Rosetta and Data Curation in the context of research: first insights

Haifa, 11th September 2011

Andreas Kirstein

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

OVERVIEW

1. Introduction
2. Survey „Handling research data“ : first results
3. Status quo of data handling
4. Needs of researchers at ETH Zurich
5. Workflows and integration with Rosetta
6. Formats and other issues
7. Current and future work

ETH ZURICH – THE UNIVERSITY

ETH Zurich

- **Swiss Federal Institute of Technology** Zurich
- Founded in 1855
- One of the leading international universities 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

INFRASTRUCTURE

ETH Zurich infrastructure

- **ETH-Bibliothek**
  - Information, library and collection management for ETH Zurich
  - Main library and special libraries in departments
  - Special collections including ETH Archives
  - Library IT-Services dedicated to ETH-Bibliothek‘s applications

- **ETH Zurich IT Services**
  - IT infrastructure management
  - IT service provider
  - Storage management

→ Aim: central services + local services managed centrally

PROJECT BACKGROUND AND SCOPE

- **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)**

SURVEY „HANDLING RESEARCH DATA“ : FIRST RESULTS

Online survey with all professorships or research groups per department

Department	abbr.	Number of professorships or research groups per department	Rate of return (%)
Environmental Sciences	D-UWIS	33	63
Agricultural and Food Sciences	D-AGRL	19	78
Humanities, Social and Political Sciences	D-GESS	23	83
Earth Sciences	D-ERDW	17	94
Chemistry and Applied Biosciences	D-CHAB	50	65
Civil, Environmental and Geomatic Engineering	D-BAUG	40	100
Biosystems Science and Engineering	D-BSSE	10	80
Mathematics	D-MATH	15	87
Materials Science	D-MATL	12	100
Management, Technology and Economics	D-MTEC	20	89
Mechanical and Process Engineering	D-MAVT	*in progress*	
Computer Science	D-INFK	*in progress*	
Architecture	D-ARCH	*in progress*	
Biology	D-BIOL	*in progress*	
Information Technology and Electrical Engineering	D-ITET	*in progress*	
Physics	D-PHYS	*in progress*	

Questionnaire consists two parts

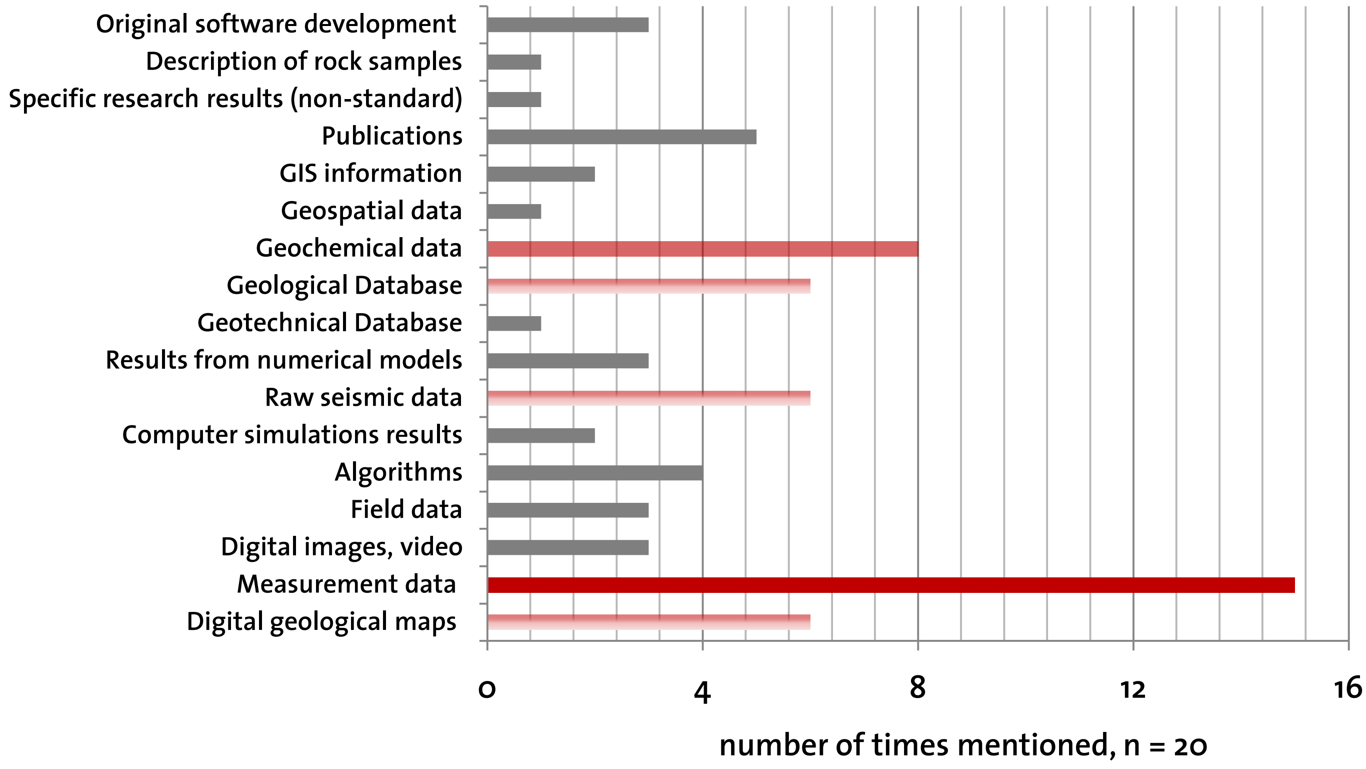
First part: how research data are handled generally, questions are based on a survey carried out by the working group "Research data" of the Leibniz community

Second part: specific questions about research data, questions based on the paper „Conducting a Data Interview“ from Witt & Carlson, Purdue University Libraries, 2010

