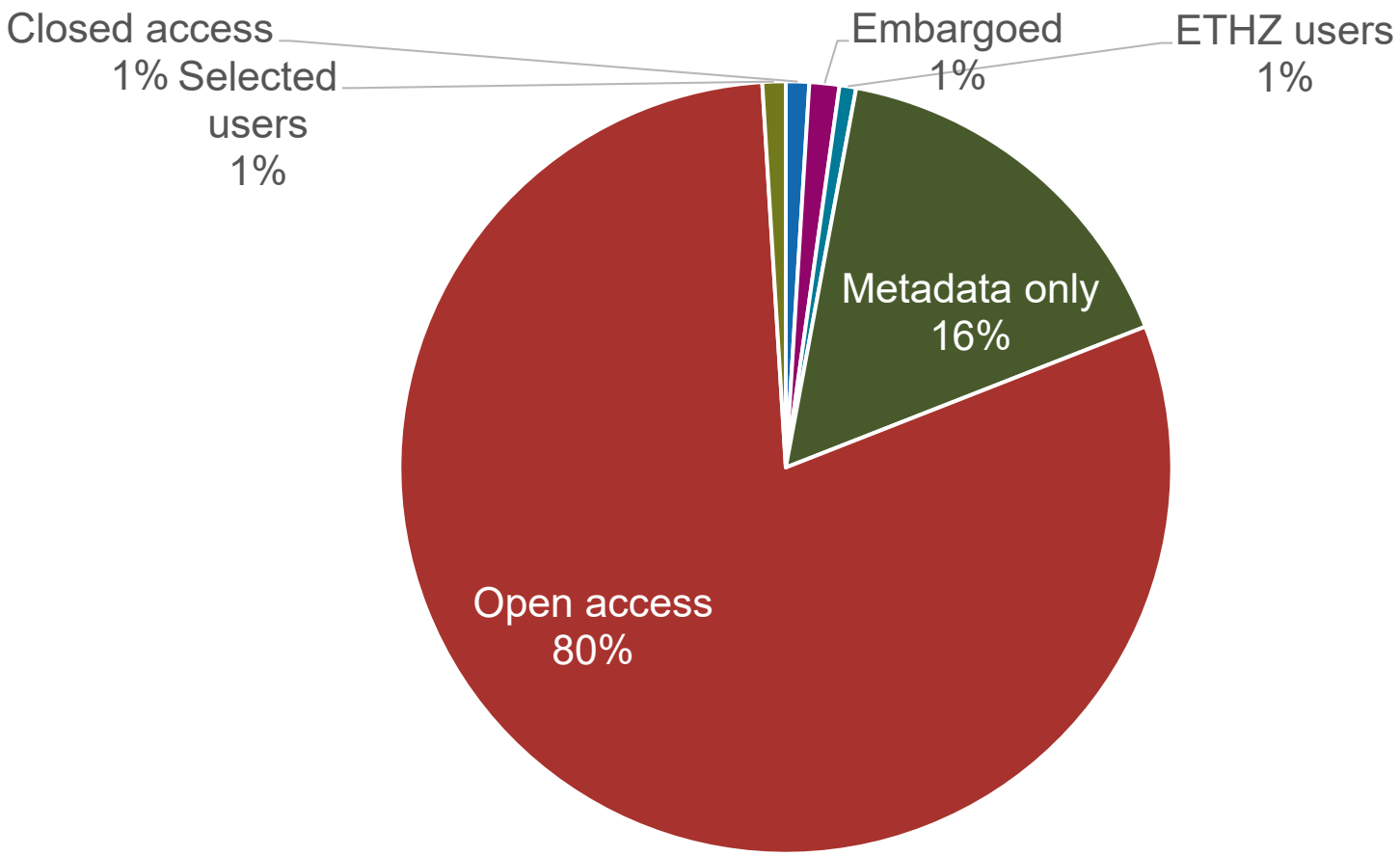
Optional metadata



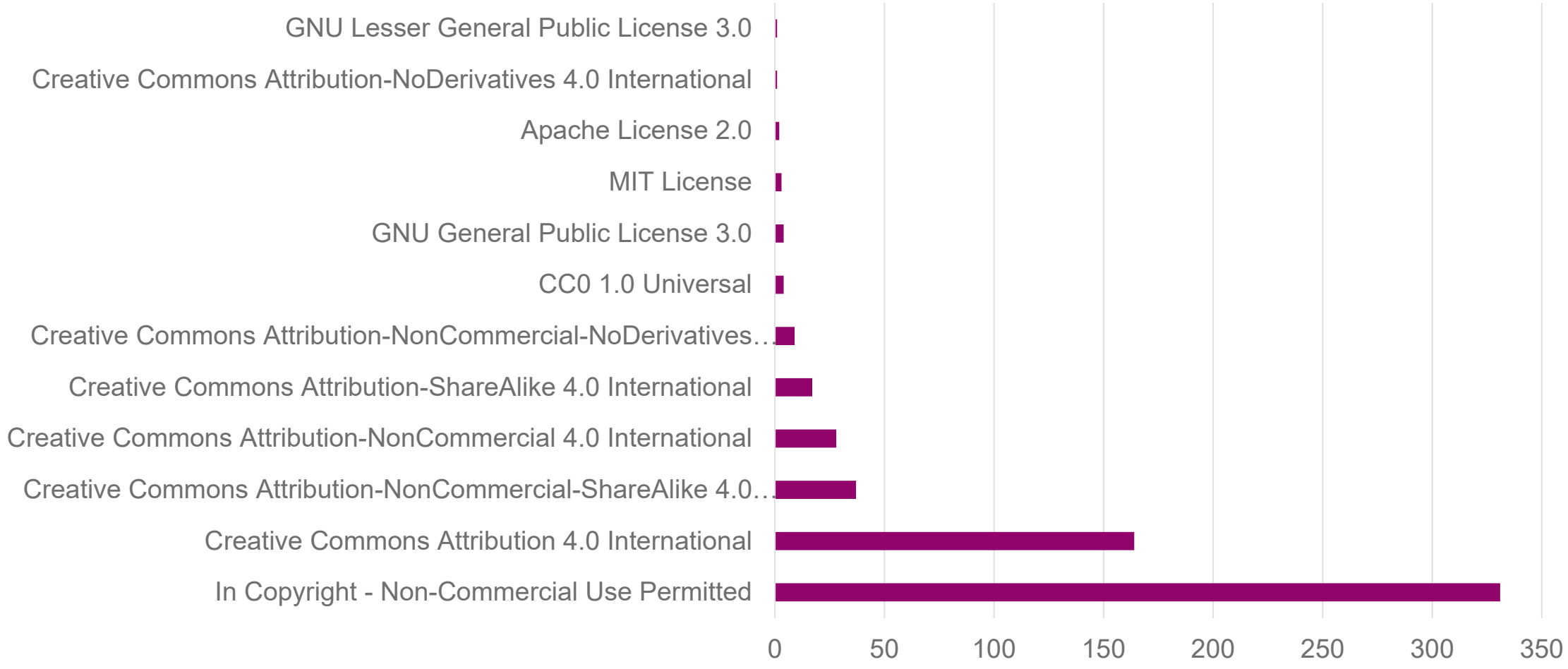
File types (types used in > 10 items)



Availabilty



Licences


GNU Lesser General Public License 3.0
Creative Commons Attribution-NoDerivatives 4.0 International
Apache License 2.0
MIT License
GNU General Public License 3.0
CC0 1.0 Universal
Creative Commons Attribution-NonCommercial-NoDerivatives…
Creative Commons Attribution-ShareAlike 4.0 International
Creative Commons Attribution-NonCommercial 4.0 International
Creative Commons Attribution-NonCommercial-ShareAlike 4.0…
Creative Commons Attribution 4.0 International
In Copyright - Non-Commercial Use Permitted
0
50
100
150
200
250
300
350

3. Quality assurance and compliance checks

What does ETH Library do in terms of quality assurance?

Metadata

- Check submitter metadata for consistency with repository rules and spelling errors
- Check whether related data and publications are linked correctly
- Add formal metadata (access rights, format, size, publisher,…)

What does ETH Library do in terms of quality assurance?

