



# Structure and Collections in Rosetta – from Ingest to Delivery

Rosetta Advisory Group, 6th Annual Meeting

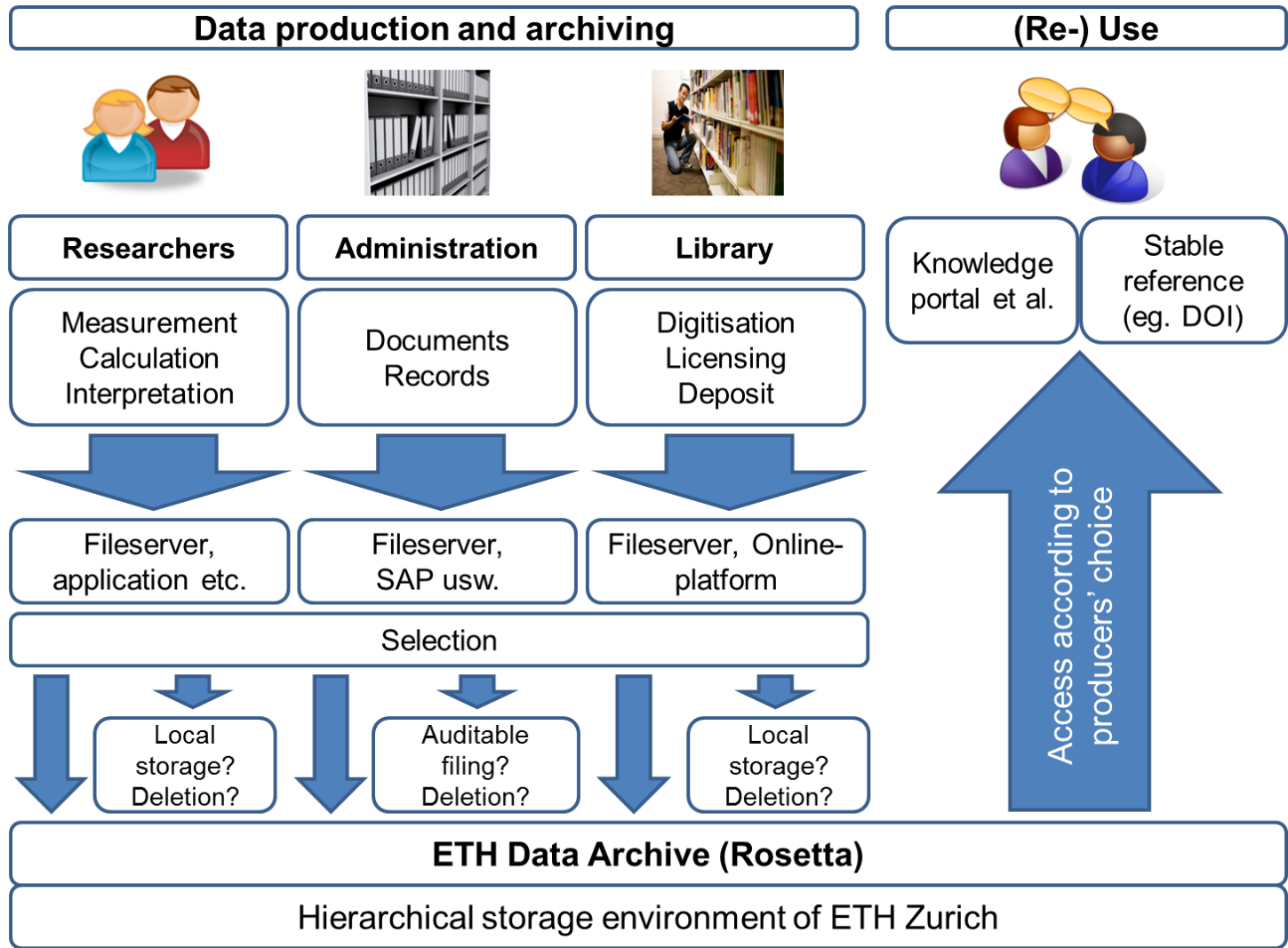
Leuven, 3 June 2015

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

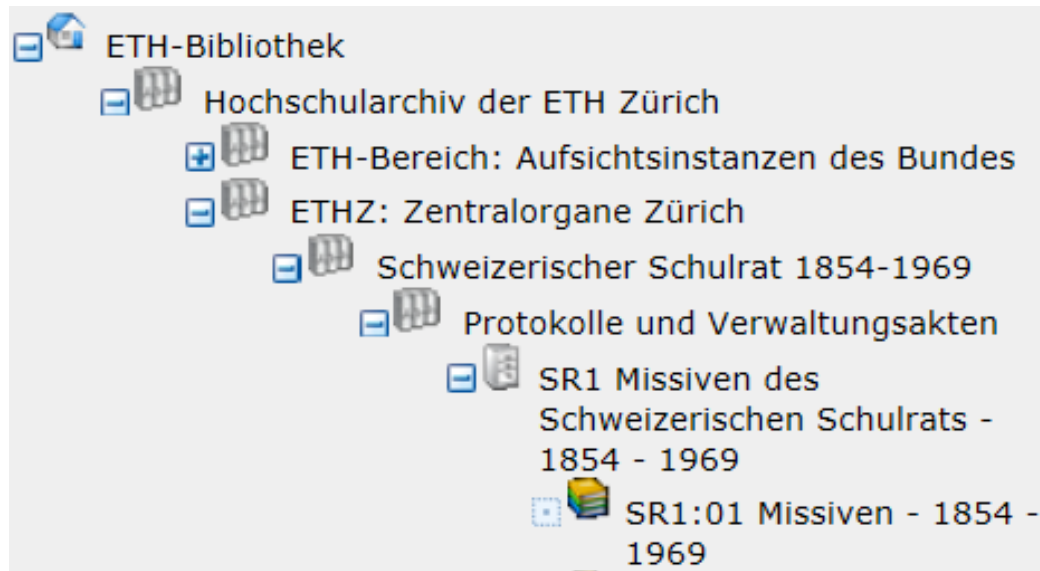
- **Folder structure contains information which is**
  - relevant for **intellectual re-use**
