"Same same but different"

research data and digital preservation at ETH Zurich

“Same same but different”

Research data and digital preservation at ETH Zurich

Rosetta Roadshow 2011
6th to 9th September
Dr. Matthias Töwe

ETH-Bibliothek
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

OUTLINE

1. ETH Zurich and ETH-Bibliothek
2. Background and objectives of our project
3. First results of a survey with researchers
4. Pilot projects
5. Expectations concerning the role of Rosetta
6. Strategic options for ETH-Bibliothek

ETH-Bibliothek
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

THE UNIVERSITY: ETH ZURICH

ETH Zurich

- **Swiss Federal Institute of Technology** Zurich
- Founded in 1855
- University for **technology and the natural sciences**
- More than **16‘000 students from 80 countries**, 3‘500 among them are doctoral candidates
- More than **400 professors** teach and conduct research in the areas of engineering, architecture, mathematics, natural sciences, system-oriented sciences, and management and social sciences

THE UNIVERSITY LIBRARY: ETH-BIBLIOTHEK

- **Central university library** for ETH Zurich with four **special libraries**
- Swiss centre for **technical and scientific information**
- **Special collections** including ETH Archives, Image Archive, Map Collection, ETH‘s Collection of Prints and Drawings and others
- **Hosting nationwide services** for university libraries, e.g.
  - NEBIS library network (http://www.nebis.ch)
  - Consortium of Swiss Academic Libraries (http://lib.consortium.ch)
  - Swiss electronic library e-lib.ch (http://www.e-lib.ch)
  - National digitization projects with currently about 3 million pages:
    - http://www.e-rara.ch (rare books)
    - http://retro.seals.ch (Swiss journals)

ETH-BIBLIOTHEK: THE PHYSICAL SPACE

ETH-Bibliothek – the physical library

- Physical holdings (by end of 2010): 7'617'000, including
  - 2'858'000 print volumes
  - 2'213'000 reports (print and microfiche)
  - 1'853'000 physical images
  - 216'000 manuscripts (ETH Archives)
  - 22'000 archival boxes (ETH Archives)
  - 5'330 subscribed print journals
- Usage in 2010:
  - 282'000 print documents on loan
  - 112'000 articles in document delivery

ETH-BIBLIOTHEK: THE DIGITAL SPACE

ETH-Bibliothek – the digital library

- Digital holdings (by end of 2010): 292'000, including
  - 158'000 electronic images
  - 63'000 e-books
  - 13'000 licensed e-journals
  - 24'000 full texts in inst. repository (ETH E-Collection)
- Usage in 2010:
  - 3'365'000 downloads of e-journal article full texts
  - 912'000 visits of e-book full texts
  - 2'025'000 downloads of full texts from ETH E-Collection
  - 527'000 visits to databases

ETH-Bibliothek
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

PROJECT BACKGROUND AND SCOPE

ETH-Bibliothek – the hybrid library

- ETH-Bibliothek's services are hybrid
- They will remain hybrid for a long time
- Weights shift continously
- Usage numbers underline growing dependence on digital resources
- Reliability, citability and persistence of digital resources do not yet match their importance

→ ***We need to do something…***

PROJECT BACKGROUND AND SCOPE

... and we could do something for others, as well:

- **ETH Zurich guideline on integrity of research**
  - Project managers must ensure that („primary“) data is kept for as long as is appropriate for the discipline
  - No common tool is available to support this
- **Irreplaceable data**
  - Unique observational data in long continuous timelines
  - Other data which is to be used for comparative research
- **Published and / or referenced data**
- **Administrative records from ETH Archives**
- **Library materials (born digital theses, digitization masters)**

ETH-Bibliothek
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

ROLES IN ETH ZURICH'S INFRASTRUCTURE

Overall aim: local services managed centrally

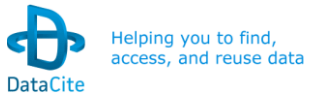
ETH-Bibliothek

- **Information, library and collection management** for ETH Zurich
- **Library IT-Services** dedicated to library's applications
- **Mandate for Digital Curation** from the university board
- **Focus on research data**, but responsibility for all data

