“Forever is composed of Nows"

Long-term preservation of research data in an academic library

Author(s):
Töwe, Matthias

Publication date:
2012

Permanent link:
https://doi.org/10.3929/ethz-a-007362251

“Forever is composed of Nows”:

Long-term preservation of research data in an academic library

UKSG 2012
Glasgow, 26th/27th March 2012
Dr. Matthias Töwe
ETH Zurich, ETH-Bibliothek

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

OUTLINE

1. Background: issues and objectives
2. Current project
3. Roles
4. Vision
5. Limitations
6. «Nows» and caveats

BACKGROUND (I)

Challenges

- **Research** process as a whole **relies on digital data**
- Data can only be used in a **defined technical environment**, which usually remains stable for only a few years
- **Good scientific practice** requires retention of data in usable form
- **Funding organisations** require data management plans (e.g. NSF, DFG)

BACKGROUND (II)

Challenges

- **Re-use of data** becomes increasingly important and should be facilitated
- Data which cannot easily be reproduced and has **permanent relevance** must remain available
- **Published or referenced** supplementary material must be citable and remain available
- Researchers want to **retain control** of their data

MAJOR RISKS

- **Data loss**

  → Data cannot be found

- **Loss of readability**

  → Data cannot be rendered due to technical reasons (most often obsolescence of one required component such as application, operating system, hardware)

- **Loss of interpretability**

  → Data cannot be interpreted and used in a scientifically correct manner due to a lack of semantic information

DATA LOSS

Data Loss

→ Data cannot be found because...

- Their location of storage is not known
- File or folder structures were changed without documentation
- Intransparent redundancies and versions exist
- Persons originally responsible cannot be conctacted
- Offline-media are stored in unknown locations
- Offline-media were damaged by deterioration
- Reading devices für offline-media are no longer available

→ ***Recovery*** *«ex post» might even be possible, but* ***effort/cost will only be justified in exceptional cases***

LOSS OF READABILITY

Loss of readability

→ Data cannot be rendered because…

- **File formats** are not recognized by current software or are not rendered correctly
- **Software** required for rendering or even editing data is no longer available
- Available older software cannot be run on current **operating systems and/or hardware**

→ ***Recovery*** *«ex post» might even be possible, but* ***effort/cost will only be justified in exceptional cases***