Files

- Virus check
- Check readability (open files with common viewer / tool, random sample for large collections)
- Detect file formats with DROID
- Check whether file formats are compatible with the chosen retention period
- Add new formats and support level to file formats registry
- Check whether file names, folder names and structure are comprehensible

→ Quality assurance does not involve repository staff editing / manipulating files

→ Repository staff contacts submitters with recommendations, submitters decide whether they apply the recommended changes and resubmit their files

Compliance checks

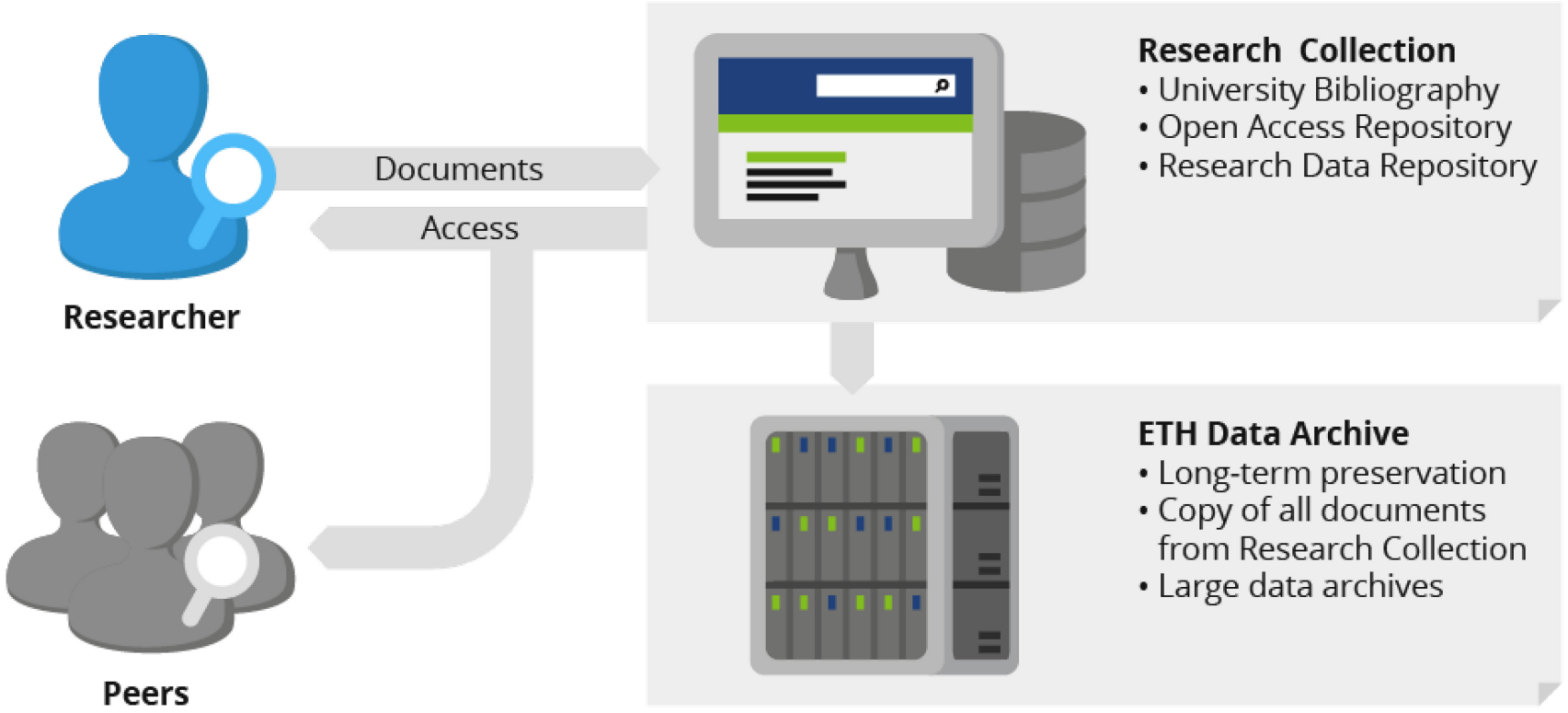
- In principle, according to Research Collection's Terms of Use, end users are responsible for compliance with the following laws and policies and confirm non-violation of these norms when submitting their data:
  - Swiss Copyright Act
  - Guidelines for the Financial Exploitation of Research Results at ETH Zurich
  - ETH Zurich Compliance Guide

- However, in terms of risk management and as a service to ETH researchers, ETH Library does inform users if repository staff detects violations of these norms during their quality assurance tasks.

