ETH BIBLIOTHEK

Structure and Collections in Rosetta – from Ingest to Delivery

Rosetta Advisory Group, 6th Annual Meeting

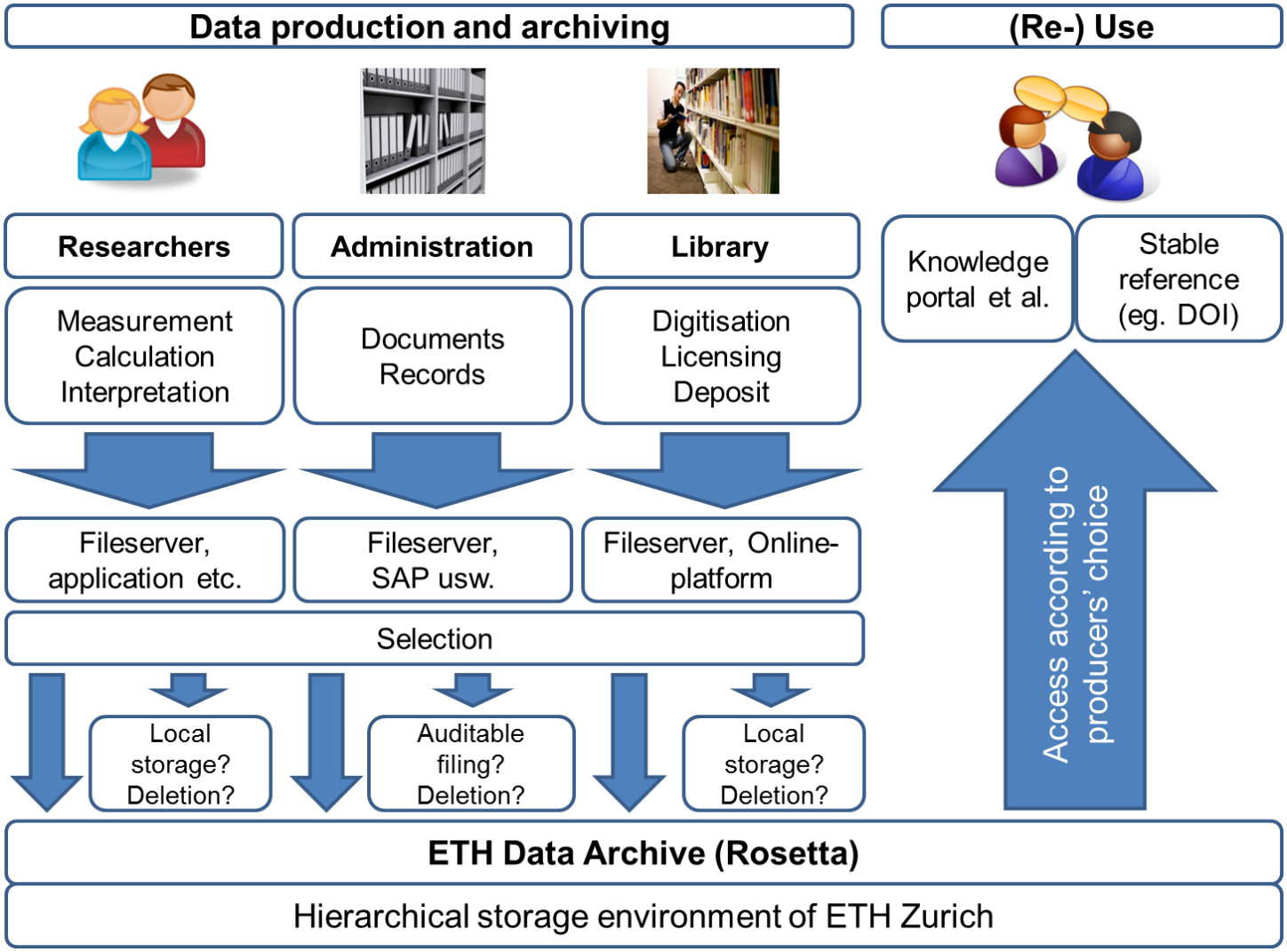
Leuven, 3 June 2015

Matthias Töwe, Head Digital Curation, ETH-Bibliothek, ETH Zurich

Overview

- Use of Collections
- Source and Pre-Ingest
- Technical Analyst Workbench
- Discovery
- Access Rights
- Delivery
- Issues