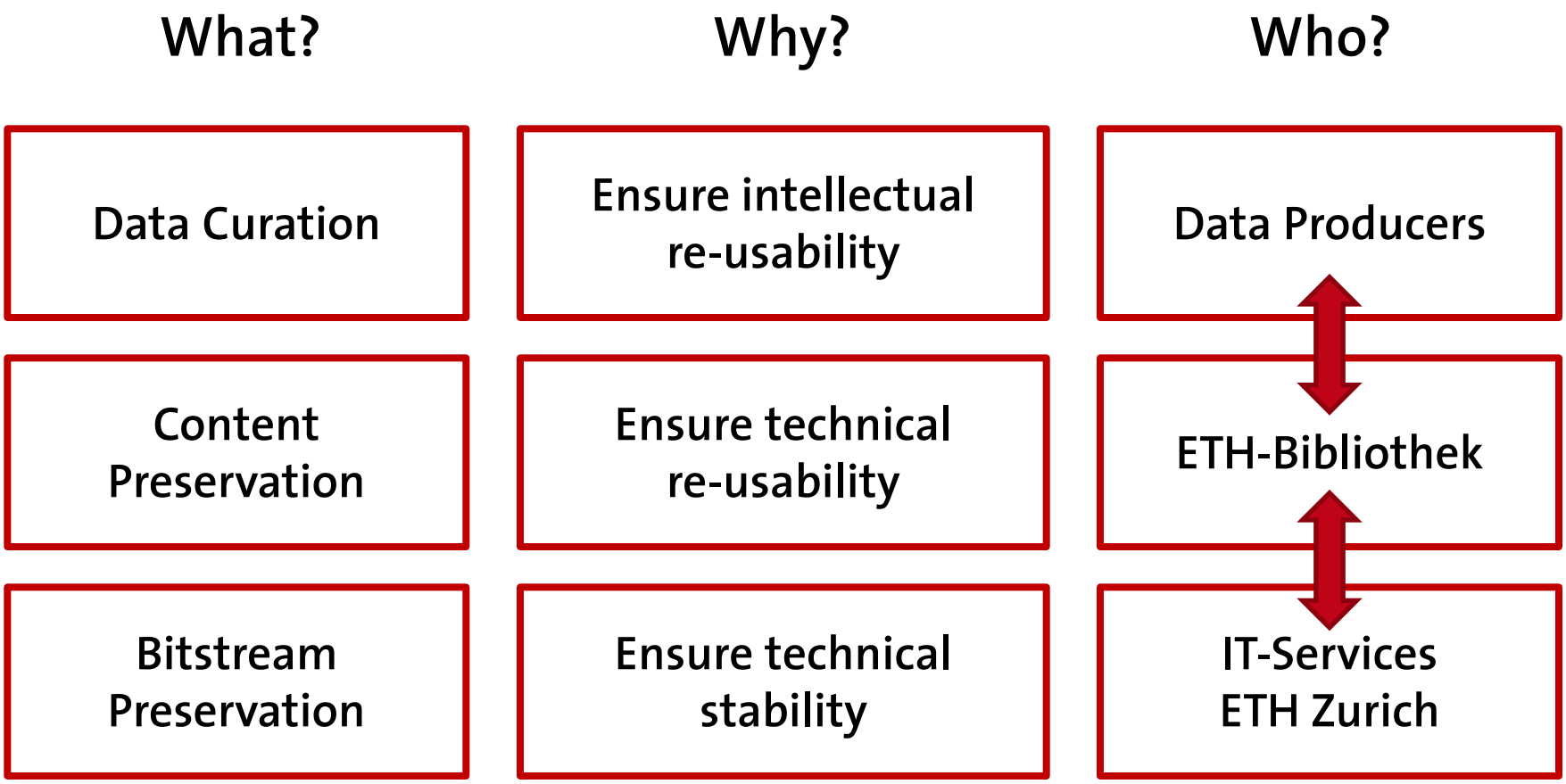
LOSS OF INTERPRETABILITY

Loss of interpretability

→ Data cannot be interpreted and used in a scientifically correct way because **semantic information is missing**, e.g. about...

- **Sample** taking and preparation
- **Methods of measurement** or data collection
- Known **errors and corrections**
- **Level of data processing**
- **Methods of analysis** and algorithms used
- ...

WHAT WE MEAN BY CURATION



Adapted after Jens Ludwig, Wissgrid

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“	

MISPERCEPTIONS

(**Many**) people including **potential partners** (IT, research)

- Tend to mix up long-term **storage** (bitstream preservation) and **long-term preservation** (keeping data usable)
- **Take preservation for granted**, once reliable storage is in place
- **See the need to change and improve current practice in data management** with the option of long-term preservation

«OUR» ROLE AND «THEIRS»

- **Can we actually «raise awareness»** with researchers?
- **Is it really useful to bother researchers** with this?
- Researchers should be provided with a convincing **added value service which makes their lives easier**
- There are **researchers with a high level of awareness and concern**
- **Best start with those who actually want a change**

COULDN'T RESEARCHERS DO IT THEMSELVES?

- **Data management and digital curation handled by researchers themselves:**
  - **Possible** in principle
  - **Time consuming**
  - Supportive of research productivity
  - **Not productive research in itself**

WHY DOES ETH-BIBLIOTHEK BOTHER?

- **Infrastructure services such as ETH-Bibliothek and IT services**
  - **Support** the research process
  - Can offer services to **ease workload of routine tasks** for researchers
  - Rely on scientists to define their requirements
  - **Rely on researchers to document their data** according to community needs
  - **Exploit synergies** in order to make data storage and curation more efficient within ETH Zurich as a whole

WHY THE LIBRARIES?

- Reputation of scientific libraries as **long-lived/permanent institutions**
- **The concept of organising and managing information** is seen as a task, where librarians might contribute
- Building on former (obviously **positive) track record, there should be a basis of trust**