Compliance management

Topic	Risk	Mitigation
Copyright	• Third-party-copyrighted material included in data collection • Conflicting licence statements for overall data collection vs. individual files	• Inform submitter of potential copyright violations or licence incompatibilities
Exploitation of research results	• Software not registered and licenced according to ETH guidelines	• Inform submitter about ETH Software Licencing Policy
Protection of personal data	• Non-anonymised personal data included in dataset without explicit permission by human subjects	• Advice users to consult with data protection experts on a case-by-case basis **before upload**

How to reconcile publishing and preservation requirements



How to reconcile publishing and preservation requirements

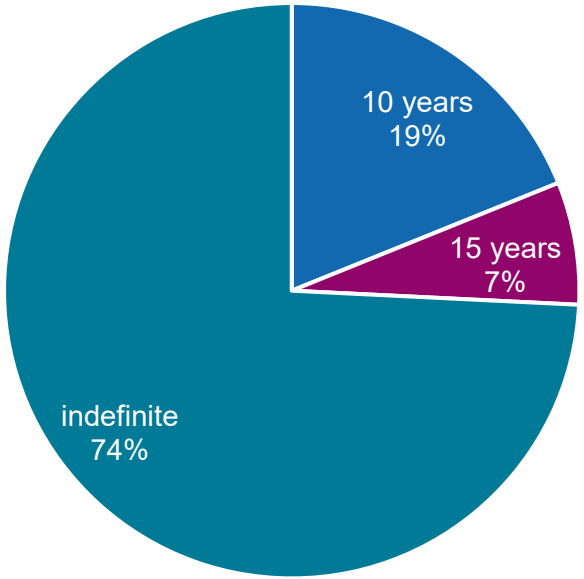
Requirements for preservation

- Submit data in formats suitable for long term preservation
- Either researchers or repository staff must invest resources in file format conversion

User priorities for data publishing

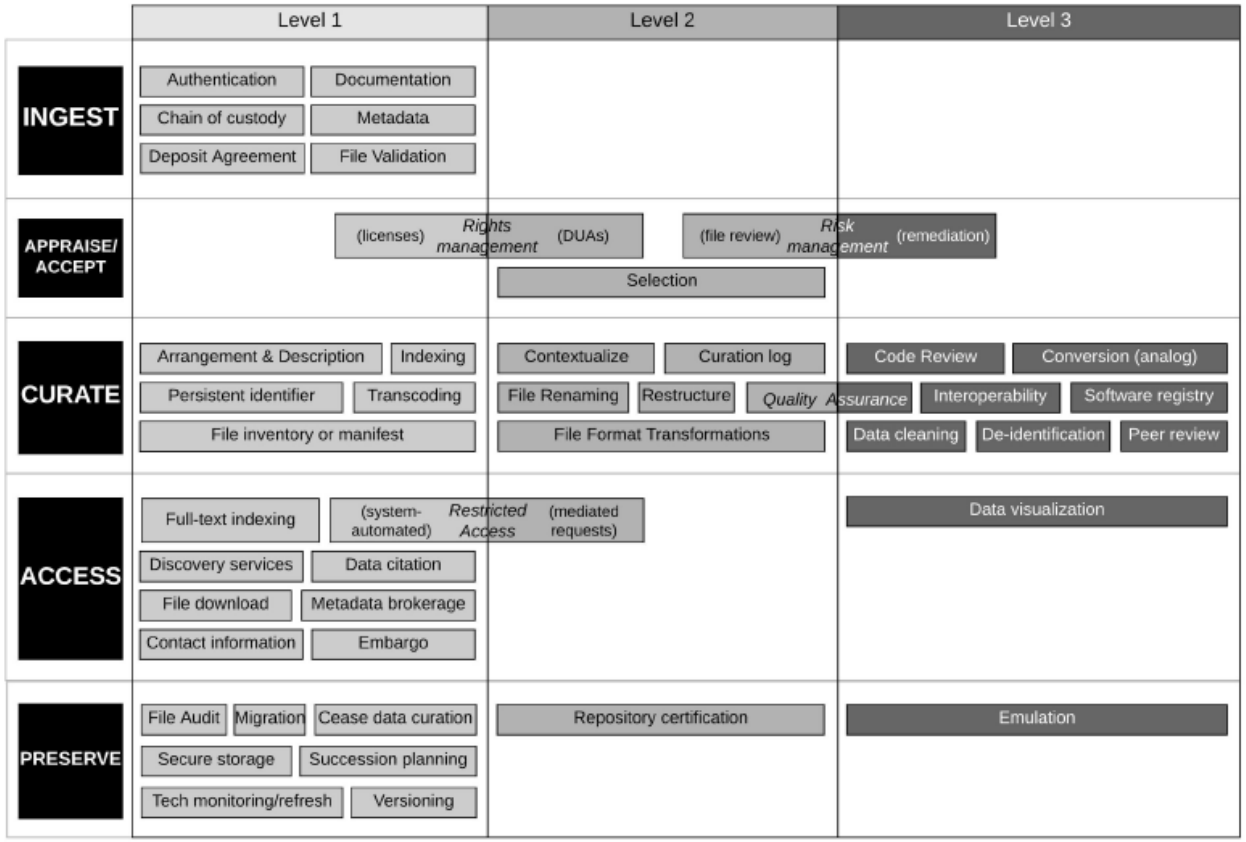
- Publish data for immediate reuse by peers and/or peer reviewers
- Comply with data management regulations of funders and institution
- Invest minimal time in data preparation
- «Better safe than sorry»: users choose indefinite retention period even if file formats are not suitable for long term preservation