ETH Zurich IT Services

- **IT infrastructure** management
- **IT service** provision including virtual environment for Rosetta
- **Storage** management
  → New storage infrastructure from 2011 onward aiming at:
  Robustness, scalability, transparent costing, efficient hierarchical storage management, rule-based storage according to pre-defined classes of data

PURPOSES

- **Re-use of data**
  - Accountability, re-assessment
  - Citability of data referred to in publications
    → DOI-registration already operational 

  - Re-use could range from granting access to known colleagues to fully *Open Data*

- **Open Data?**
  - Very sensitive issue to many scientists
  - Not necessarily related to long-term preservation

SURVEY OF ALL RESEARCH GROUPS

What do our potential customers need and want?

- Online survey with **all** professors
- Questions on produced and used **data**, current **data handling** and possible offers from the library.
- **Personal reminders** and talks
- Several departments with **answers from about 80%** of addressees
- **Extremely heterogeneous** degrees of awareness, understanding and willingness
- Depending on disciplines and persons

ETH-Bibliothek
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

STATUS QUO OF DATA HANDLING

Results from the survey and from interviews on current solutions:

- **„None“:** Data on a file system and/or on offline media – only group leader can retrieve anything manually
- **Managed on- and offline storage** including conversion to open formats (e.g. doc to rtf or txt) and periodic migration to new media
- **Supported applications** on group level
  Capture data when produced, support handling, analysis and visualisation, but not long term preservation in a narrower sense

→ ***There is awareness that data needs to be taken care of***

→ ***Preservation must not be mixed up with initiatives for Open Data***

ETH-Bibliothek
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

RESEARCHERS NEEDS

Researchers...

- want to keep **full control** at least of who accesses their data - even though they might theoretically be in favour of Open Data
- need to **re-arrange and select** data prior to ingest, add **documentation** and **legal documents**
- need to **edit metadata and add data** to ongoing series, e.g. annually
- are **interested in support for preservation and quality control** (checklists, feedback on metadata...)
- need to keep certain data for **limited periods** (e.g. 10 to 12 years)
- see **archiving needs** often **related to** data and materials used for **publications** and want to **persistently reference** them
- make mixed statements concerning a **data policy** for ETH
- want **no additional workload**

ETH-Bibliothek
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

POSITION WITHIN THE LIFECYCLE

- **Previous assumption:**
  - *„Researchers won‘t want us to interfere with their research and data management: Let‘s support them with advice and checklists during data production and otherwise only take over data at the end of the pipe.“*
- **Statements from pilot partners:**
  - *„We need to be able to do some degree of data management before/while submitting material to the long-term archive and don‘t want to use an additional tool for that.“*

→ ***In some cases we might be closely in touch with the research process and its data management.***

→ ***A good thing for the library, but demanding with respect to staffing and software.***

