ETH zürich

ETH Library

Langzeitarchivierung von Forschungsdaten: Erfahrungen aus der Schweiz

Presentation

Author(s):
Töwe, Matthias

Publication date:
2025-09

Permanent link:
https://doi.org/https://doi.org/10.3929/ethz-c-000784994

Rights / license:
Creative Commons Attribution 4.0 International

This page was generated automatically upon download from the ETH Zurich Research Collection.
For more information, please consult the Terms of use.

ETH zürich

ETH-Bibliothek

Langzeitarchivierung von Forschungsdaten: Erfahrungen aus der Schweiz

Dr. Matthias Töwe
Forschungsdatenmanagement und Datenerhalt,
ETH-Bibliothek, ETH Zürich
25. September 2025, Frankfurt am Main

Inhaltsverzeichnis/Agenda

1. ETH Zürich und ETH-Bibliothek
2. „What‘s in a name?“
3. Lernkurve und Erkenntnisse
4. Policies und Communities
5. Workflow für Publikation und LZV von Forschungsdaten
6. Offene und neue Fragen
7. Ausblick

ETH Zürich

- Gegründet 1855
- 16 Departemente in den Fachgebieten
  - Architektur und Bauwissenschaften
  - Ingenieurwissenschaften
  - Naturwissenschaften und Mathematik
  - Systemorientierte Naturwissenschaften
  - Management- und Sozialwissenschaften
- 26’198 Studierende
- 525 Professorinnen und Professoren (FTE)
- 10’680 Mitarbeitende (FTE)

(Stand 2024)

ETH-Bibliothek

- Gegründet 1855
- Hauptbibliothek der ETH Zürich
  - inkl. drei Bibliotheken in Departementen
- Nationales Zentrum für technische und naturwissenschaftliche Information
- Sammelgebiete
  - Architektur und Bauwissenschaften
  - Ingenieurwissenschaften
  - Naturwissenschaften und Mathematik
  - Systemorientierte Naturwissenschaften
  - Management- und Sozialwissenschaften

ETH-Bibliothek

- Alte und Seltene Drucke
- Bildarchiv
- Graphische Sammlung
- Hochschularchiv der ETH Zürich
- Karten
- Kunstinventar der ETH Zürich
- Materialsammlung
- Max Frisch-Archiv
- Sammlung Sternwarte
- Sammlung wissenschaftlicher Instrumente und Lehrmittel
- Thomas-Mann-Archiv

„What‘s in a name?“

Langzeitarchivierung vs. Datenerhalt vs. Langzeitverfügbarkeit

- Implikationen von «Langzeit**archivierung**» (LZA) problematisch in der **Kommunikation**
- «**Datenerhalt**»: Versuch, auf kontinuierlichen **Prozess mit aktiven Massnahmen** zu verweisen
- «Langzeit**verfügbarkeit**» (LZV) macht das Ziel noch klarer und stellt **Nutzung in den Mittelpunkt**
  - Höhere Erwartungen an Auffindbarkeit und Online-Zugriff

«Langzeit»?

- Grundvoraussetzung für *ETH Data Archive*: Es muss eine Aufbewahrung für ***mindestens*** 10 Jahre geplant sein.
  - Aufgabenteilung mit dem institutionellen Repositorium *Research Collection*

Lernkurve

Beginn der LZV-Aktivitäten (2011/2012)

- Lösung gedacht für **das wissenschaftliche und kulturelle Erbe der Hochschule**
  - Doktorarbeiten und andere **Inhalte des Repositoriums**
  - Masterfiles der **Digitalisierung**
  - **Geschäftsunterlagen** des Hochschularchivs
  - **Forschungsdaten**

Frühe Erkenntnis

- LZV kann nur mit dem arbeiten, was noch ankommt
- Voraussetzungen müssen viel **früher und von der Datenentstehung an** geschaffen werden
- Unterstützung für das **Forschungsdatenmanagement** etablieren, um Ausgangslage zu verbessern