SURVEY „HANDLING RESEARCH DATA“ : FIRST RESULTS

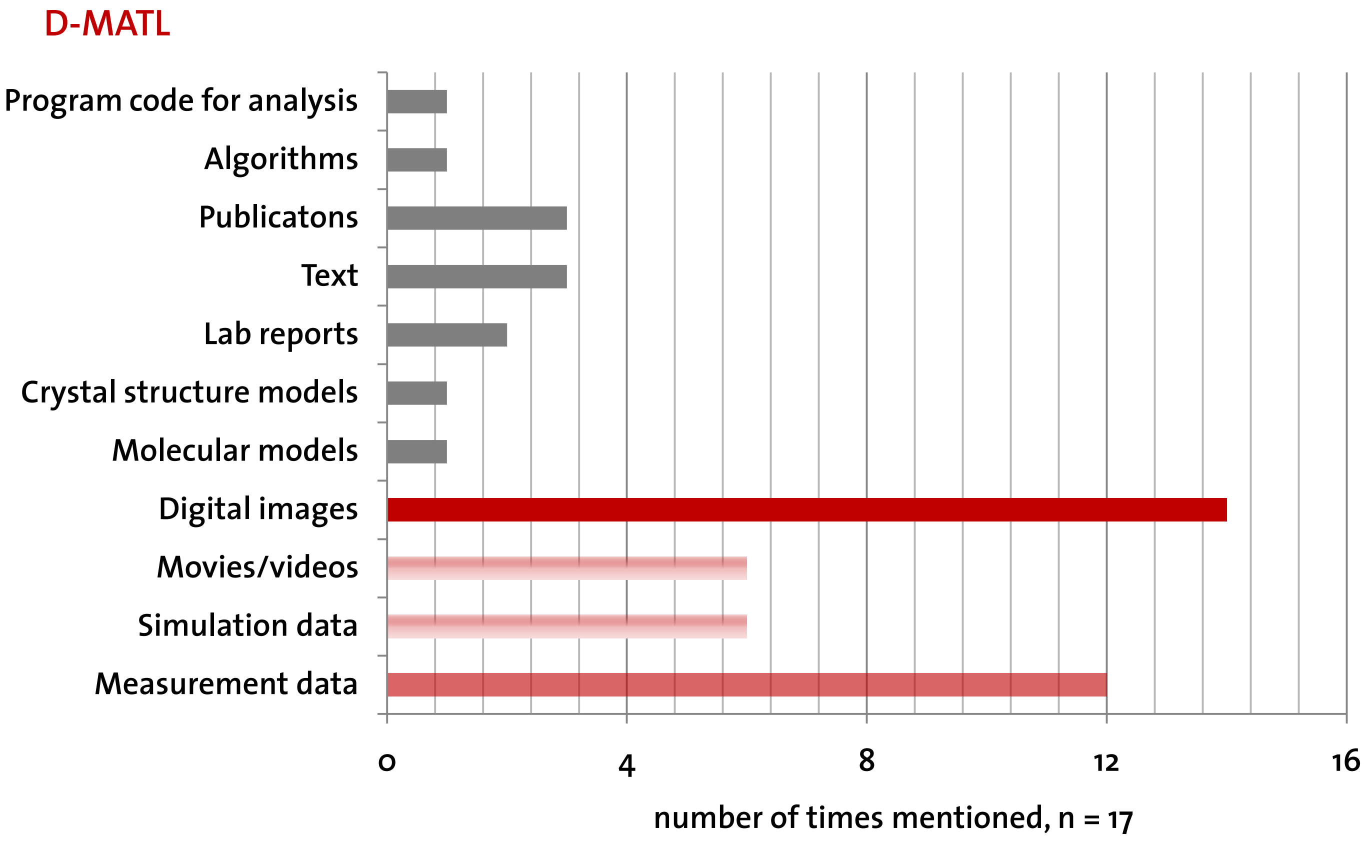
With which types of data do researchers in your discipline work?