ETH-Bibliothek
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

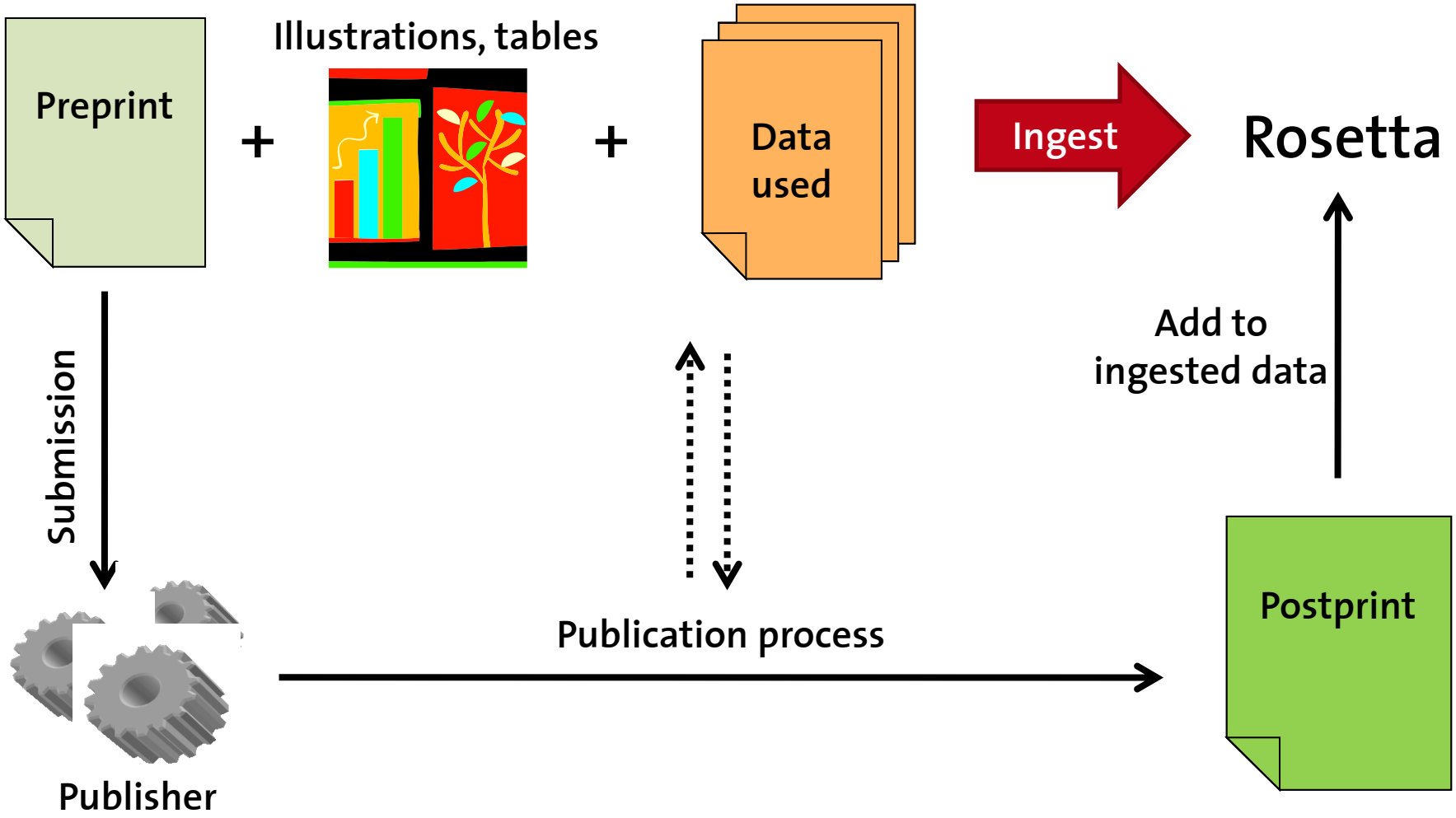
CURRENT STEPS

- **Pilot phase with four to five pilot partners**
- **Different requirements**
- **Proposed workflows**
  - Manual data management around a publication ready manuscript
  - (Group) datasets as raw material for following publications
  - Automatic import from an existing data management application

→ ***Next step: definition of use cases with pilot partners and discussion with Ex Libris***

ETH-Bibliothek
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

EXAMPLE WORKFLOW



PARALLEL PROJECTS

Pilot with ETH Archives

- ETH Archives operated by ETH-Bibliothek
- Legal obligation to archive digital administrative records
- Heterogeneous data producers