  - required for **re-use in automated processing** in a defined algorithm
- **Directory structure should be preserved in a form which is accessible to both data producer and potential future re-users**

# 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»:

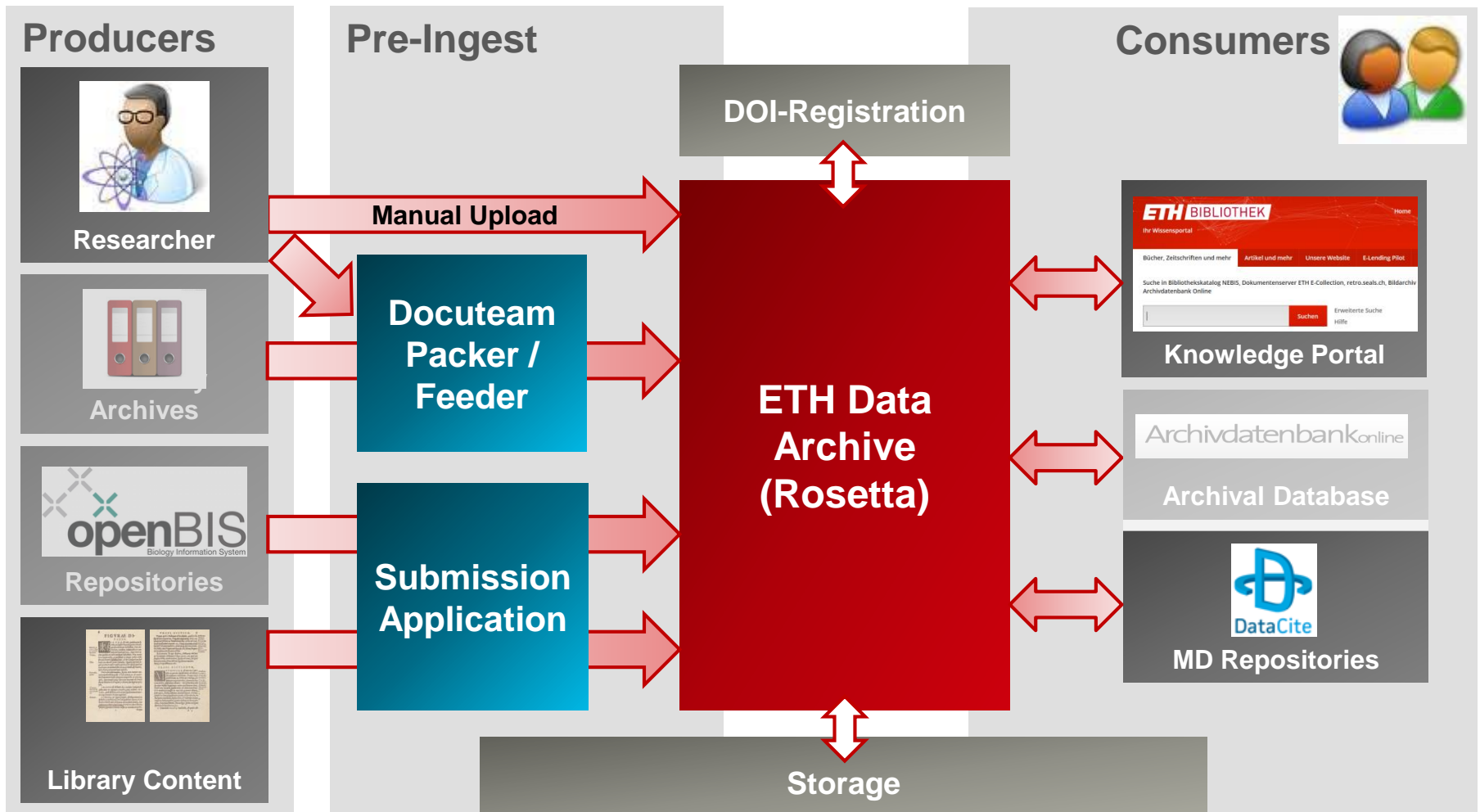
- no re-use of individual files expected, format identification possible, preservation capabilities restricted