How to reconcile publishing and preservation requirements



- **New approach** as of 2020:
  - ✓ Chosen retention period is no longer used as indicator for how long we must «keep the files readable» but only for how long we must «keep the files» (e.g. bitstream preservation)
  - ✓ Submitters need to activate a checkbox if they are actually interested in keeping their data readable over the long term
  - ✓ If activated our team provides recommendations on how to convert the submitted files so they become suitable for long term preservation

Defininig the scope of your quality assurance / data curation activities



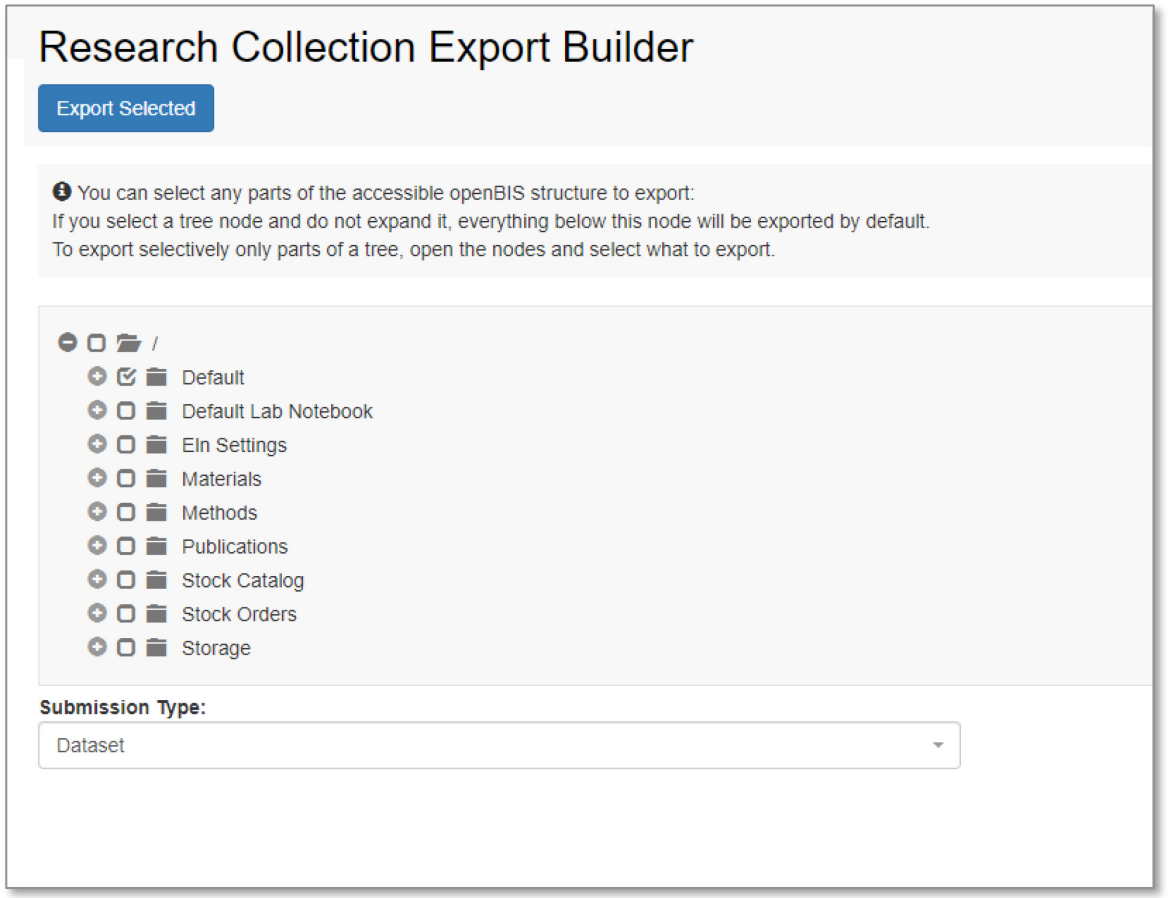
Lafferty-Hess, Sophia, Julie Rudder, Moira Downey et al. (2020). Conceptualizing Data Curation Activities Within Two Academic Libraries. In: *Journal of Librarianship and Scholarly Communication* 8(1). eP2347. https://doi.org/10.7710/2162-3309.2347