Vision for diverse data types



View of Research Data Producers

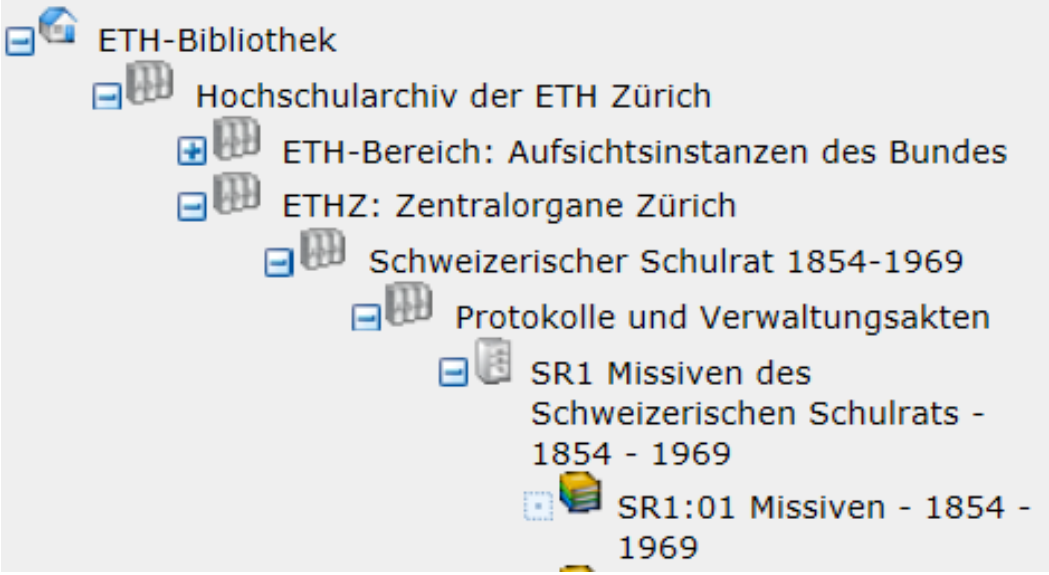
- **Folder structure contains information which is**
  - relevant for **intellectual re-use**
  - required for **re-use in automated processing** in a defined algorithm
- **Directory structure should be preserved in a form which is accessible to both data producer and potential future re-users**

Typical customer requests

- *«I want to submit a **manuscript for an article**. The editor demands that **raw data** is deposited in a **repository.** What can I do?»*
- *«Up to now we have been archiving **data to our doctoral theses on CD-ROM** and we would like to find a better solution. How can we do this? -- Oh, and our professor is retiring in n months.»*
- *«We want to **link** from an article **research data which we would like to be publicly available**. We have analysed the data with our own methods and now other groups might want to look into them with their methods. How can we do this?»*

View of University Archives for Administrative Records

- In Continental European archival tradition, **hierarchical structure is used to reflect provenance in particular**
- **A dedicated Archival Information System is used to manage this information (internal database plus separate public view)**



Why manage Collections in Rosetta?

- **Collection information is not only relevant for immediate access**
- **Preserve context information** which is contained in structure
- **There is no CMS available** outside of Rosetta and Rosetta is metadata master (e.g. Research Data)
- A CMS exists, **but the DPS should hold the structural information** as part of its exit strategy
- **Enable navigation** between IEs and collections in Rosetta **without having to switch to the CMS every time**