Pilot with ETH E-Collection

- Institutional repository
- Materials belonging to ETH‘s scientific heritage

→ Get to know the system with known data

POSSIBLE LATER PROJECTS

Materials from mass digitization

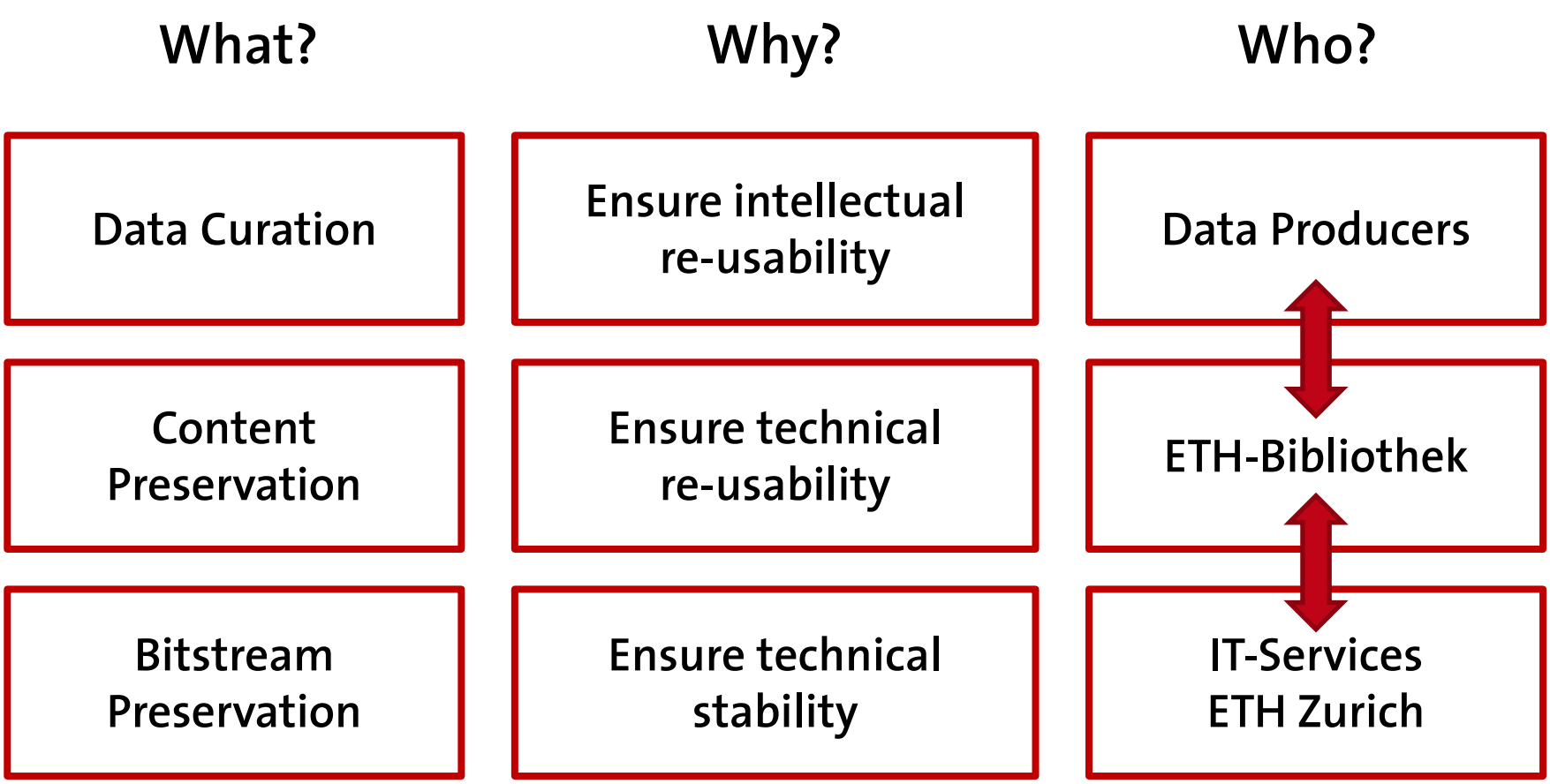
- Image master files
- Metadata files which can be recreated only with massive expenditure of time and money

→ *Digital preservation to be **integrated in existing workflows** on running applications*

→ *Further analysis to evaluate the **potential of Rosetta to take over functions** of existing platforms*