4. New developments

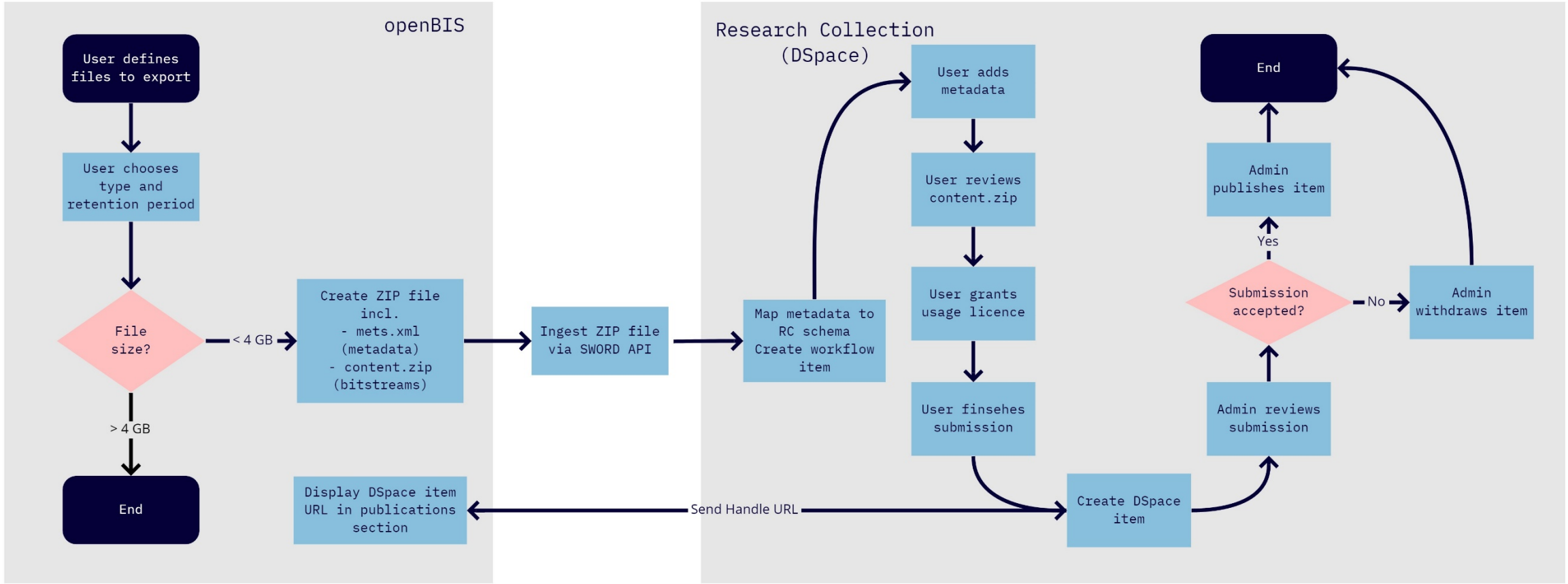
Integration with openBIS

Motivation

- Researchers at ETH can use **openBIS** to manage their daily research and data.
- Researchers use the **Research Collection** to publish their data related to a publication.
- Facilitating the transfer of selected data + metadata from openBIS to the Research Collection benefits researchers who use openBIS for their active data management and provides for an **integrated solution** supporting data management, data publishing, and data preservation.