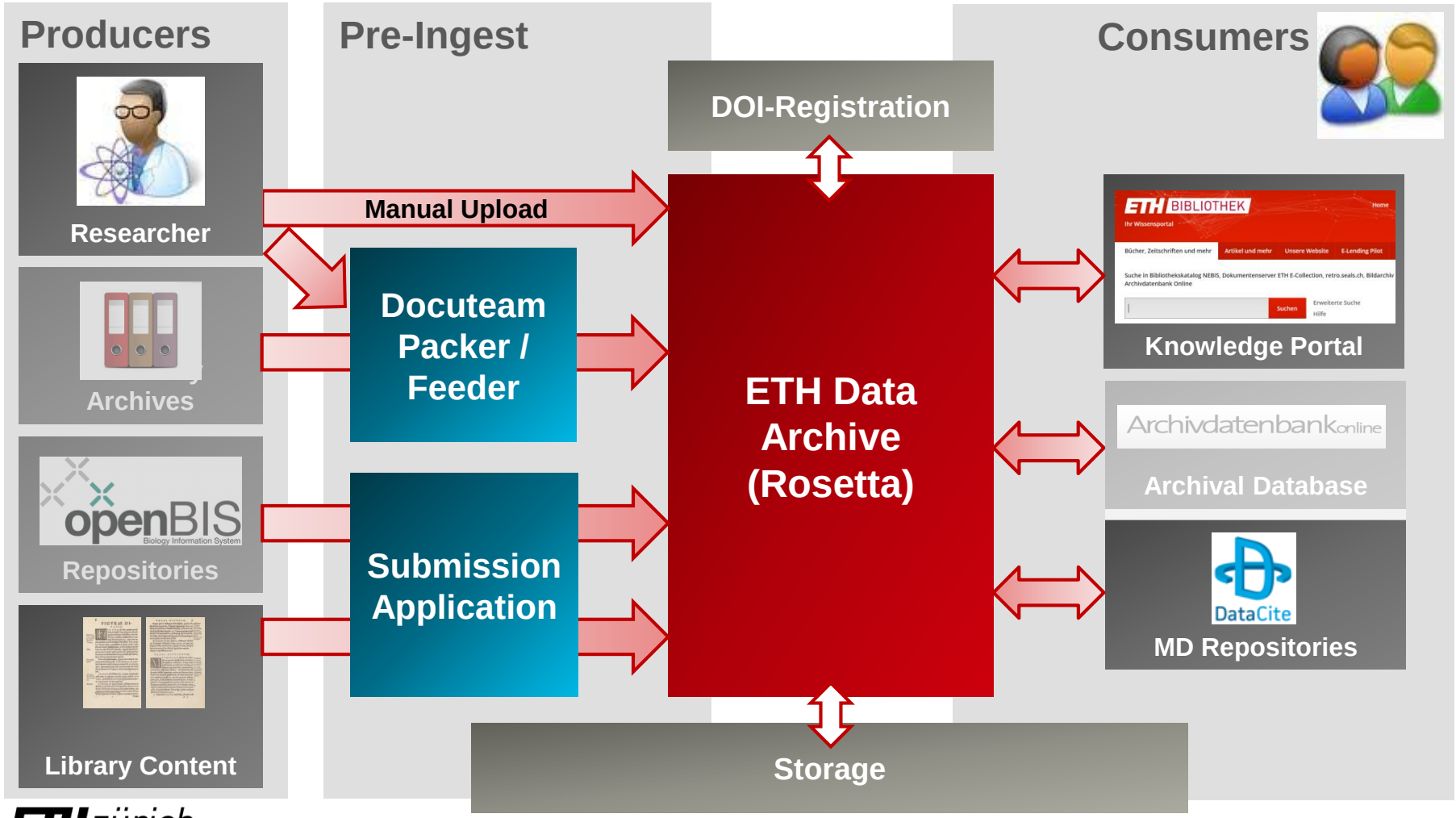
Reflect structure in Rosetta

- **Possibilities**
  - **Complex IE**
    > Structure reflected in ToC [case open]
  - **Rosetta Collections**
    > Reflecting at least the original folder structure
    > Enrichment for creating additional collections based on metadata > facets
  - **Simple IE with Container File**
    > «Quick and a bit dirty»:
    no re-use of individual files expected, format identification possible, preservation capabilities restricted

Methods for Deposit and Upload

- **Manually**:
  Web dialogue for upload and metadata capture
- **Semi-automatically**:
  Batch-Upload of files with existing metadata (CSV)
- **Automatically**:
  - Adapted Submission Application packages structured files with existing metadata in XML-format
  - Submission Application can also be implemented as interface with an existing source application
- **Automatically after manual preparation** in docuteam packer (local viewer and editor for file-structure and metadata)