Herausforderung Forschungsdaten

- Sinnvolle LZV setzt umfassende **Dokumentation von Daten und Kontext** voraus
  - *«Garbage in – garbage out»*
- Services zur LZV stehen am **Ende der «Nahrungskette»**…
  - …und können kaum direkt Einfluss auf die Datenerzeugung nehmen
- Forschende sind meist die einzigen, die spezifische Meta-Informationen liefern können
  - Ihre **Prioritäten** richten sich in den seltensten Fällen nach der LZV
- Wachsende Erwartung an «**machine actionability**»
  - Kann das «Garbage»-Problem verstärken
- *«All kinds of everything»* - **Heterogenität** in jeder Hinsicht
  - Nach Fächern und Communities; hinsichtlich Methoden, Datentypen und -mengen, Infrastrukturen, Offenheit, Üblichkeit der Nachnutzung…
- Statische Einzelfiles sind zunehmend nicht die **relevante Einheit** oder können nicht allein stehen
  - Wachsende Abhängigkeit von anderen Daten und von Code

LZV-Wünsche und das echte (Forschenden-)Leben

Beispiele:

Wünsche aus Sicht LZV	In der Praxis
Offene, stabile, archivtaugliche **Dateiformate**	Für bestimmte Zwecke sind **proprietäre Formate unentbehrlich** und Konvertierung in alternative Formate bedeutet **Informationsverlust**
Technische Analyse auf Dateiebene	U.U. sehr viele Einzelfiles in gleichen Formaten > grosser Overhead an technischen Metadaten Aus Gründen der praktischen Handhabung werden **Container** mit vielen Einzelfiles erzeugt
Umfassende **Kontextdokumentation** in menschen- und maschinenlesbarer Form	«Ich habe hier noch ein Readme-File»
Unkomprimierte Daten	«Und **wer zahlt** den Speicher?»

Rückenwind

Schweizerischer Nationalfonds u.a.

- Anforderungen für Datenmanagementpläne (**DMP**s) ab ca. 2016
- Schrittweiser Aufbau von Informations- und Unterstützungsangeboten zu DMPs durch ETH-Bibliothek und Scientific IT Services

Institutionelle Research Data Management Guidelines (RDMG)

- Längerer Erarbeitungsprozess in einer institutionellen **Arbeitsgruppe Open Science** ab 2018
- In Kraft seit Juli **2022**
- Machen **Rahmenvorgaben für Datenpublikation, Aufbewahrungsfristen, Verantwortlichkeiten** usw. – aber ohne Zwang zur Ablieferung von Daten in Research Collection oder ETH Data Archive
- Vermeiden detaillierte Top-Down-Vorgaben
- Betonen **Rolle der Fach-Communities**, um zu angepassten, praktikablen Lösungen zu kommen

Rolle der Communities

Optimistische Erwartung

- **Forschende einigen sich selbst auf Best Practices**, implementieren diese für sich und fordern dies auch von Kolleginnen und Kollegen ein

Wahrscheinlicher

- Nur wenige fachliche Communities verfügen über Strukturen, um sich auf bindende Standards zu einigen
- Es stellt sich jeweils eine **Balance zwischen Anforderungen** von Geldgebern, Herausgeberinnen und den **Ressourcen und Interessen** von Forschenden ein – mit Ausreissern
- Es wird einen **Kosmos unterschiedlicher Lösungen** geben für FDM und damit i.d.R. auch für die LZV mit einer Bandbreite von Antworten auf konkrete Fragen
  - Was wird **wie lange** als **relevant** bewertet?
  - Wer setzt die **Kriterien** oder nimmt sogar Bewertung vor?
  - Was bedeutet eine solche Bewertung für die tatsächliche **Aufbewahrungsdauer**?
  - In welcher **Form** sollen (welche) Daten sinnvollerweise vorgehalten werden?
  - Welche **Erwartungen** hat die Community **aneinander** im Hinblick auf Dokumentation und Zugänglichkeit?