D-ERDW



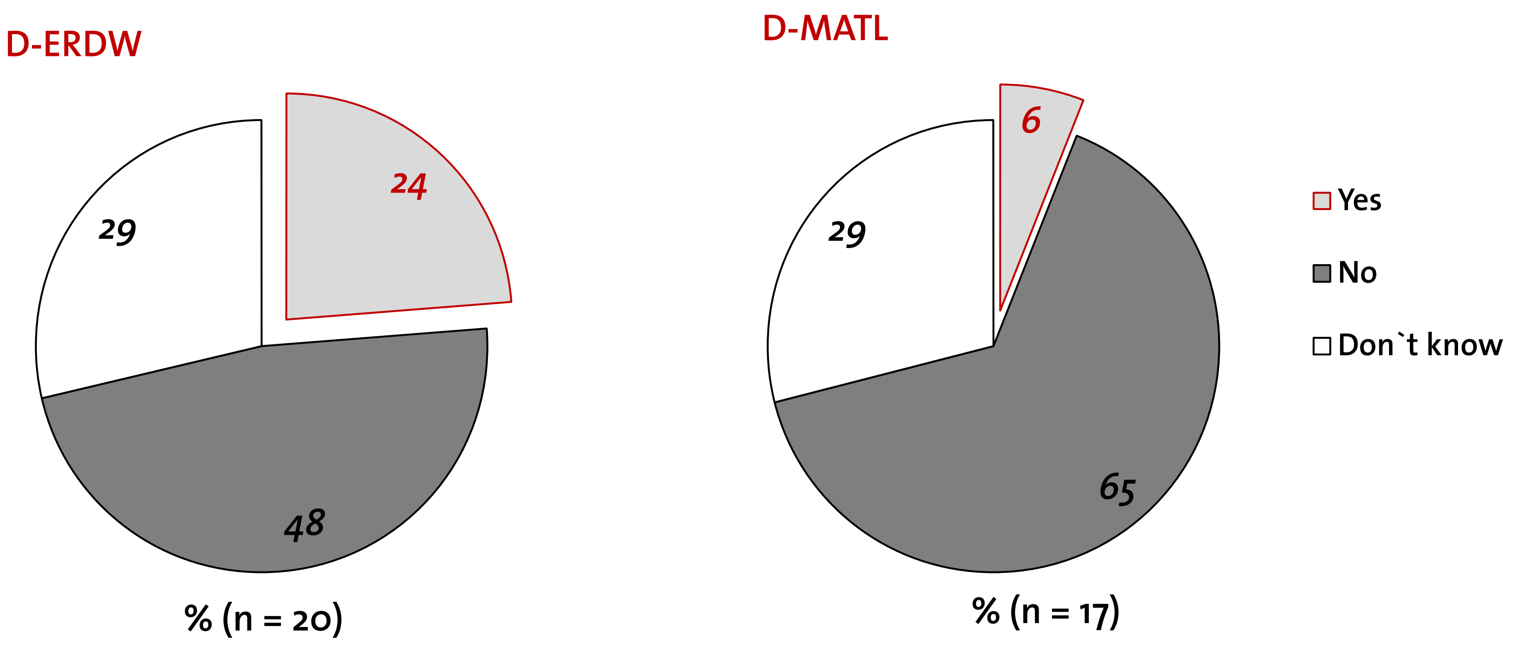
SURVEY „HANDLING RESEARCH DATA“ : FIRST RESULTS

With which types of data do researchers in your discipline work?



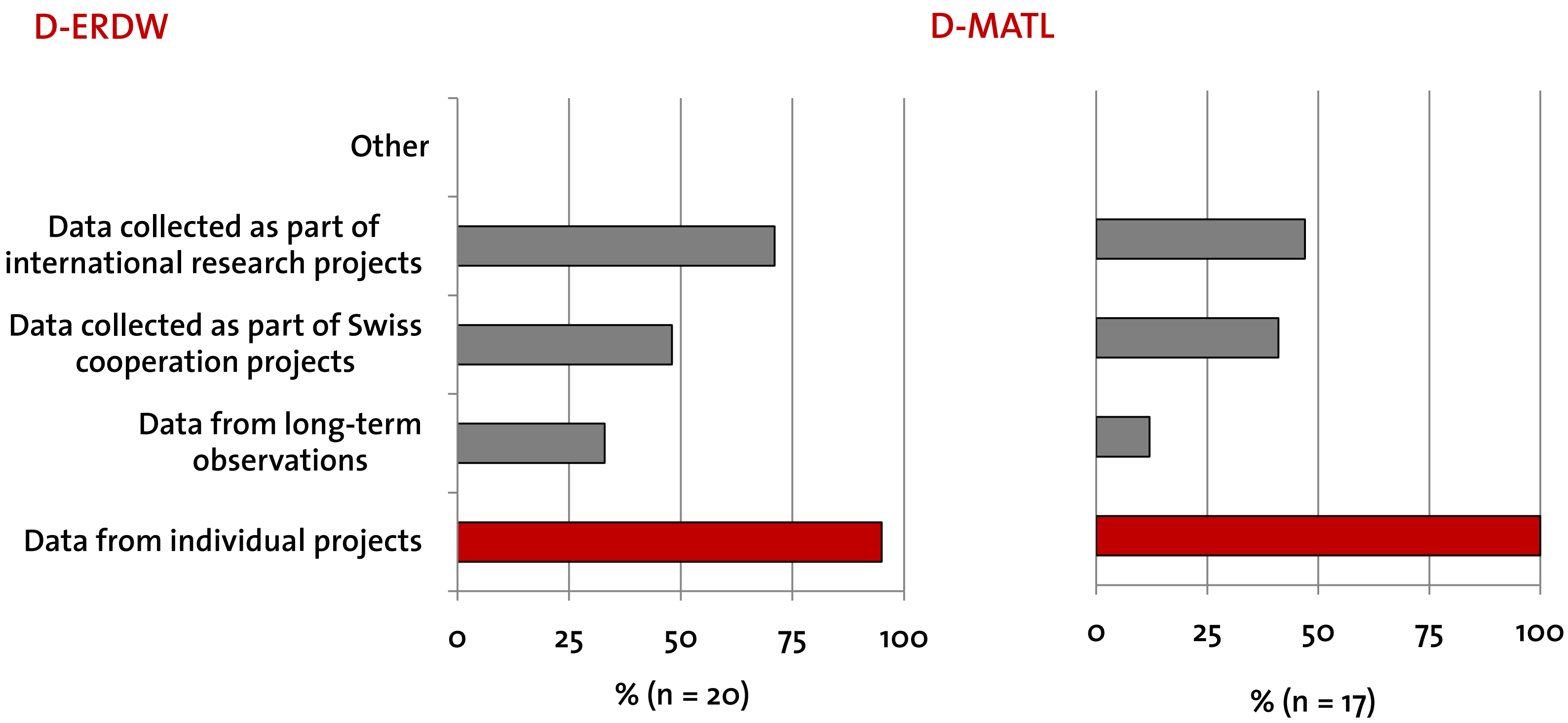
SURVEY „HANDLING RESEARCH DATA“ : FIRST RESULTS

In your discipline, are there metadata standards for describing research data?



SURVEY „HANDLING RESEARCH DATA“ : FIRST RESULTS

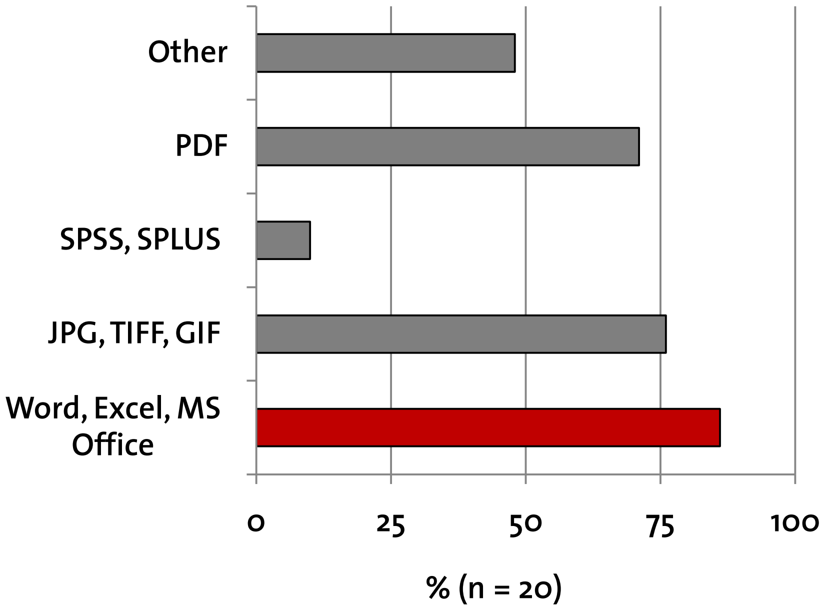
What is the main type of research data worked on by your research group?