Rosetta as part of the systems landscape



(Illustration: Franziska Geisser)

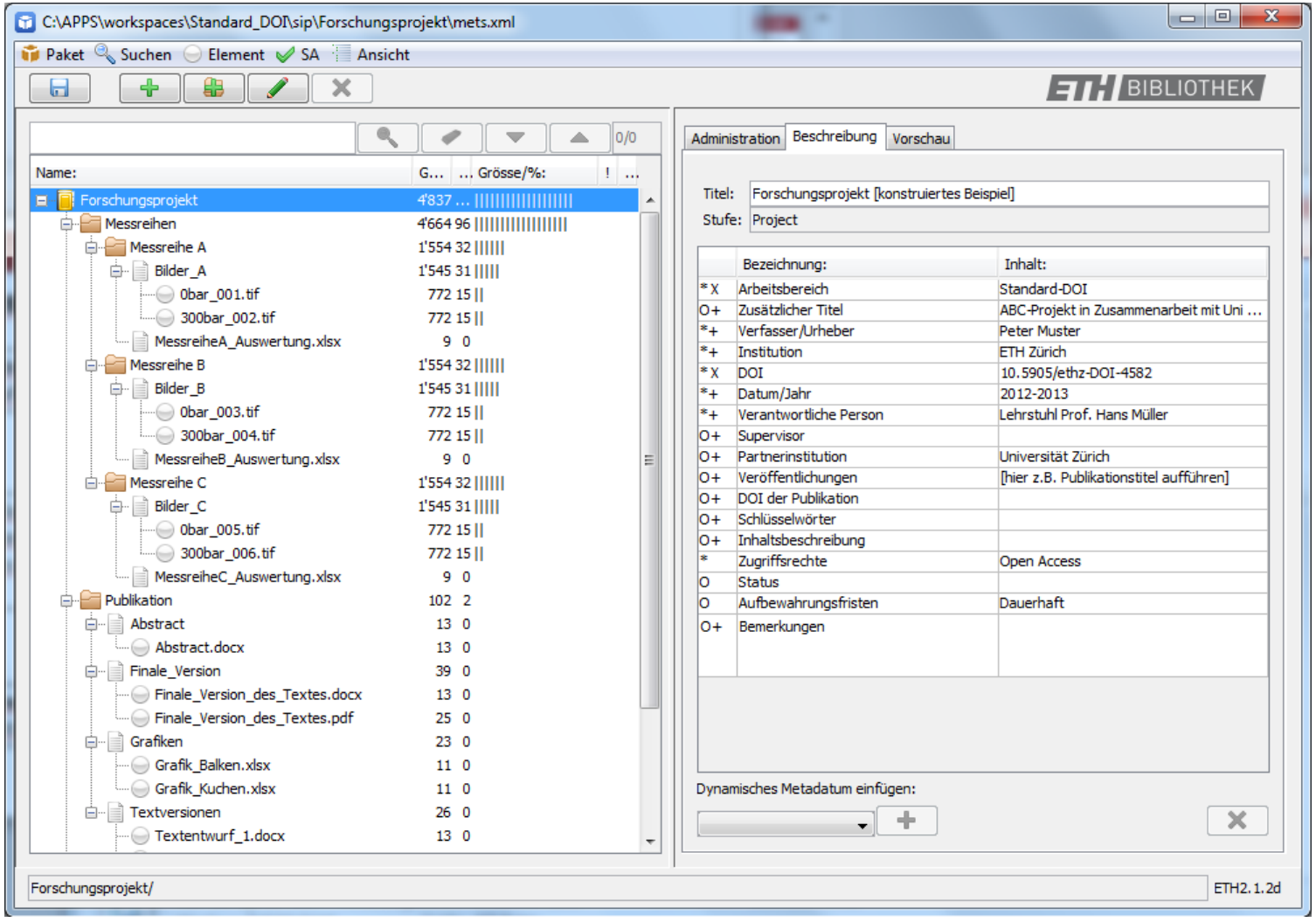
Collections in Sources and in Pre-Ingest I