Workflow für Publikation und LZV von Forschungsdaten

Forschung → *Publikation* → *Langzeitverfügbarkeit*

Research Collection (RC)

Pakete < 50 GB: Metadaten und Daten in RC

Pakete > 50 GB: Metadaten in RC

Periodischer **Export** → **ETH Data Archive**

Upload durch Forschende oder Data Stewards

Daten in libdrive

via IT-Support

Long Term Storage (LTS)
Interne Speicherung auf Tape ohne strukturierte Metadaten und direkten Zugriff;
keine Lenkungsabgabe

libdrive
- Lokale Cloud
- 10 Jahre
- Offener Zugang

LZV-Option im Publikationsworkflow der Research Collection

- LZV explizit anwählen im Upload-Prozess der Research Collection
- Team der Research Collection prüft Inhalt und empfiehlt gegebenenfalls weitere Massnahmen (z.B. für Dateiformate gemäss Empfehlung: https://unlimited.ethz.ch/x/8QeSCw)

Langfristige Nutzbarkeit sicherstellen:

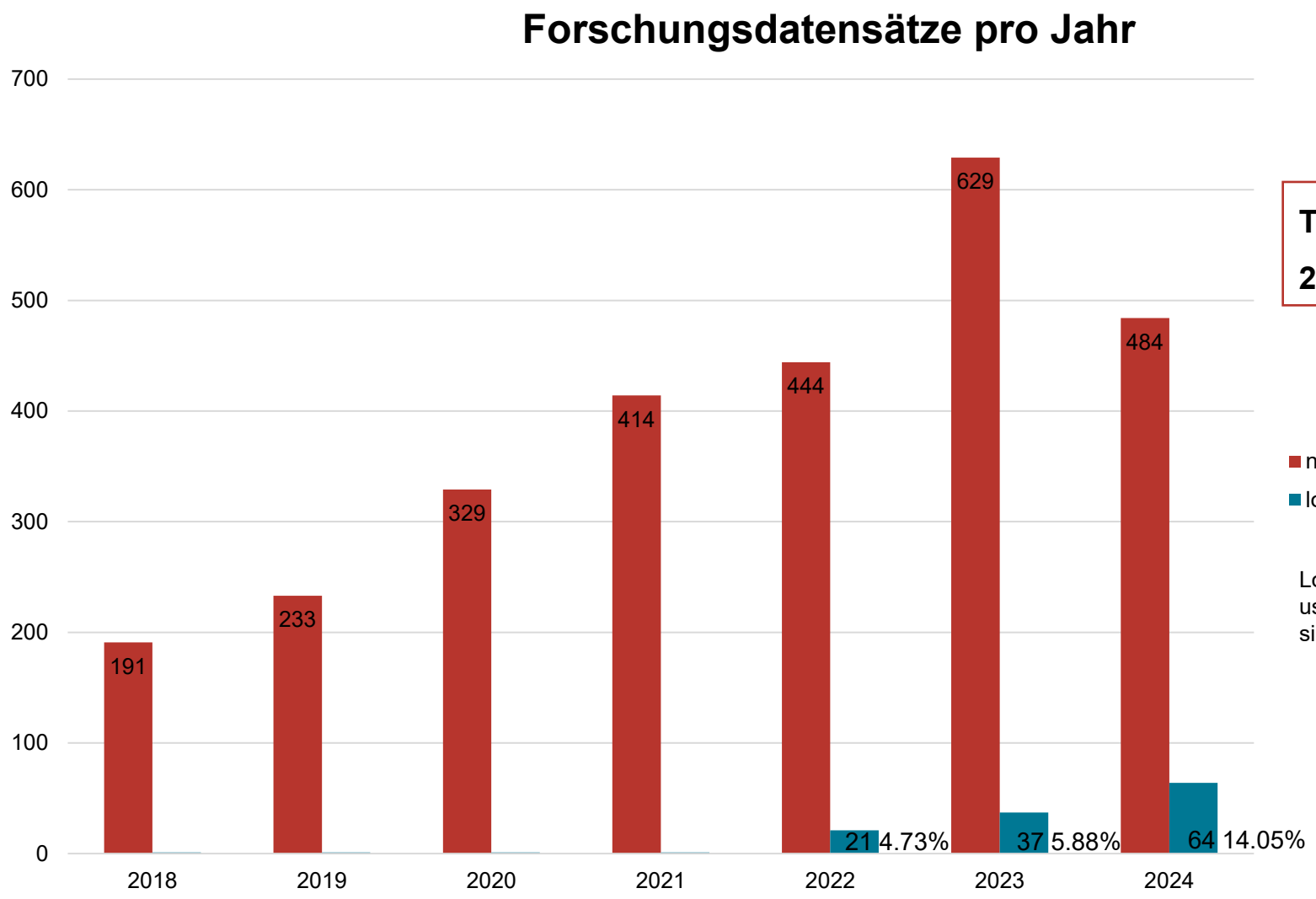
☑ Ich lege Wert darauf, dass meine Daten langfristig (mehr als 10 Jahre) nutzbar bleiben.