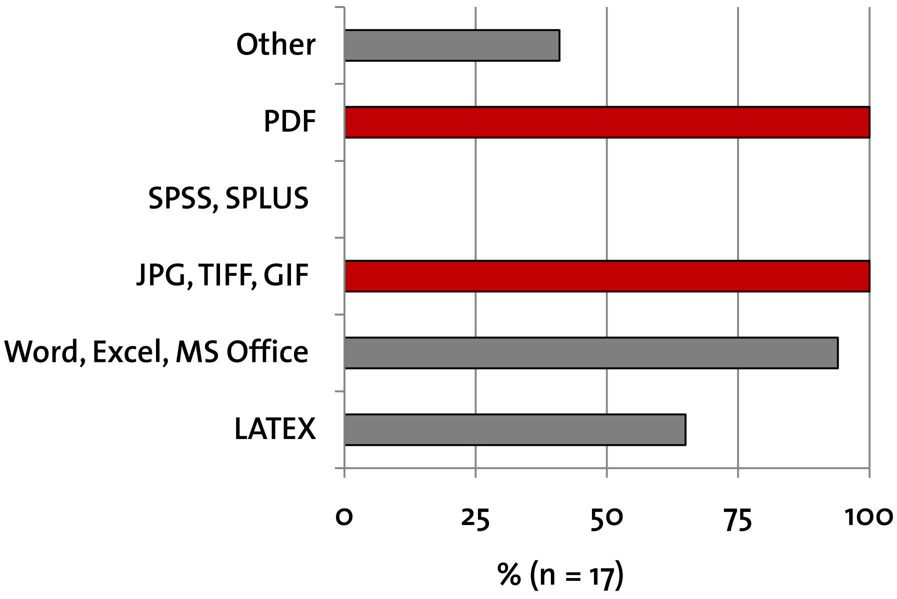
SURVEY „HANDLING RESEARCH DATA“ : FIRST RESULTS

Which formats do you and your research group work with?

D-ERDW



D-MATL

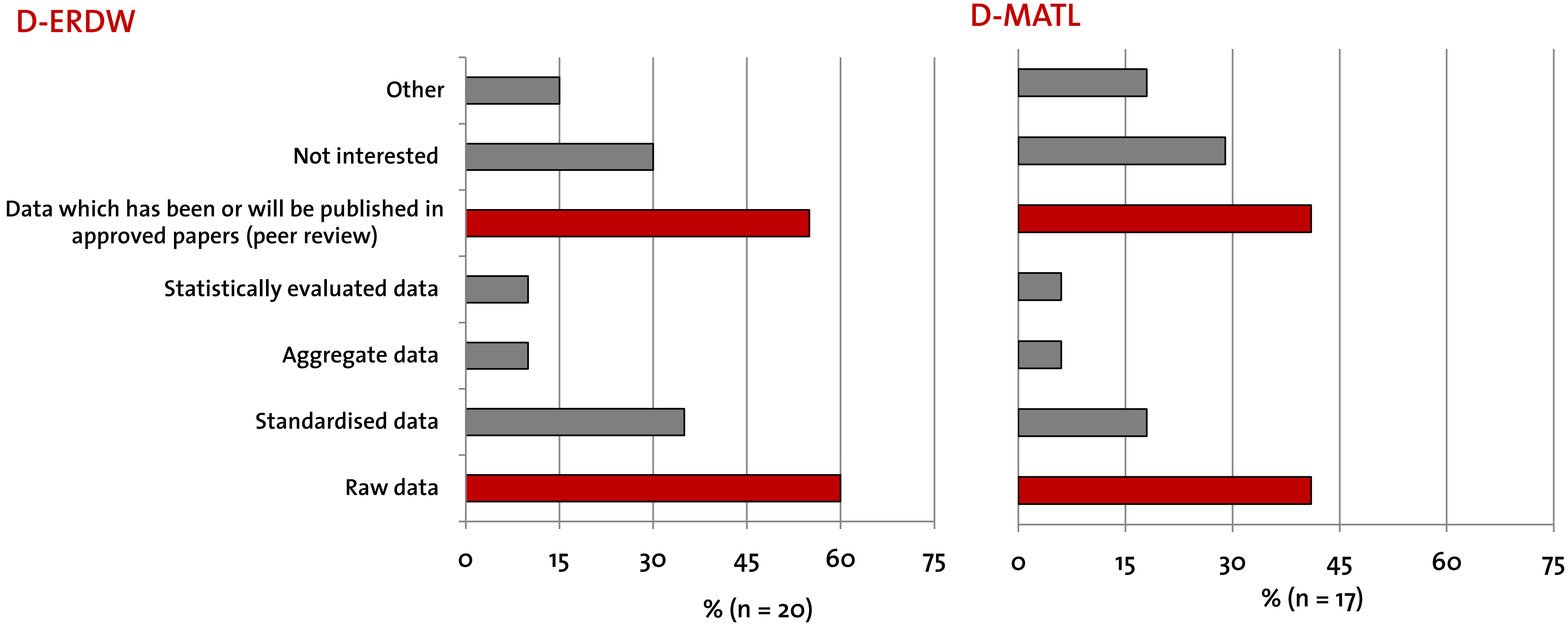


Other: GIS, ACCESS, FILEMAKER, POSTGRES, ILLUSTRATOR, PHOTOSHOP, InDesign, DREAMWEAVER, MATLAB, FILEMAKER PRO, ASCII .....

Other: ADOBE ILLUSTRATOR + PHOTOSHOP, MATLAB, ASCII, VMD, NAMD, MATHEMATICA .....