- **Directory structures on the file system**
  - **Each file belonging to only one collection**
  - often **no metadata**, sometimes available from external source
  - Use **directory loader**
  - Use **ftp upload**
  - Use **CSV-Upload**
  - **IE holds reference to collection** in dcterms:isPartOf

Collections in Sources and in Pre-Ingest II

- **SIP with folder structure created in docuteam packer**
  - **Each file belonging to only one collection**
  - **Metadata on folder and file level**
    > facilitate creation of logical collections in Rosetta
  - Use **docuteam feeder as submisson application**

Collections in Sources: docuteam packer



More Collections from Enrichment

- **Creation of «new» logical collections in enrichment** based on metadata fields
  - Motivated by practical reasons – easier handling and navigation
  - Reflect additional context information while it is still available
- **Potentially:**
  Source system bundling files (or other objects) based on different criteria such as technical, descriptive or administrative metadata

Technical Analyst Workbench

- **TA operations via UI can be applied to one file at a time only** or files need to be selected individually (e.g. replace file or even «skip»)
- **Impractical when ingesting thousands of files**
- **«Remember my decision» is applied with reservations**
  - **Requires re-run and can have side-effects** in cases where the created rule's scope is too wide

Comparison of Upload-Processes

Method	Metadata	Structural view	Download / Export	Use case
ZIP/Tar Upload	Level IE = Package, via MD Form	1 File, view structure in ZIP tool	Download whole package in Viewer	Research group «locker»
Directory Loader (Aurigma)	Level IE = Package, via MD Form	Logical Structmap nice in theory, but not reliable, possibly due to browser issues?	Download: individual files only (format and browser dependent); Export: folder structure is duplicated and unusable [open case]	LTP; structure and diversity of format issues not too broad; Limits in size and number unclear
CSV Upload	Level SIP, Collection, IE, Representation, File, via CSV-File	Depending on options; files in one IE are «flat»	Download: individual files only (format and browser dependent)	Too complicated and error prone for manual process; rather use docuteam packer or automated CSV creation
FTP Upload	Level IE = Package, MD in METS.xml	[Not tested due to open case]	Download: individual files only (format and browser dependent);	Similar to Directory Loader, but less restricted? No folder structures possible?