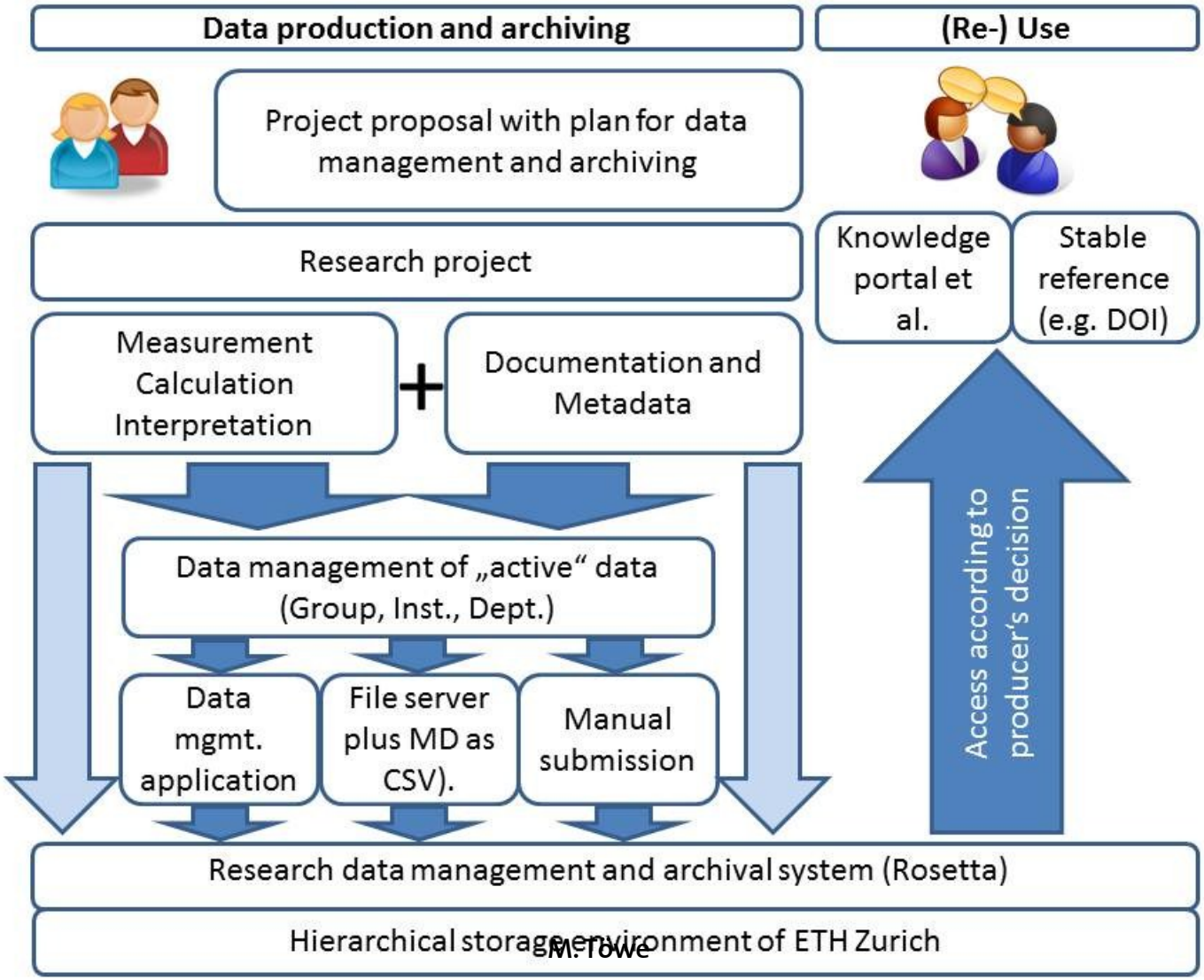
→ ***Survey at ETH Zurich confirmed that researchers see a role for ETH-Bibliothek here***