SURVEY „HANDLING RESEARCH DATA“ : FIRST RESULTS

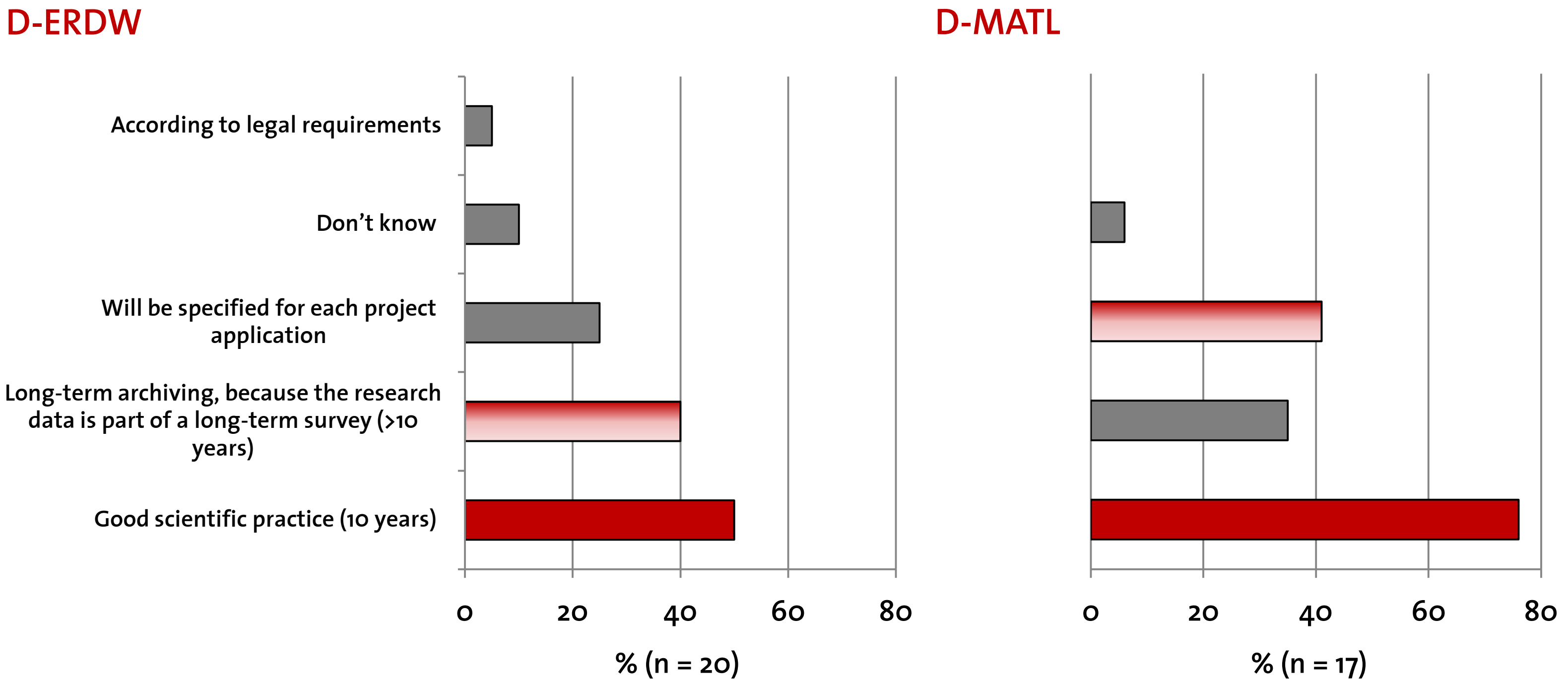
If the ETH-Bibliothek were to provide you with a database for storing your research data, at which level would you or your research group like to save data?



→ *Pilot projects are focused on managing and archiving research data published in papers (status preprint) or data sets.*

SURVEY „HANDLING RESEARCH DATA“ : FIRST RESULTS

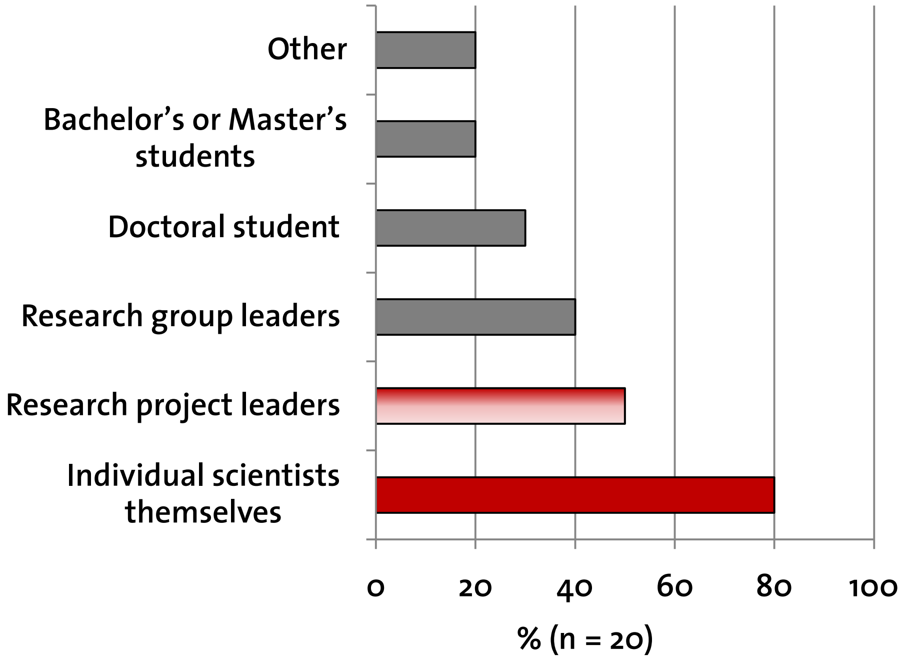
How long a period do you or your research group have in mind for storing data?