# 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



# 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 submission application**

# Collections in Sources: docuteam packer

The screenshot shows the docuteam packer application window. The left pane displays a hierarchical file tree for a project named 'Forschungsprojekt'. The right pane shows the 'Administration' tab with a metadata form.

**File Tree (Left Pane):**

Name:	Größe/%
Forschungsprojekt	4'837 ...
Messreihen	4'664 96
Messreihe A	1'554 32
Bilder_A	1'545 31
Obar_001.tif	772 15
300bar_002.tif	772 15
MessreiheA_Auswertung.xlsx	9 0
Messreihe B	1'554 32
Bilder_B	1'545 31
Obar_003.tif	772 15
300bar_004.tif	772 15
MessreiheB_Auswertung.xlsx	9 0
Messreihe C	1'554 32
Bilder_C	1'545 31
Obar_005.tif	772 15
300bar_006.tif	772 15
MessreiheC_Auswertung.xlsx	9 0
Publikation	102 2
Abstract	13 0
Abstract.docx	13 0
Finale_Version	39 0
Finale_Version_des_Textes.docx	13 0
Finale_Version_des_Textes.pdf	25 0
Grafiken	23 0
Grafik_Balken.xlsx	11 0
Grafik_Kuchen.xlsx	11 0
Textversionen	26 0
Textentwurf_1.docx	13 0

**Metadata Form (Right Pane):**

Administration Beschreibung Vorschau

Titel: Forschungsprojekt [konstruiertes Beispiel]  
 Stufe: Project

Bezeichnung:	Inhalt:
* X Arbeitsbereich	Standard-DOI
O+ Zusätzlicher Titel	ABC-Projekt in Zusammenarbeit mit Uni ...
*+ Verfasser/Urheber	Peter Muster
*+ Institution	ETH Zürich
* X DOI	10.5905/ethz-DOI-4582
*+ Datum/Jahr	2012-2013
*+ Verantwortliche Person	Lehrstuhl Prof. Hans Müller
O+ Supervisor	
O+ Partnerinstitution	Universität Zürich
O+ Veröffentlichungen	[hier z.B. Publikationstitel aufführen]
O+ DOI der Publikation	
O+ Schlüsselwörter	
O+ Inhaltsbeschreibung	
* Zugriffsrechte	Open Access
O Status	
O Aufbewahrungsfristen	Dauerhaft
O+ Bemerkungen	

Dynamisches Metadatum einfügen: [Dropdown] + [X]