ETH-Bibliothek
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

FUNCTIONAL LEVELS



Adapted after Jens Ludwig, Wissgrid

CHARACTERISTICS OF MATERIALS

Research data	Library objects
• Only partly known context	• Usually known context
• Numerous formats	• Few formats
• Often no standard formats	• Mostly standard formats
• Third party production, can only be influenced over time	• Production can be influenced rapidly
• Few formalized metadata	• Formalized metadata
• Rights often not clear (ownership, use, publication)	• Copyright situation usually clear
• Public access not necessarily wanted	• Usually public access is desired
• Reservations with some producers	• In-house: a certain degree of trust can be expected – or decreed

ETH-Bibliothek
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

RISKS PRIOR TO OBSOLESCENCE

Serious risks need to be addressed during data production

- Unclear or frequently changing responsibilities

  → ***Loss of meta-information***

- Missing or incomplete documentation

  → ***Loss of contextual information***

- Haphazard directory and file structures

  → ***Versioning issues, uncontrolled redundancies***

→ *Occurrence of these and further risks* ***can make long-term preservation of any data questionable***

→ ***Research is particularly prone to these risks*** *due to its dynamic development and the high mobility of staff*

ETH-Bibliothek
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

OBSERVATION

- Meaningful re-use of research data will rely heavily on **contextual information and structural relations**
- **Exhaustive documentation** is required
- There is a **need to appraise, select and re-arrange** objects prior to ingest and later in time

→ ***Treatment and ingest of research data might have more in common with challenges in administrative archives than with those in typical library collections***

→ ***Trying to keep this in mind for synergies in future development***

ETH-Bibliothek
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

DIFFERENCES BETWEEN DATA TYPES?

What?	Research data	Library objects
Data Curation	Comprehensive documentation by producers required	Full control of metadata and context
Content Preservation	More and less common formats	Mainly standard formats
	Same preservation procedures apply	
Bitstream Preservation	„Any object is just bits“	

CONSEQUENCES FOR THE USE OF ROSETTA

- **Bitstream preservation relies on ETH Zurich‘s IT services‘ infrastructure**
- **Rosetta is the backbone of content preservation**
- **Rosetta is well prepared to support data curation**
  - „Classical“ metadata management
  - Support of supplementary documentation
  - Configuration of flexible workflows
  - Potential for data management prior to archiving?
  - How and where to deal with limited retention periods?

ETH-Bibliothek
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

LIMITATIONS?