Wenn Sie diese Option wählen, werden wir die langfristige Nutzbarkeit Ihrer Daten bewerten und ggf. Maßnahmen vorschlagen (z. B. Konvertierung von Dateien in verschiedene Formate). Weitere Informationen finden Sie in unserem Handbuch

Nachdem Sie die Bearbeitung Ihres Eintrag abgeschlossen haben, prüft das Research Collection Team die **Dateiformate** und **zusätzliche Anforderungen für die langfristige Nutzbarkeit** (vgl. unten). Das Research Collection Team gibt Ihnen im Anschluss eine Empfehlung ab.

Wir empfehlen vorgängig selbstständig die Forschungsdaten auf nachfolgende Anforderungen für langfristige Nutzbarkeit zu prüfen.

Forschungsdaten-Ablieferungen 2018-2024

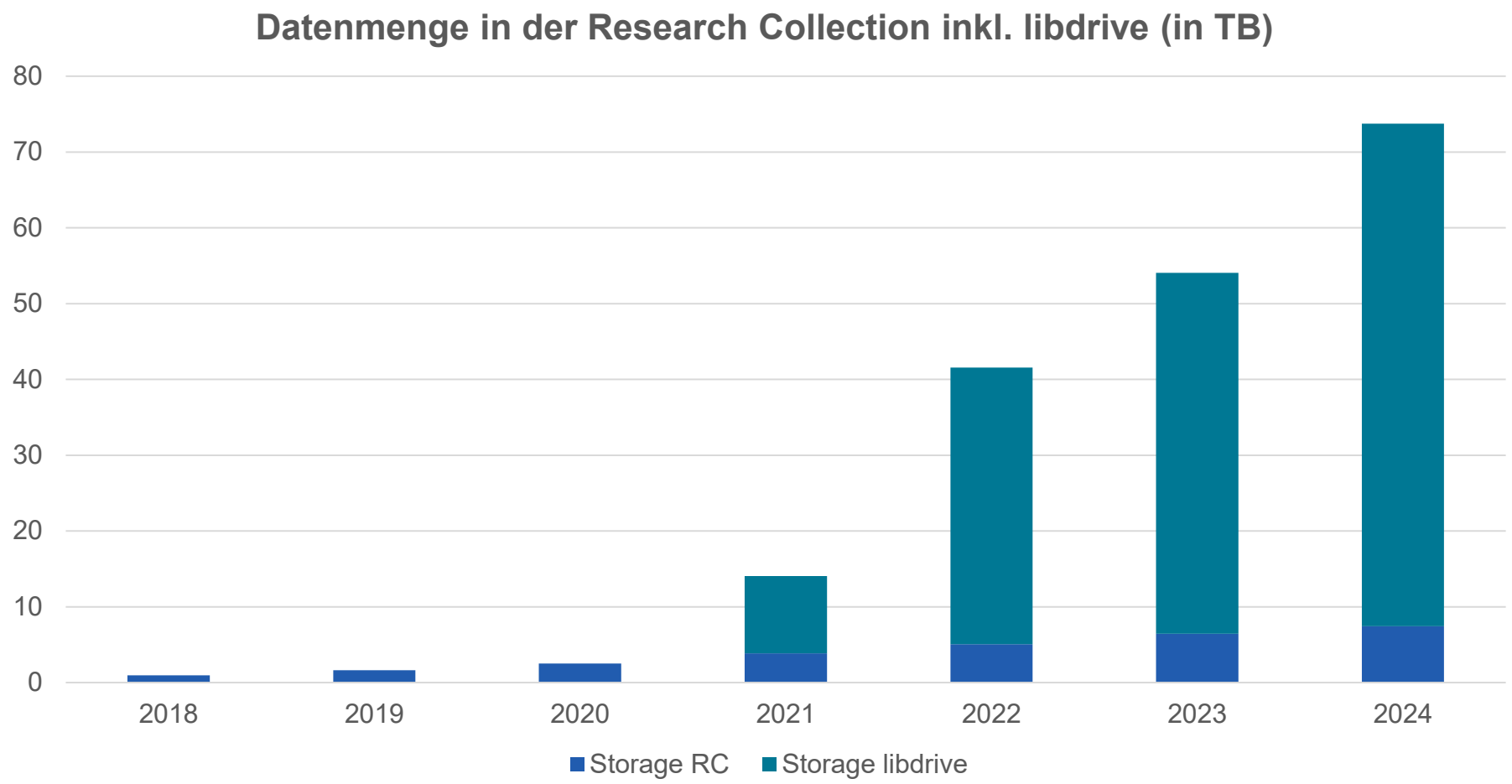


Total Datensätze Ende 2024:

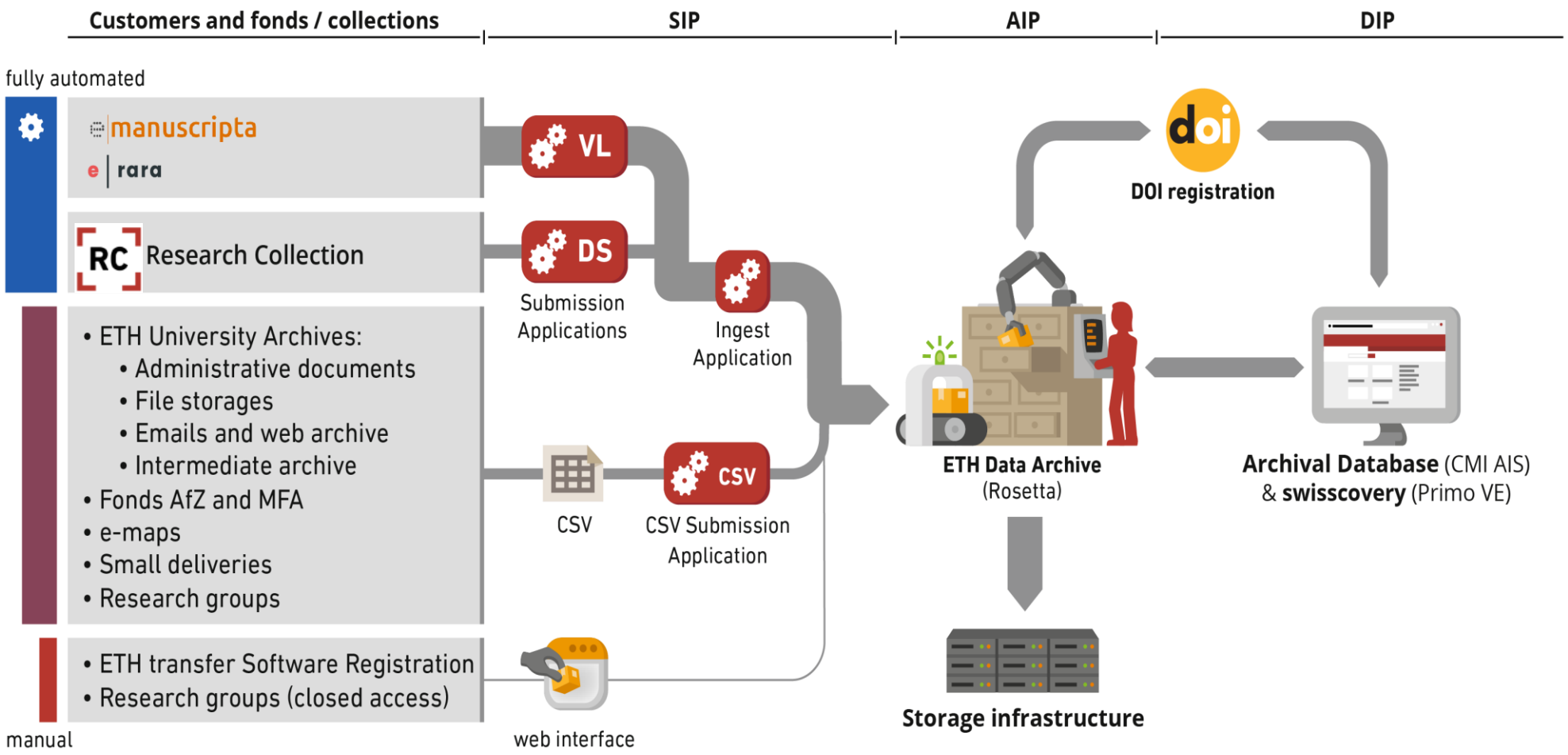
2’724 (von 272’373 Einträgen total)