NEW TASKS FOR LIBRARIES

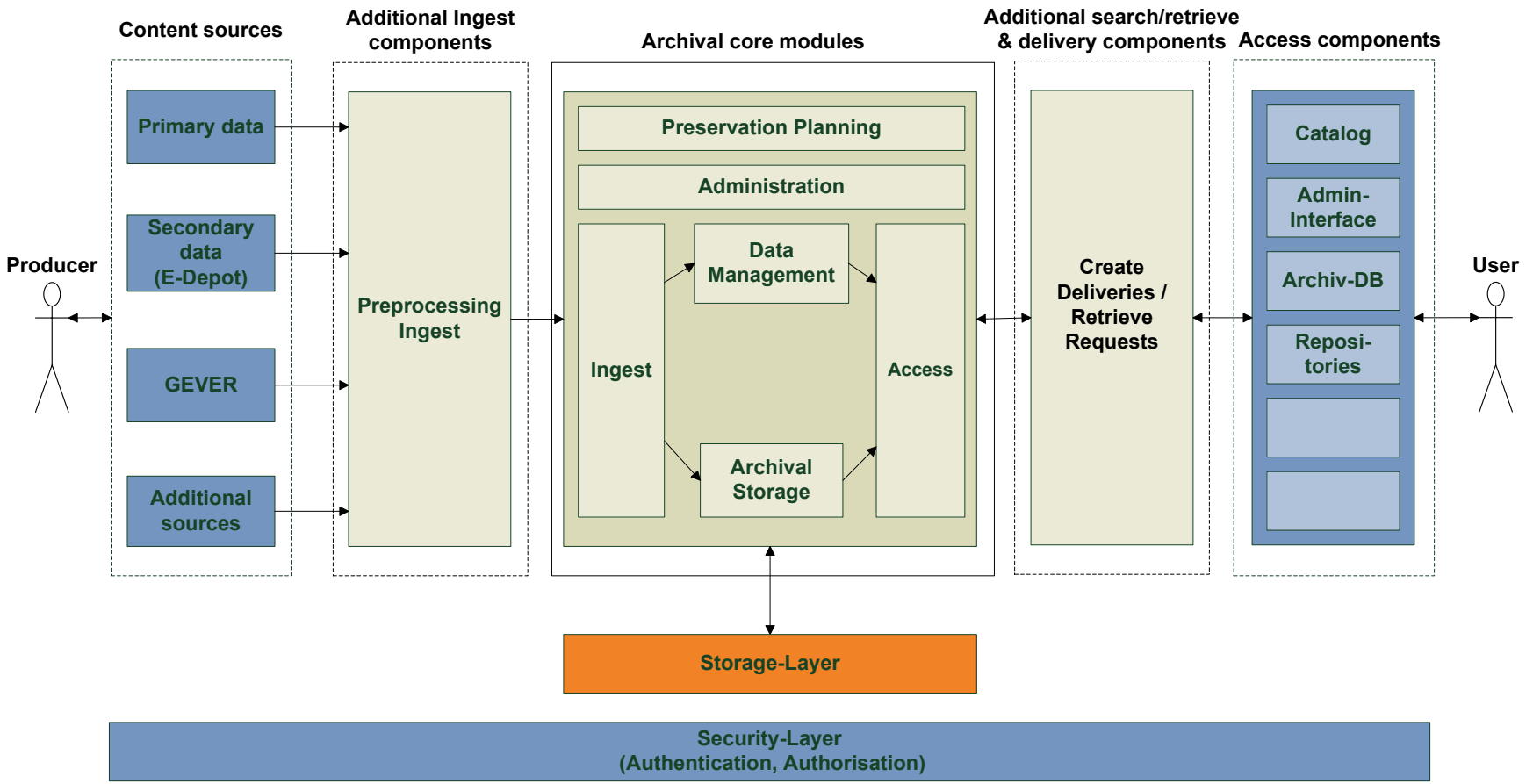
- **We can be service providers** – if we have a service to offer
- **We take on a new role:**
  - **In addition to delivering information** to researchers…
  - …we now offer **services around their own data…**
  - **…which we often cannot even make publicly accessible.**
- New tasks call for a **new professional profile** («data librarian?»)…
- …and for **new institutional cooperations** within a university