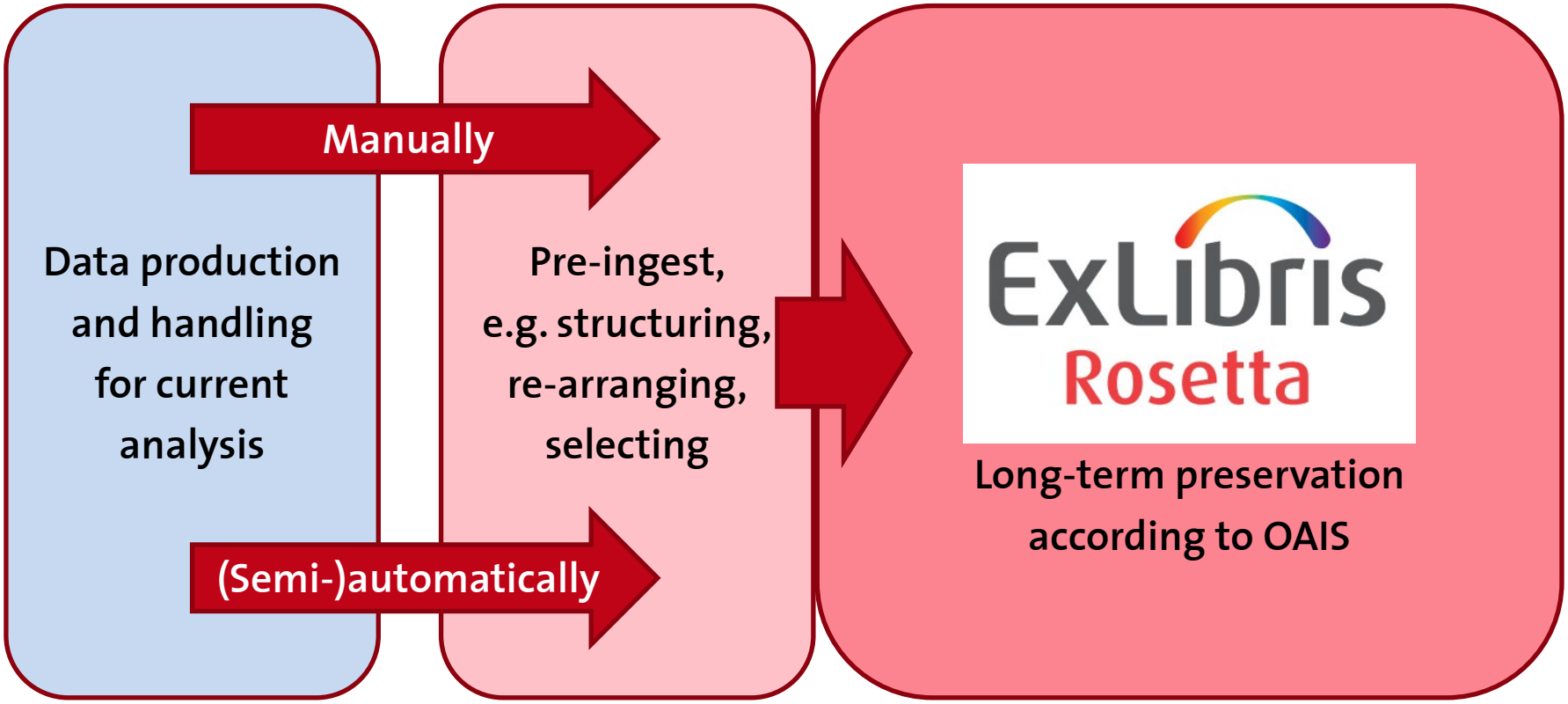
- Rosetta supports important functions with a **clear focus on long-term preservation** according to OAIS
- Rosetta can only support preservation when **adequate staffing and an active preservation management** are in place
- An **active international community** is needed to collect, manage and share information and knowledge, e.g. on formats
- **Flexibility** in data management partly **contradicts** the requirements of a **stable preservation environment**:

→ *Where should be the **interface** between data management and preservation?*

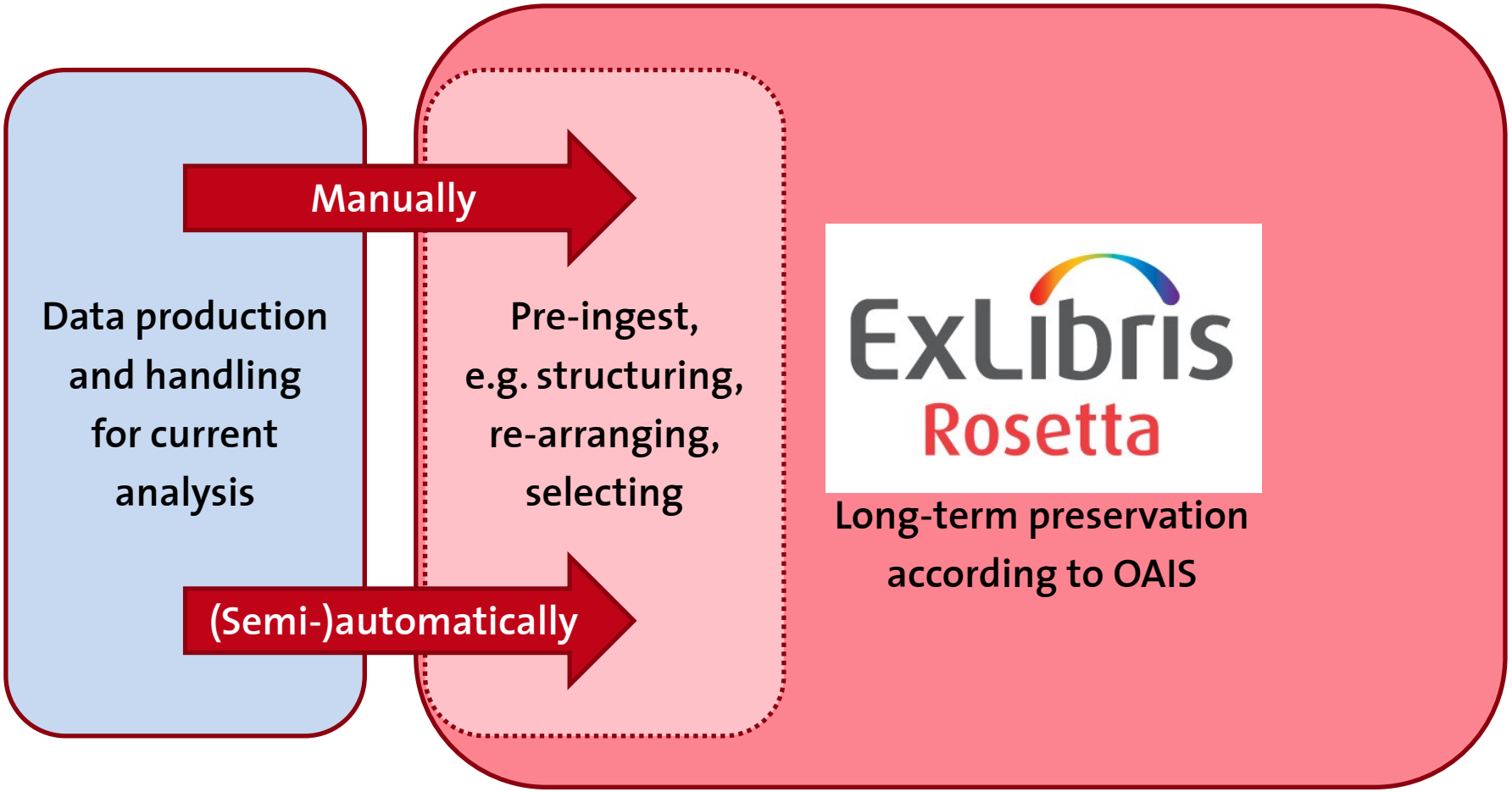
→ *What are the **technical limitations**?*

