Digital curation of research data at ETH Zurich

Presentation

Author(s):
Töwe, Matthias

Publication date:
2011

Permanent link:
https://doi.org/10.3929/ethz-a-006783383

Digital curation of research data at ETH Zurich

Institute of Integrative Biology

15th September 2011

Dr. Matthias Töwe

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

OUTLINE

1. Background: issues and objectives
2. Vision
3. Current project
4. Caveats
5. Recommendations

BACKGROUND

Challenges

- **Research** process as a whole **relies on digital data**
- **Good scientific practice** requires retention of data in usable form
- **Funding organisations** require data management plans (NSF, DFG)
- **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

→ *But:* ***Digital data can only be used in a defined technical environment****, which usually remains stable for only a few years*

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

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*

DataCite
Helping you to find, access, and reuse data

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

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

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…

- Its location of storage is not known
- File or folder structures were changed without documentation
- Intransparent redundancies and versions exist
- 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***

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

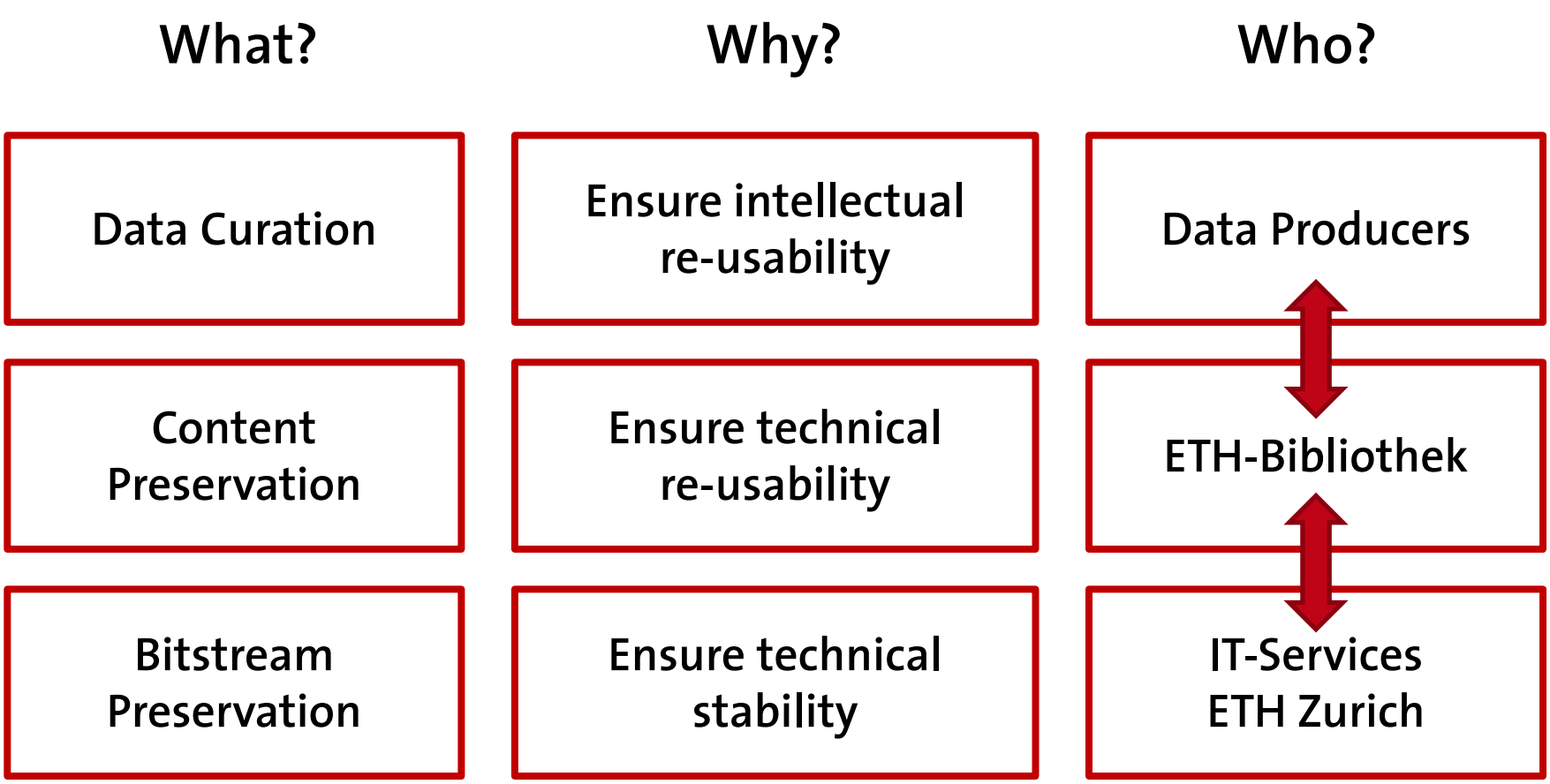
LOSS OF INTERPRETABILITY

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
- **Persons** originally responsible cannot be conctacted
- ...