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

VISION - USER'S VIEW

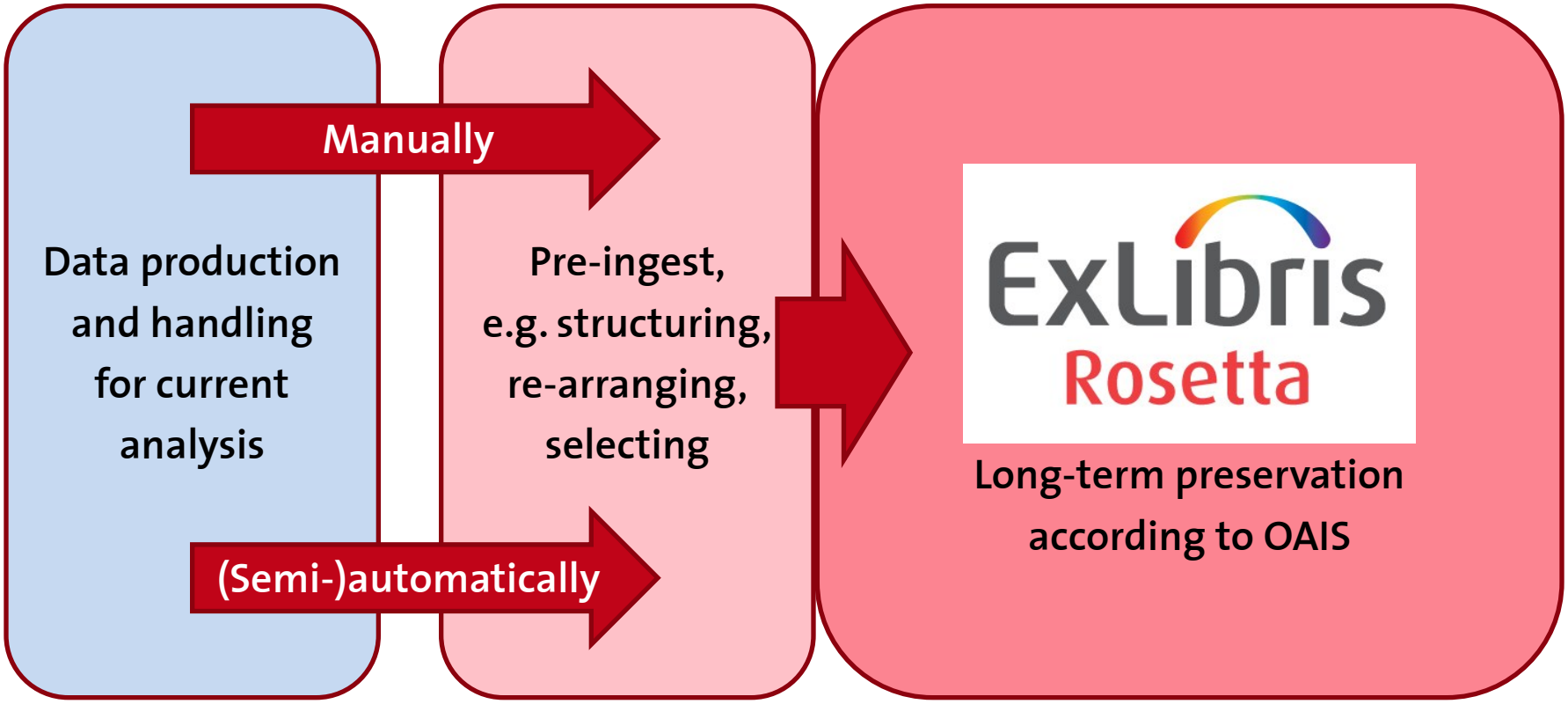


VISION – SYSTEM'S VIEW

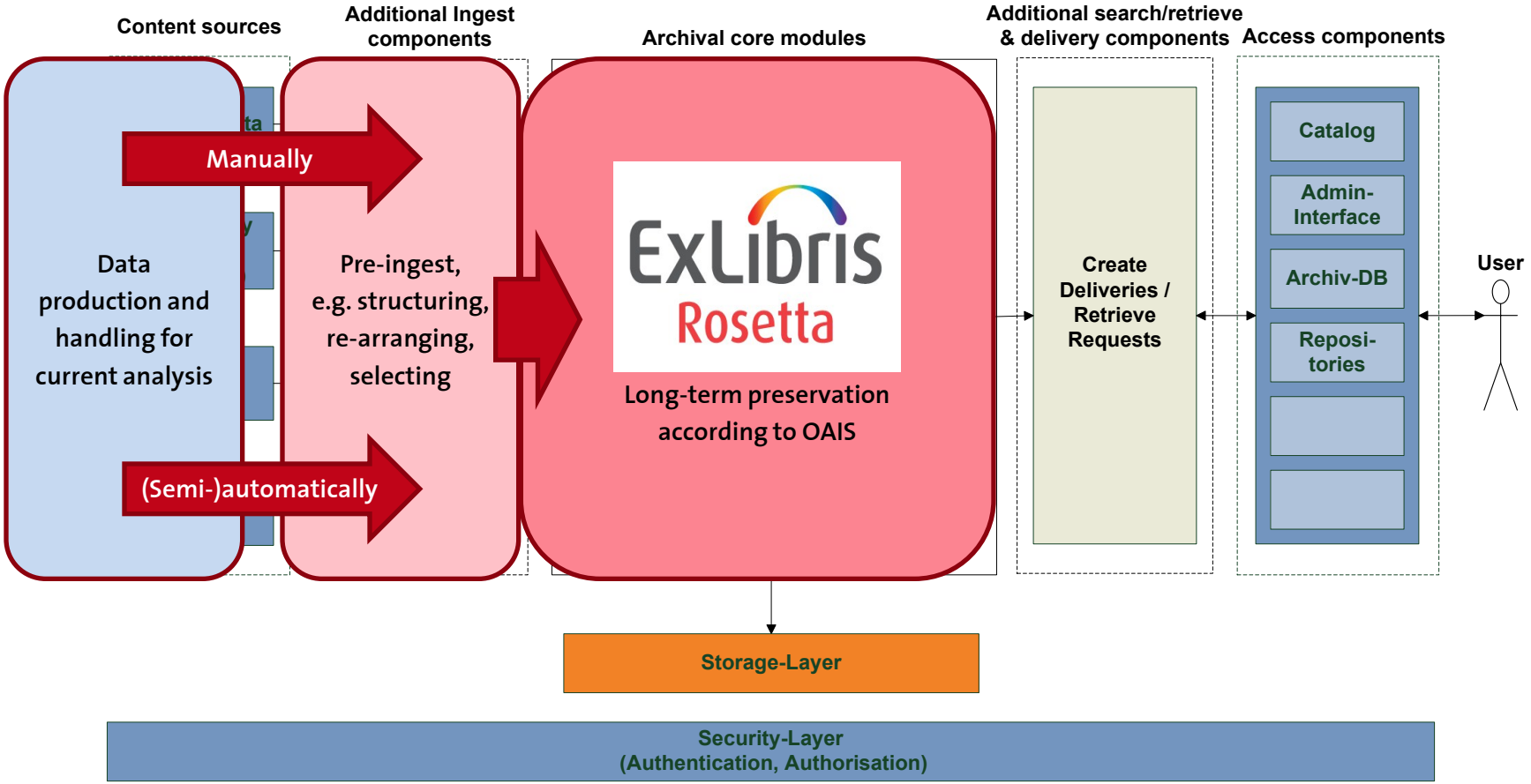


Abschlussbericht zur zweiten Phase „Pilot Langzeitarchivierung", S. 23f; Aliesch, P. et al., 2007: Projekt „Pilot Langzeitarchivierung". Intern.

ROSETTA


Manually
Data production and handling for current analysis
Pre-ingest, e.g. structuring, re-arranging, selecting
(Semi-)automatically
ExLibris
Rosetta
Long-term preservation according to OAIS

VISION



Abschlussbericht zur zweiten Phase „Pilot Langzeitarchivierung“, S. 23f; Aliesch, P. et al., 2007: Projekt „Pilot Langzeitarchivierung“. Intern.

LIMITATIONS

There are **general limits** to what we can do

- We need to **make decisions now**…
- …which **influence if and how data can be used** in future.
- **We do not know…**
  - **Who** will use data
  - **When data** will be used
  - **For which purpose** data will be used
- **«Someone» needs to commit now to paying for «eternity»**

ONLY RESEARCH DATA?

- These limitations are **not specific of research data…**
- …but they are **more pronounced** in research:
  - **High mobility** of staff
    → **fluctuation** in responsibilities
  - **Dynamic development** of methods
  - **Data management not always considered as a priority**
  - **Multitude of formats**
  - **Heterogeneity** between disciplines in methods and practices

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

THE TROUBLE WITH «NOWS»

Long-term preservation is no one-off activity

- **Each generation has to act** according to its best knowledge
- Usually, the aim is to **hand over usable data to the next generation** of curators