ETH-Bibliothek
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

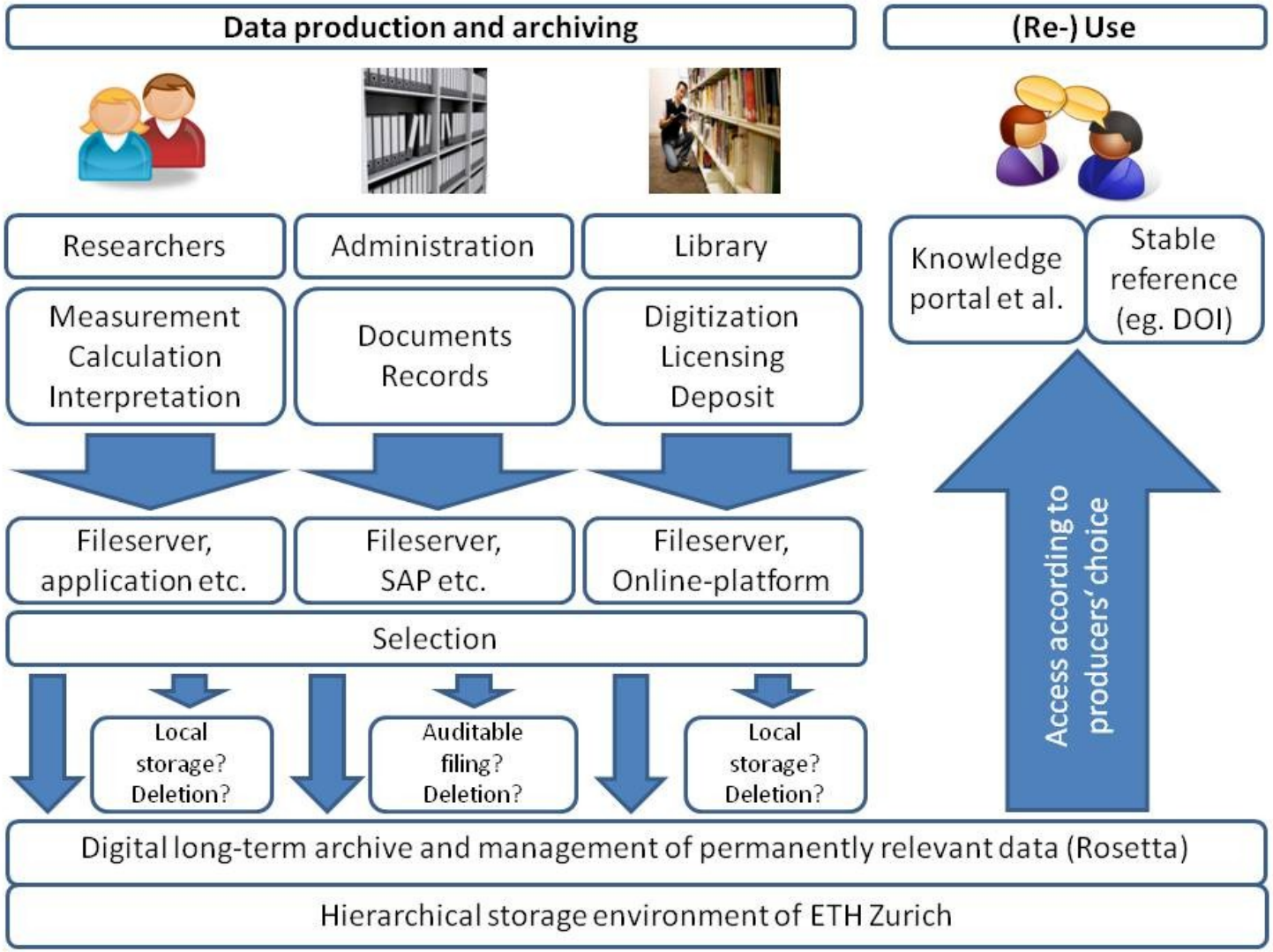
POSSIBLE WORKFLOWS AND ROSETTA



POSSIBLE EXTENSION OF ROSETTA'S SCOPE?



VISION: ROSETTA AS A COMMON BASIS



STRATEGIC QUESTIONS FOR ETH-BIBLIOTHEK

- **What kind of service can / should we offer?**
  - Limit ourselves to content preservation in a narrow sense?

  Or – in the longer run and with appropriate partners:

  - Open up to support data management or…
  - … even support other functions within the research process? …
  - … e.g. virtual research environments?

→ ***Tempting,*** *but even more* ***challenging***

→ ***Heterogeneous environment*** *prevents common solutions*

→ *Possible only with (very)* ***strong committment of ETH's board***

→ ***To be discussed*** *based on experiences from the current project*

ETH-Bibliothek
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

WORK SO FAR

- DOI-registration by ETH Zurich as member of DataCite
- **Full survey** of research groups (Profs.) at ETH Zurich and accompanying informal interviews
- Identification of **pilot partners**
- Workshops with 4 research partners on their requirements
- Work on a **manual workflow** for admininstrative records for **ETH archives**
- Check and update inventory of data hosted by the library
- Work on specification of **submission application** for library materials (institutional repository)

NEXT STEPS

- Implement **manual workflows** for research data and ETH archives
- **Identify further requirements** to be addressed in development phase until the end of 2012
- Specify and develop **submission application** for library materials
- Develop and implement **submission application** for import of research data **out of existing data management solution?**

If successful:

- **Extend coverage** to more groups
- **Convince the university's board** to grant **ongoing funding** as part of their risk management

ETH-Bibliothek
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

THANK YOU VERY MUCH!

Questions?

Dr. Matthias Töwe
Head Digital Curation
ETH-Bibliothek
Rämistrasse 101
8092 Zürich
Switzerland
+41 (0)44 632 60 32
matthias.toewe@library.ethz.ch
http://www.library.ethz.ch