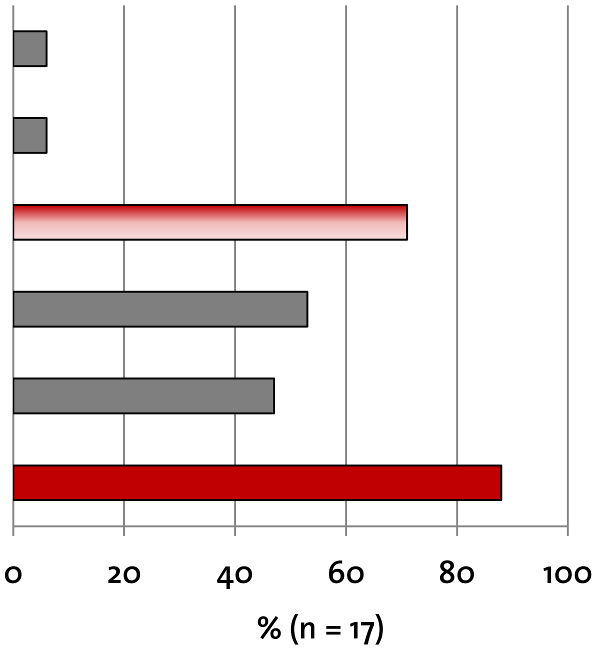
SURVEY „HANDLING RESEARCH DATA“ : FIRST RESULTS

Who is generally the owner of the research data, or who manages it?

D-ERDW

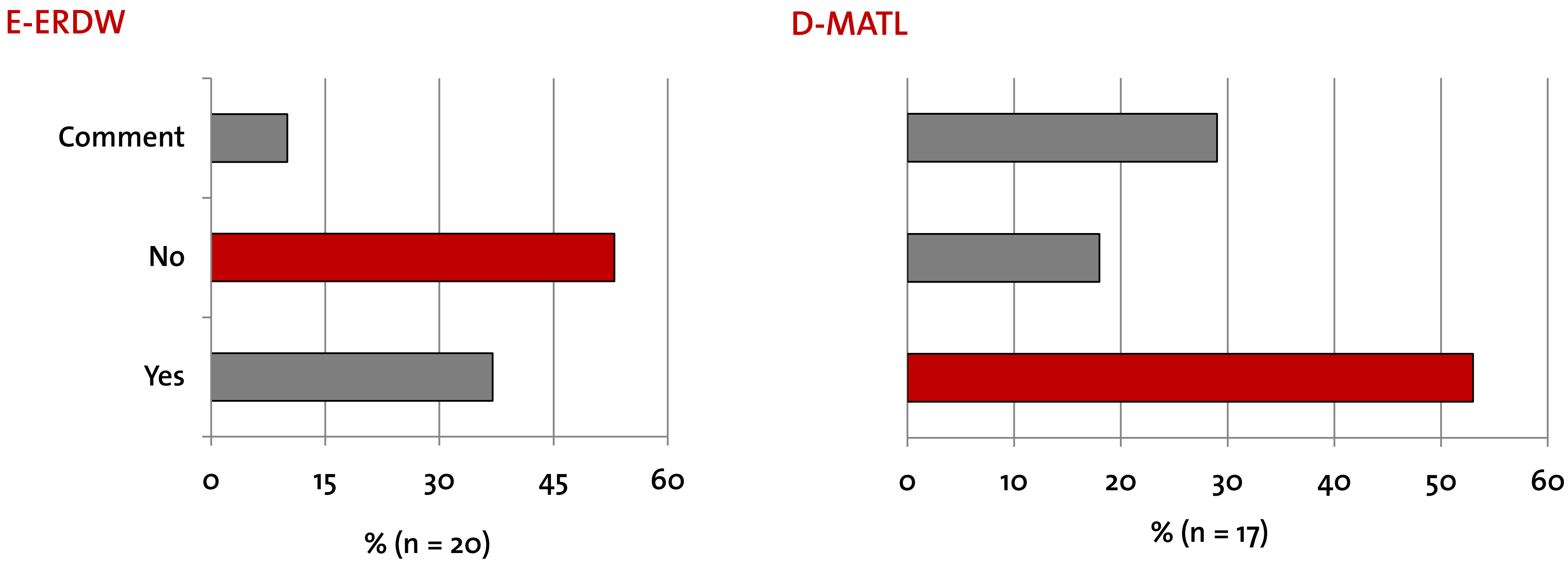


D-MATL



SURVEY „HANDLING RESEARCH DATA“ : FIRST RESULTS

In your opinion, should there be a ‘data management policy’ at ETH?



Some comments:

- *To some extent, yes, but often these policies get out of control, generating more work than they actually save. so, i'd call them "guidelines" and strongly .*
- *In principle yes, but needs to be flexible enough to not put off any data collector / producer, and to satisfy wide variety of scenarios*