Integration with openBIS



Solution for publishing large datasets

Motivation

- Users want to deposit much larger files than what we can currently accomodate (ca. 10 GB per file)

Solution

- Integrating two existing services at ETH Zurich: Research Collection (DSpace) and polybox (ownCloud)

Solution for publishing large datasets

Research Collection:

- Metadata record / landing page incl. DOI
- Link to download page on «libdrive»

libdrive:

- File upload
- File download

Solution for publishing large datasets

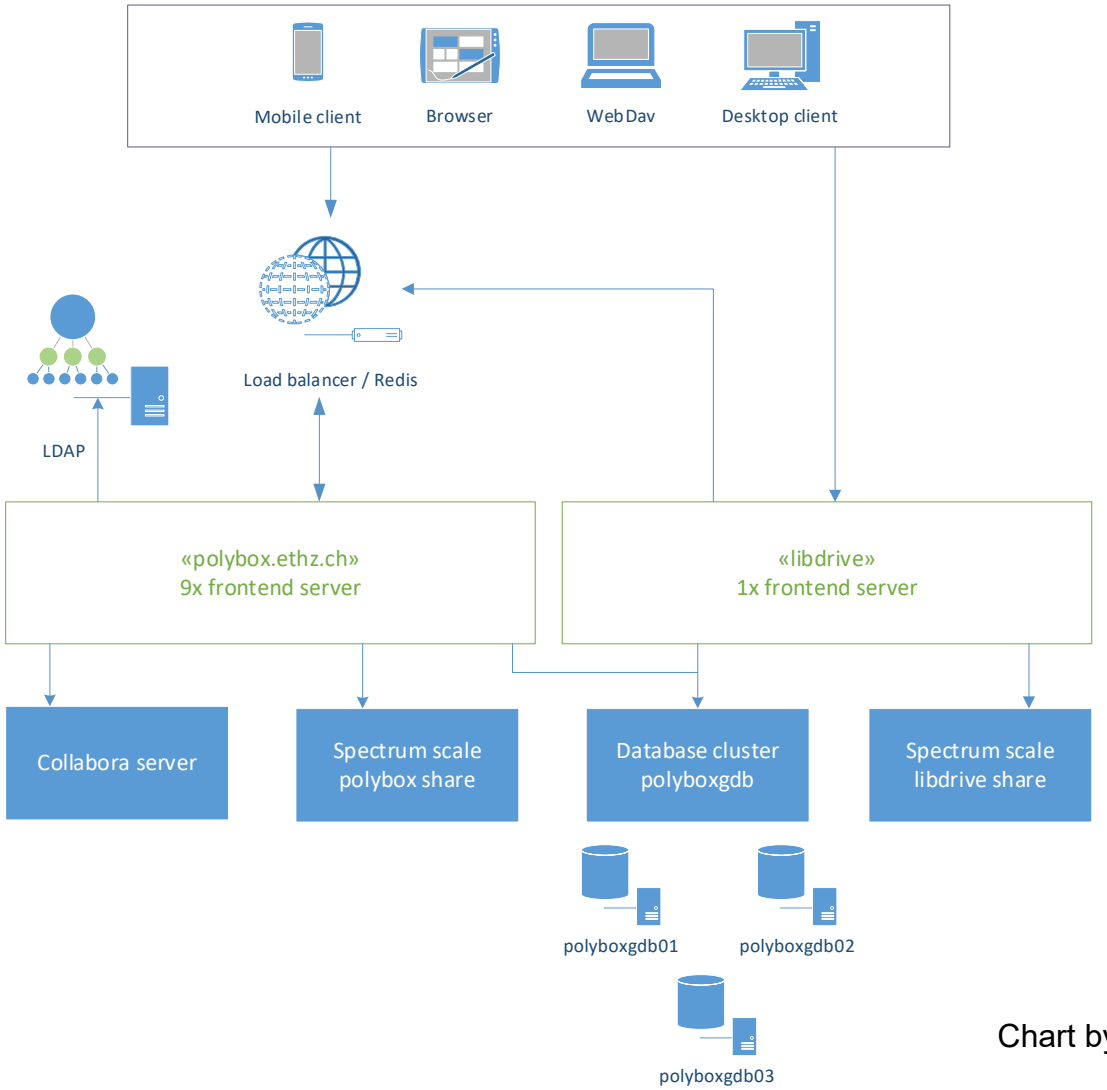


Chart by Gianluca Caratsch, ETH IT Services