- **The overall quality of the preservation chain is governed by the preservation step with the lowest quality**
- **It will be difficult to later execute an action** which was missing in the chain

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

WHAT CAN BE DONE «NOW»

Examples for the «nows» in research data

- Only now we can **communicate with data producers**
- **Find out what their needs are**
- **Define the required services**
- **Make producers document their data**
- **Discuss alternative formats** where necessary

CAVEATS

- Digital curation cannot «improve» data retroactively: **«garbage in – garbage out»**
- Therefore **researchers need to actively contribute** (e.g. documentation)
- **Who decides** about data when the producer is no longer available?
- **Data *can* be made publicly available**, but this must not be a prerequisite for its preservation

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

MORE CAVEATS

- **Written agreement** between data producer and data archive on formats, procedures and access rights
- Management of **active data not treated in current project…**
- …but **we need to provide comfortable routes to bring research data into the archive**
- **There is no absolute safety** against willful attacks: On the server level, manipulations are possible, but they won't go unnoticed

EVEN MORE CAVEATS

- «**The art of communicating with the future**»:
  - We *now* **try to minimize risks** with reasonable effort in order to avoid their occurrence in future
  - Together with producers **we can only make educated guesses** at who might want to use data for what kind of purpose
- **No «rocket science», but an ongoing task with complex dependencies and a lot of work behind**

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