STATUS QUO OF DATA HANDLING

Results from the survey and from interviews

- **„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***

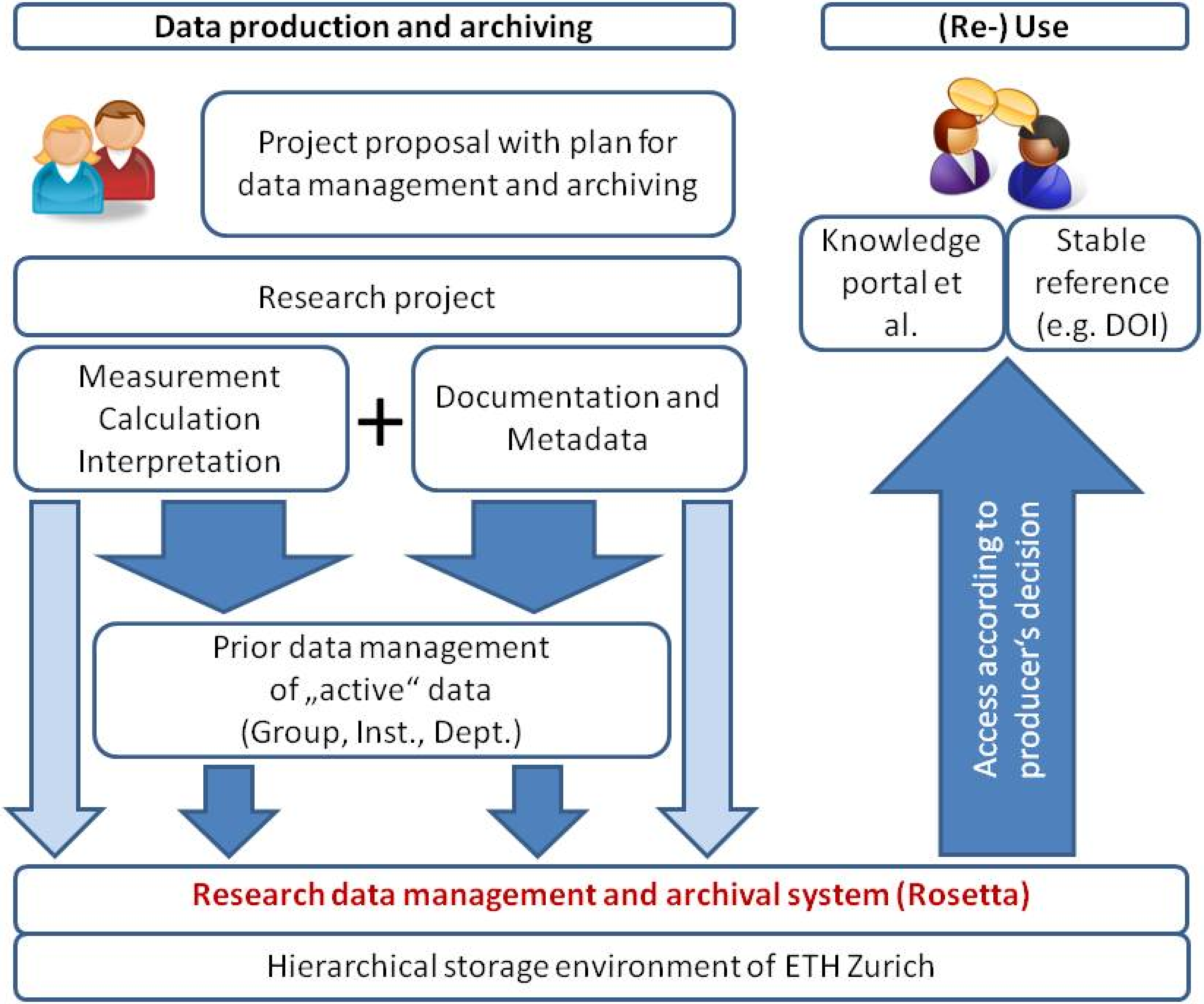
RESEARCHERS NEEDS

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

AND ROSETTA?

- 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?



Data production and archiving

(Re-) Use

Project proposal with plan for data management and archiving

Knowledge portal et al.

Stable reference (e.g. DOI)

Research project

Measurement
Calculation
Interpretation

+

Documentation and Metadata

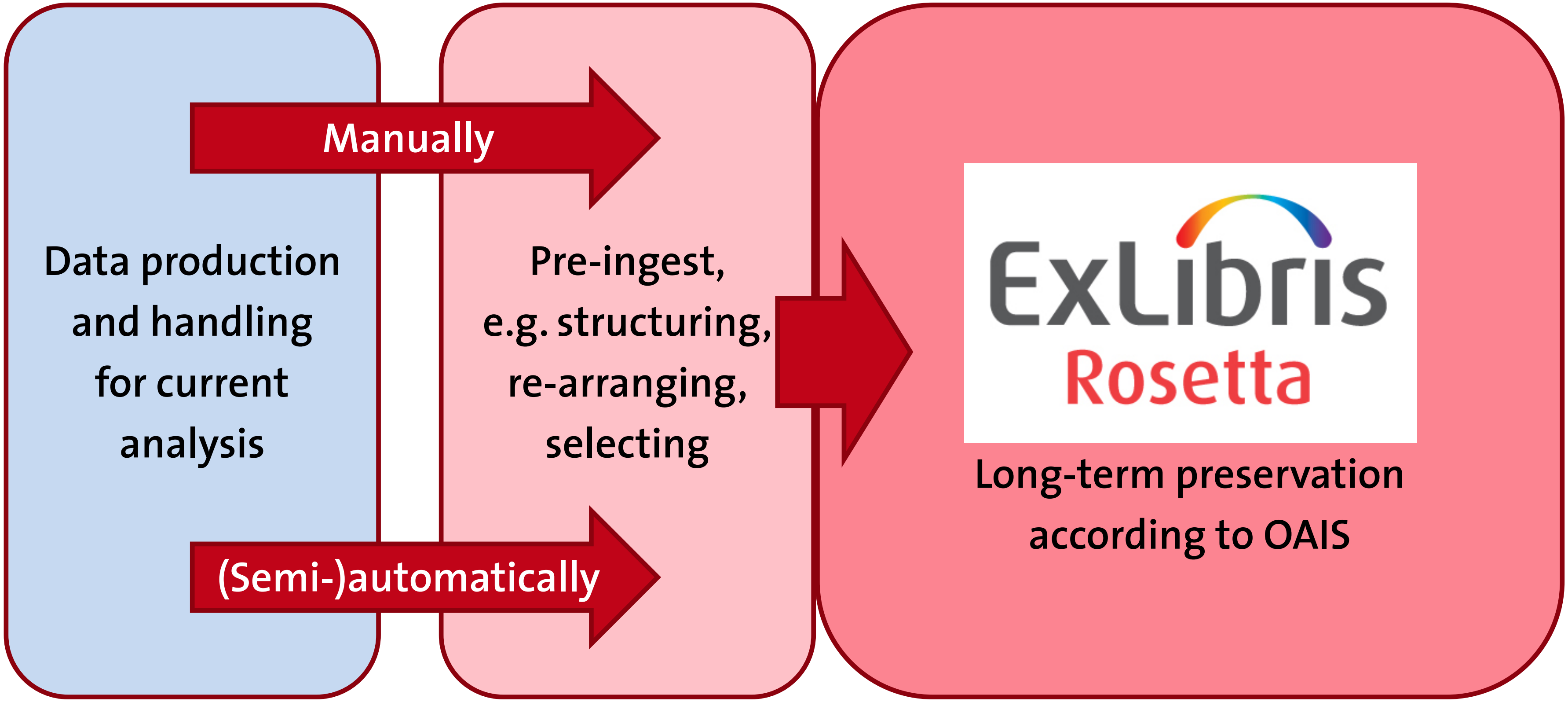
Prior data management of „active“ data (Group, Inst., Dept.)