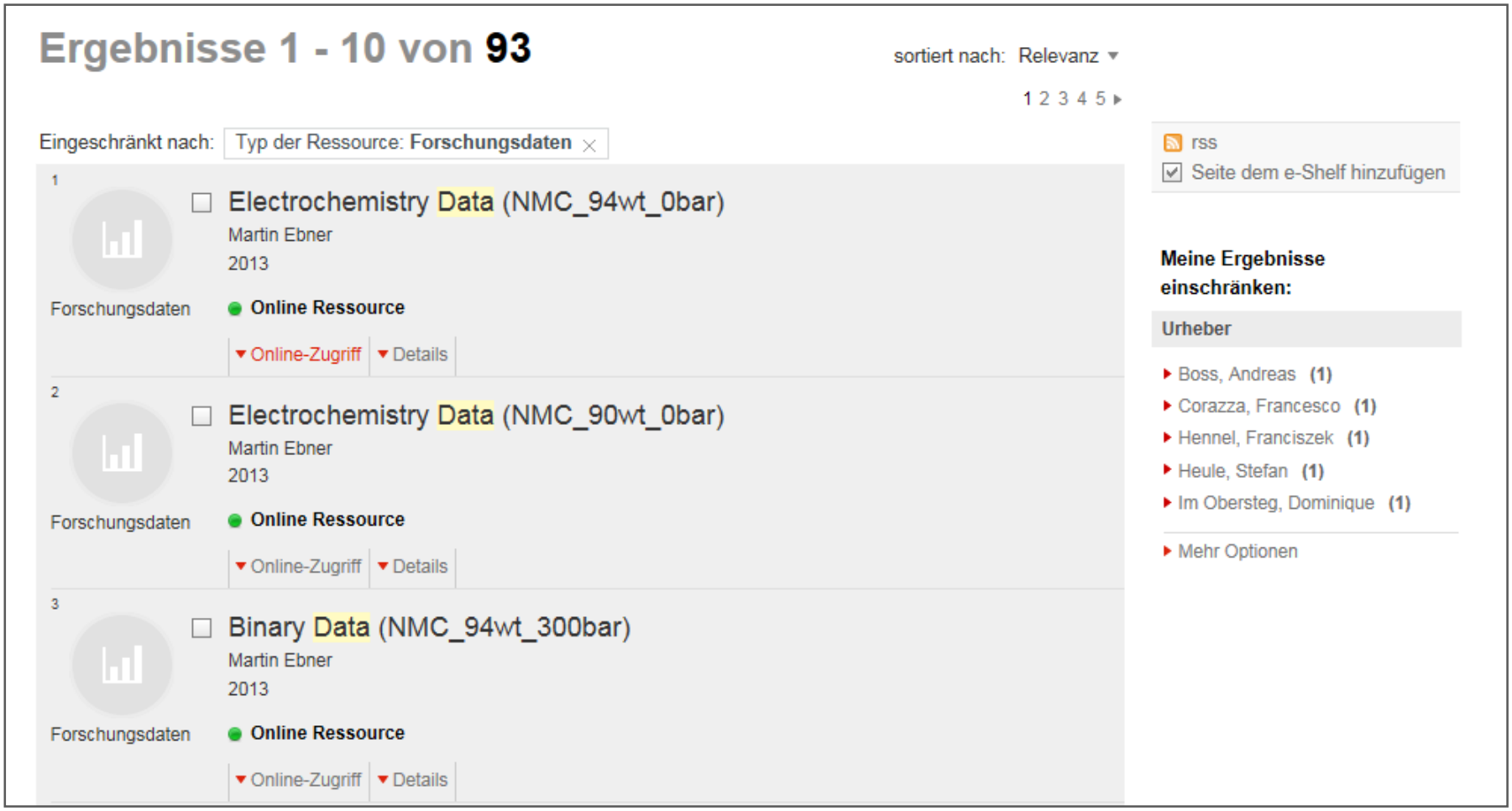
Preservation Perspective?

Method	Format Validation	Handling in TA	Use case	Preservation measures?
ZIP/Tar Upload	Yes, for Bytestreams and Container	Container can be «skipped»	Research group «locker» with a lot of undocumented formats	Bitstream preservation only. Format migration not expected and hardly possible.
Directory Loader (Aurigma)	Yes	On a file by file basis	LTP; structure and diversity of format issues not too broad; limits in size and number unclear	«As usual», depending on success of format validation and available tools.
CSV Upload	Yes	On a file by file basis	Too complicated and error prone for manual process; rather use docuteam packer or automated CSV creation	
FTP Upload	Yes	On a file by file basis	Similar to Directory Loader, but less restricted? No folder structures possible?	