Long-term usability option since 2022 !

Datenmenge


Datenmenge in der Research Collection inkl. libdrive (in TB)
80
70
60
50
40
30
20
10
0
2018
2019
2020
2021
2022
2023
2024
Storage RC
Storage libdrive

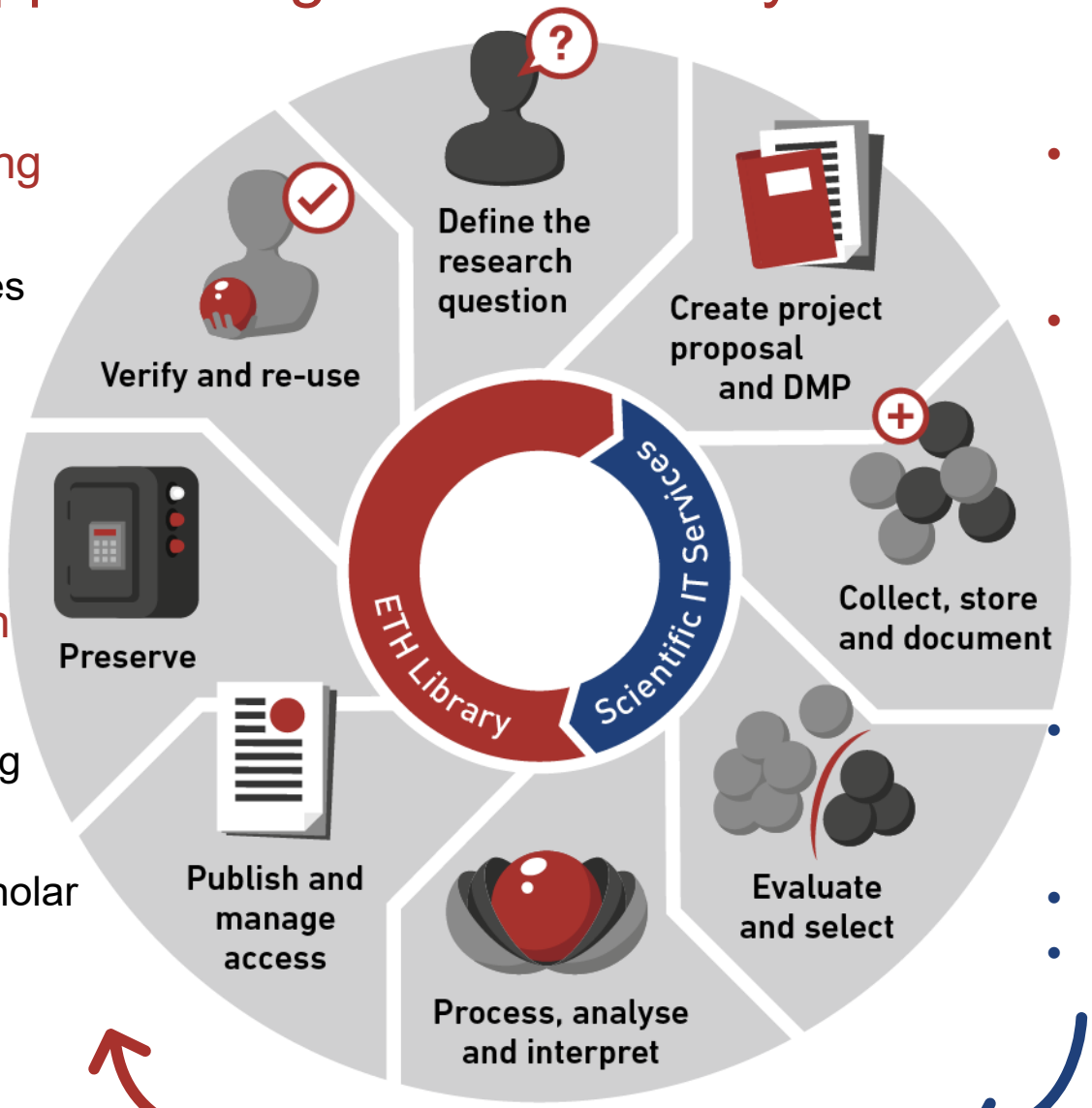
Daten *aus* der Forschung? – Es gibt auch Daten *für* die Forschung!