Access according to producer's decision

Research data management and archival system (Rosetta)

Hierarchical storage environment of ETH Zurich

POSSIBLE WORKFLOWS AND ROSETTA



Data production and handling for current analysis

Manually

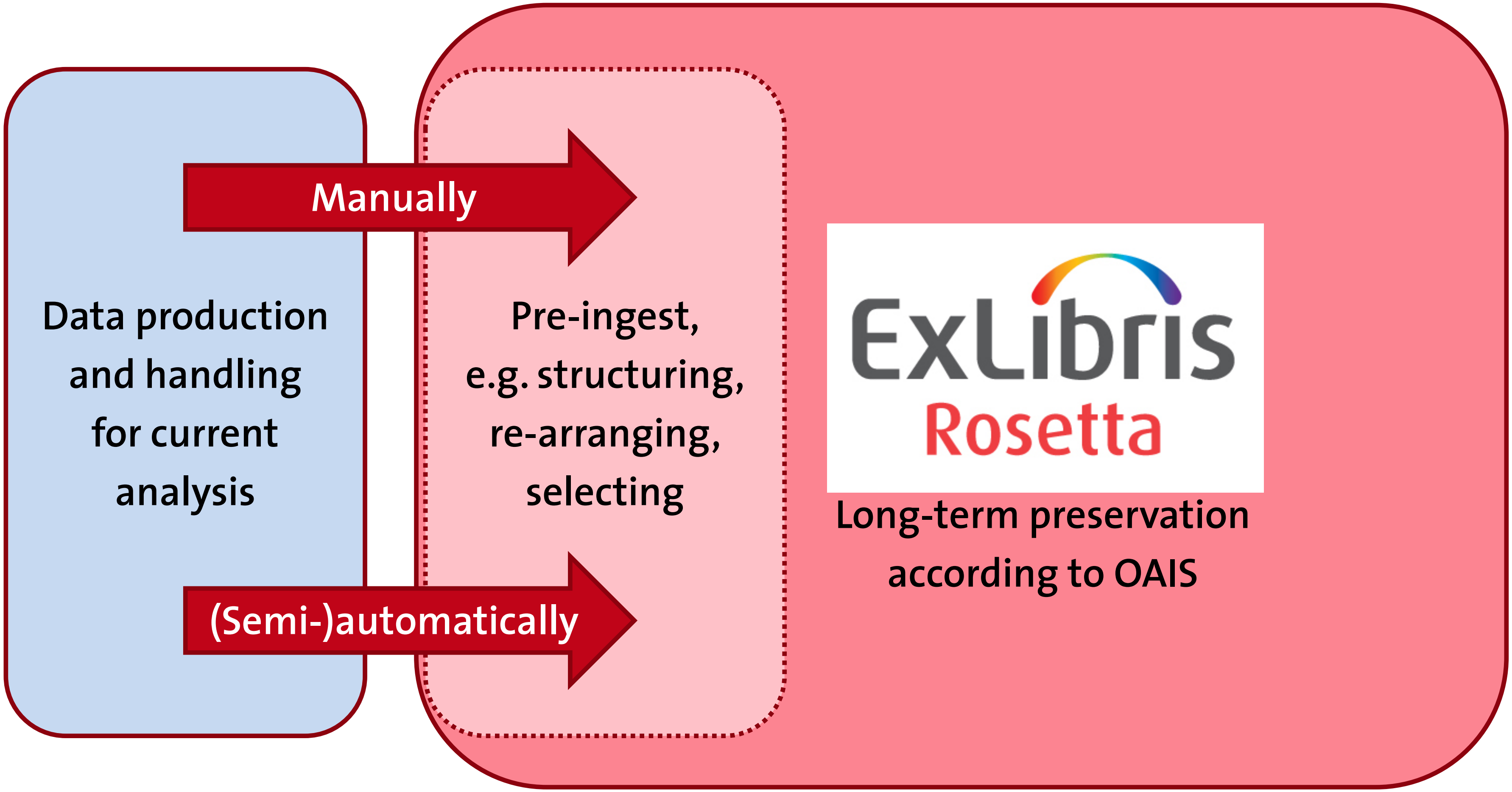
(Semi-)automatically

Pre-ingest, e.g. structuring, re-arranging, selecting

ExLibris Rosetta

Long-term preservation according to OAIS

POSSIBLE EXTENSION OF ROSETTA'S SCOPE?



Manually

(Semi-)automatically

Data production and handling for current analysis

Pre-ingest, e.g. structuring, re-arranging, selecting

ExLibris Rosetta

Long-term preservation according to OAIS

RISKS

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*

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

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*

FORMATS

A (potentially large) number of „exotic“ formats needs to be dealt with, but there might be hope:

- For analyzed data, requirements from data exchange with colleagues or from publishers for submissions can stimulate standardization
- Many formats from self programmed software are in fact ASCII-files. The challenge here is not mainly technical, but careful documentation is required to enable meaningful re-use.
- Future agreements on standards are expected within disciplines due to their specific needs for exchange, preservation and re-use

→ ***Till then, use of open standards or exchange formats needs to be further advocated***

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**
- 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
- **Convice the university's board** to grant ongoing funding as part of their risk management

THANK YOU VERY MUCH !

Questions?

Remarks?

Andreas Kirstein
Head Media and IT Services
Vice Director
ETH-Bibliothek
Rämistrasse 101
8092 Zurich
+41 44 632 26 74
andreas.kirstein@library.ethz.ch
http://www.library.ethz.ch