Solution for publishing large datasets

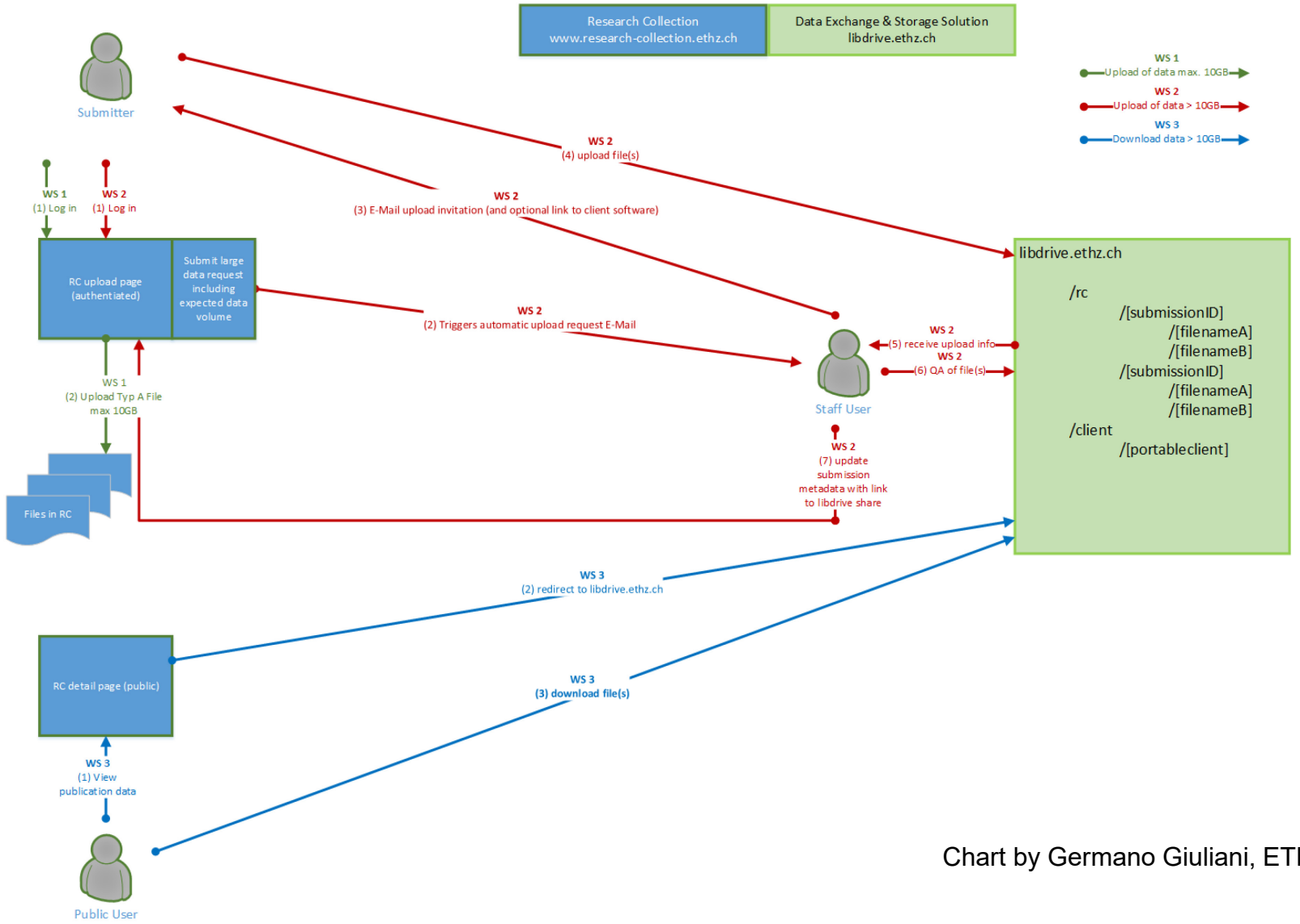


Chart by Germano Giuliani, ETH Library

Solution for publishing large datasets

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

What's next? Plans for the future

Certification

Applying for the Core Trust Seal

Geo-referencing

Improved geo-location referencing and search

Google Dataset search

Implement schema.org for inclusion of datasets

Barbara Hirschmann
Head E-Publishing
barbara.hirschmann@library.ethz.ch

ETH Library
Rämistrasse 101
8049 Zurich, Switzerland

www.library.ethz.ch