ETH2.1.2d

# 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
<b>ZIP/Tar Upload</b>	Level IE = Package, via MD Form	1 File, view structure in ZIP tool	Download whole package in Viewer	Research group «locker»
<b>Directory Loader (Aurigma)</b>	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
<b>CSV Upload</b>	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
<b>FTP Upload</b>	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?
<b>ZIP/Tar Upload</b>	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.
<b>Directory Loader (Aurigma)</b>	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.
<b>CSV Upload</b>	Yes	On a file by file basis	Too complicated and error prone for manual process; rather use docuteam packer or automated CSV creation	
<b>FTP Upload</b>	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

Ergebnisse 1 - 10 von 93 sortiert nach: Relevanz ▾


1 2 3 4 5 ▶

Eingeschränkt nach:  ×

1  **Electrochemistry Data (NMC\_94wt\_0bar)**  
Martin Ebner  
2013  
Forschungsdaten ● **Online Ressource**  
[▼ Online-Zugriff](#) [▼ Details](#)

2  **Electrochemistry Data (NMC\_90wt\_0bar)**  
Martin Ebner  
2013  
Forschungsdaten ● **Online Ressource**  
[▼ Online-Zugriff](#) [▼ Details](#)

3  **Binary Data (NMC\_94wt\_300bar)**  
Martin Ebner  
2013  
Forschungsdaten ● **Online Ressource**  
[▼ Online-Zugriff](#) [▼ Details](#)

 [rss](#)  
 [Seite dem e-Shelf hinzufügen](#)

**Meine Ergebnisse einschränken:**

**Urheber**

- ▶ [Boss, Andreas](#) (1)
- ▶ [Corazza, Francesco](#) (1)
- ▶ [Hennel, Franciszek](#) (1)
- ▶ [Heule, Stefan](#) (1)
- ▶ [Im Obersteg, Dominique](#) (1)

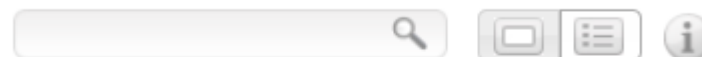
---

▶ [Mehr Optionen](#)

# Collection Viewer

## IE Access Rights enforced; Navigation on / off

NMC\_90wt\_0bar



1 - 5 of 5 Records

Title ▲

PID

1	 Binary Data (NMC_90wt_0bar)	IE105518	<a href="#">View</a>
2	 Electrochemistry Data (NMC_90wt_0bar)	IE105512	<a href="#">View</a>
3	 Labeled Particles (NMC_90wt_0bar)	IE105524	<a href="#">View</a>
4	 Particle Statistics (NMC_90wt_0bar)	IE105515	<a href="#">View</a>
5	 Raw Data (NMC_90wt_0bar)	IE105521	<a href="#">View</a>

# Collection Management Controlled by Ownership



- Software Registration (ETH transfer)
  - Closed Source Software
  - Open Source Software
  - Supplementary Material
  - X-ray Tomography
    - NMC\_90wt\_0bar
    - NMC\_92wt\_0bar
    - NMC\_94wt\_0bar
    - NMC\_96wt\_0bar
    - NMC\_90wt\_300bar
    - NMC\_90wt\_600bar
    - NMC\_92wt\_300bar
    - NMC\_92wt\_600bar
    - NMC\_94wt\_300bar
    - NMC\_94wt\_600bar
    - NMC\_96wt\_300bar
    - NMC\_96wt\_600bar
    - NMC\_90wt\_2000bar
    - NMC\_92wt\_2000bar
    - NMC\_94wt\_2000bar

Collection Name	Supplementary Material	Collection Description	Created by AssignCollectionByDCTask (DC source: 'Supplementary Material')
Collection ID	18575311	Creation Date	16.07.14
Modification Date	16.07.14	Publish	No
External System	-	External Id	-
Allow Navigation	No		

[View Collection](#)

Contents

Metadata

1 - 4 of 4 Records

	Title	PID	
1	Optimization of short broadband RF pulse...	IE154555	<a href="#">View</a> <a href="#">Info</a>
2	Elkin_Ecol_Apps_data_storage_2015	IE157604	<a href="#">View</a> <a href="#">Info</a>
3	Diss_ETH_21709_Kirf_Mathias_B_Digital_Ap...	IE157607	<a href="#">View</a> <a href="#">Info</a>
4	Data used in the paper "Recording large ...	IE157704	<a href="#">View</a> <a href="#">Info</a>

# 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](http://www.library.ethz.ch/Digital-Curation)

[data-archive@library.ethz.ch](mailto: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](mailto:matthias.toewe@library.ethz.ch)