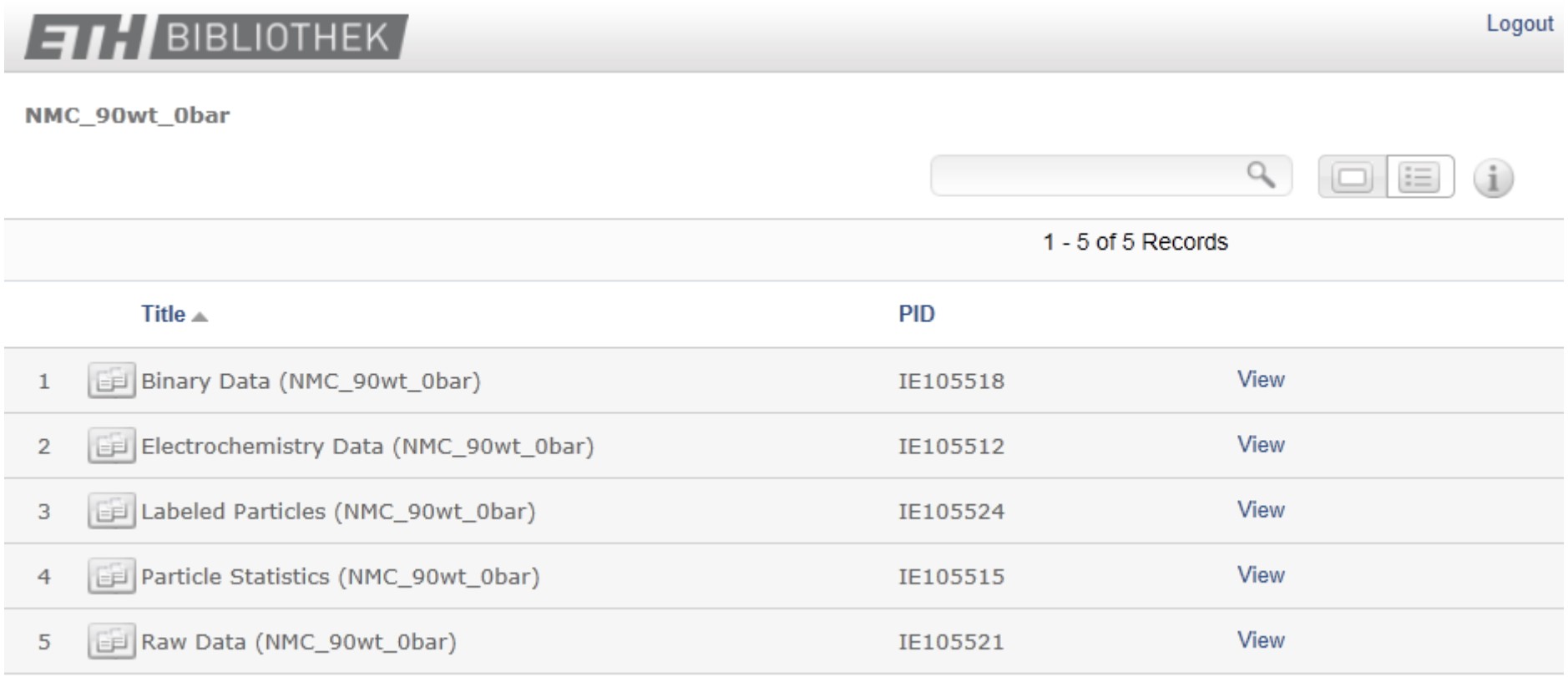
Publishing und Delivery

- «**Discovery**» of IEs and Collections is performed outside of Rosetta:
  - OPAC / Primo
  - Archival database
  - Laboratory Information and Management System?
- «**Publishing**» via OAI-PMH, SRU Request or Web Service
- «**Delivery**» from Rosetta:
  - Depending on defined Access Rights
  - Navigation in Collection Viewer switched on or off