VL = Visual Library, source system of digitized material from e-rara and e-manuscripta
DS = DSpace, underlying software of the Research Collection

Services and support along the Data Lifecycle at ETH Zurich

- Preservation and Archiving Services
  - ETH University Archives
  - ETH Data Archive
  - Preservation planning
  - Metadata preparation
- ETH Research Collection
- Publication services
  - Open access publishing
  - DOI registration
  - Indexing in Google Scholar



- Reference management software: licences and courses
- DMP guidance, templates, and review:
  - http://www.snf.ch
    → DMP Guidelines
  - https://documentation.library.ethz.ch/x/WwKsB
    → DMP Instructions
- Active data management services (storage, wikis, openBIS etc.)
- High-performance computing
- Research with strictly confidential data

ETH library: https://library.ethz.ch

Advice / Training / Infrastructure

Scientific IT Services: https://sis.id.ethz.ch/

Illustration CC-BY, Andres Bucher

Offene und neue Fragen

Wie gehen wir um…

- …mit dem Bedarf von Forschenden, die **Zugangsgewährung für nicht-öffentliche Daten** an uns zu delegieren?
- …mit der Langzeitverfügbarkeit von **sensiblen Daten** generell?
- …mit der Langzeitverfügbarkeit von Daten, die **nach heutigem Stand als anonymisiert** gelten, aber aufgrund ihres Charakters in Kombination mit anderen Daten zumindest theoretisch eine Identifizierung erlauben könnten – vielleicht auch erst in der Zukunft?
- …der zukünftigen **Bewertung** von Daten, die ursprünglich nur für zehn Jahre abgeliefert wurden?
- …der wirklich **systemunabhängigen** Definition der teils aufwändigen Prozesse?
- …unserem eigenen Anspruch, Daten **funktional** zu erhalten, d.h. technisch (Bitstrom), **logisch** und **semantisch**?

Wie weiter?

- **Kuratierung** stärken – wie weit kann die Rolle der Bibliothek hier reichen?
  - oder ist die Frage eher: **Wer macht was**?
- **Pipelines** mit stärkerer Integration mit LIMS/ELN-Systemen zur weiteren Automatisierung schaffen? Diese könnten analog zur Einführung von Records Management Systemen in der Verwaltung frühzeitig für eine Strukturierung und Kontextualisierung sorgen.
- Müssen wir uns stärker um **Vollständigkeit** bemühen oder darf es eine wissenschaftsgetriebene **Selektivität** geben?
- **Akzeptieren**, dass sowohl Individuen als auch ganze Fächer hinterherhinken werden – die LZV ist kein Selbstzweck.

Dr. Matthias Töwe
Leitung Forschungsdatenmanagement und Datenerhalt
matthias.toewe@library.ethz.ch

ETH-Bibliothek
ETH Zürich
Rämistrasse 101
8092 Zürich

library.ethz.ch