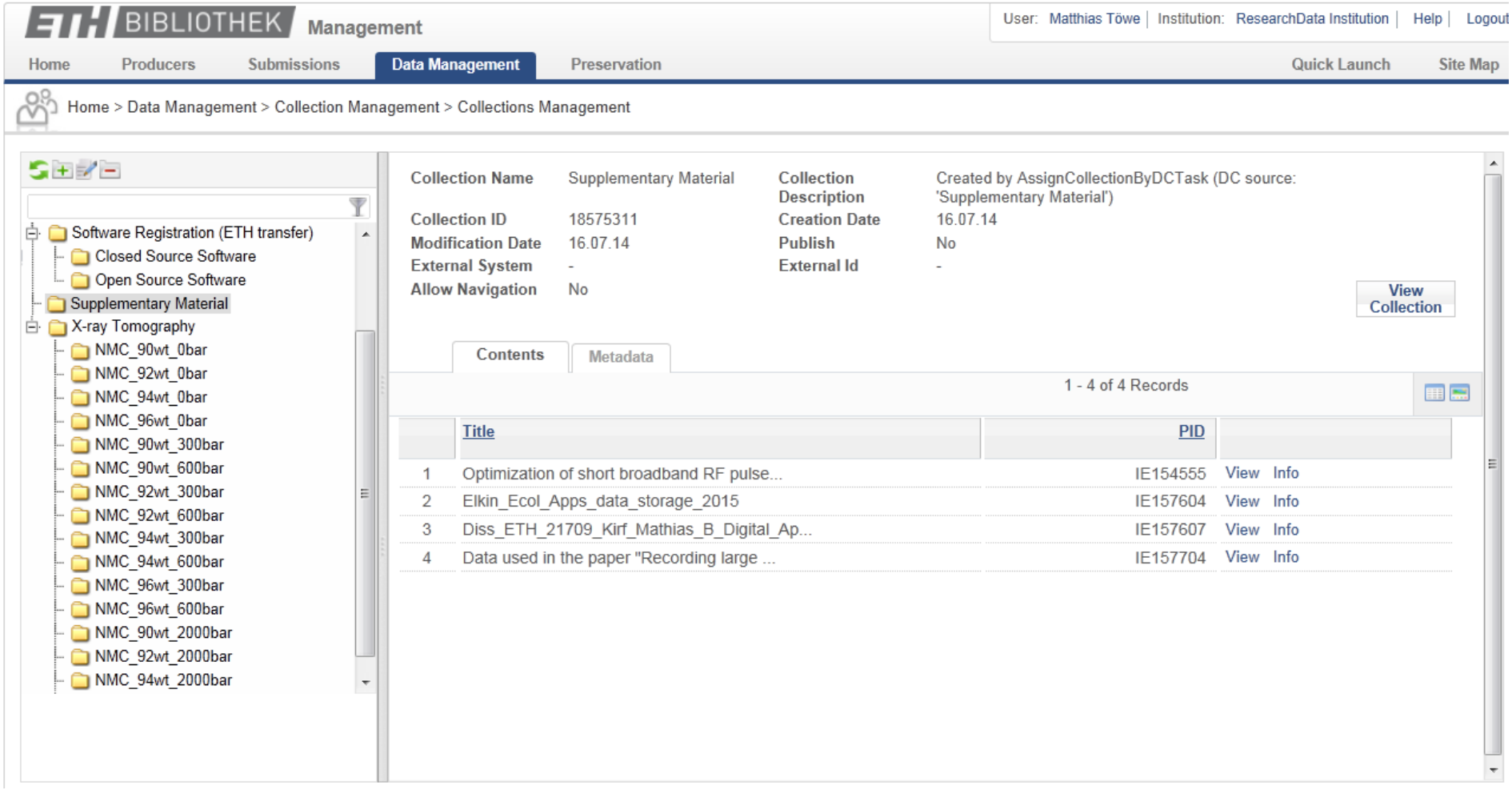
Discovery in Knowledge portal (Primo): Only IEs are published in «flat» form



Collection Viewer
IE Access Rights enforced; Navigation on / off



Collection Management Controlled by Ownership



Access Rights

- **We welcome suggestions** how to achieve the following (see examples on next slide):
  - Hide embargoed Collections' metadata when navigating in Rosetta
  - Handle bytestreams of a complex IE with shared descriptive metadata, but with individual Access Rights

Access Rights - Collections

- Manage Access Rights exclusively in external CMS?
- Use case University Archives
  - The **existence of certain collections must not be displayed** for several decades. (This is not about IEs contained within the collections, but about the collections themselves.)
  - **How hide embargoed collections' metadata** from view?
    Archival Information System allows control of what is published to Primo, but collection metadata can in principle be seen in Collection Management

Access Rights - File

- **Complex IE** with individual files having different access rights, for example:
  - **Full text** of the PhD thesis (parts of which have already been published in scholarly journals): no access (staff only)
  - **Abstract** of the PhD thesis: open access
  - **Raw data** underlying the PhD thesis: access restricted to research group or university members

    or
  - **Digitized book** is publicly available as a whole, but it contains an illustration that is subject to third party copyrights. So only the **page containing the illustration needs to have an access restriction**.

Questions?

www.library.ethz.ch/Digital-Curation

data-archive@library.ethz.ch

Dr. Matthias Töwe
Head Digital Curation
ETH-Bibliothek
Rämistrasse 101
8092 Zurich
044 632 60 32
matthias.toewe@library.ethz.ch