FUNCTIONAL LEVELS



Adapted after Jens Ludwig, Wissgrid

WHY DOES ETH-BIBLIOTHEK BOTHER?

- **Digital curation handled by researchers themselves:**
  - Possible in principle
  - Time consuming
  - Supportive of research productivity, not productive research in itself
- **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

COMPLEMENTARY APPROACHES

Storage concept (IT services ETH Zurich)

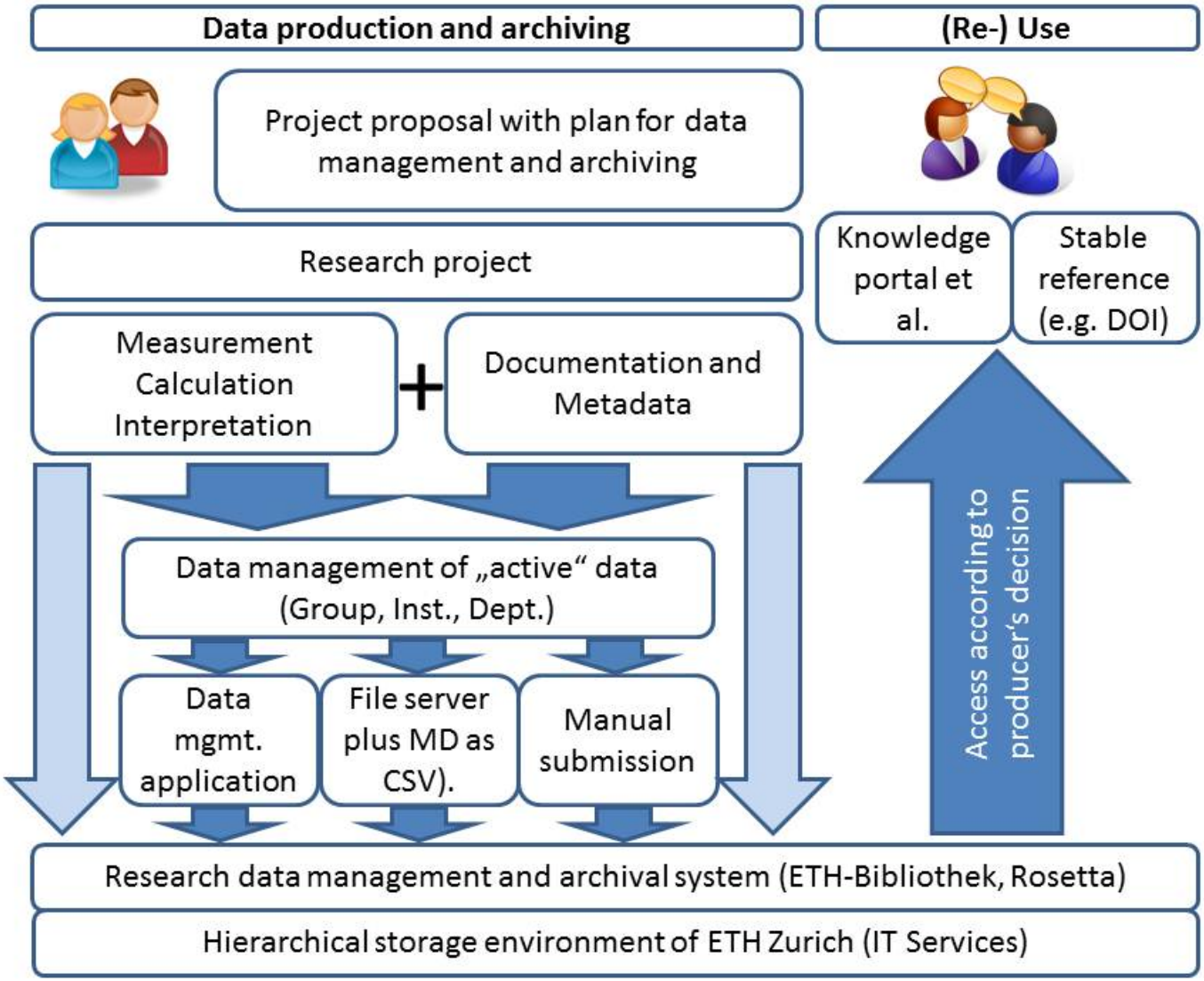
- Powerful **hierarchical storage** of all data of ETH Zurich
- Control through **classification** of data
- **Efficient operation** of the necessary infrastructure

Digital Curation (Digitaler Datenerhalt, ETH-Bibliothek)

- **Active preservation management of data** relevant for the longer term or in permanence in order to ensure their availability and usability.
- Fokus: documented **research data, direct cooperation with researchers**
- **Interfaces** with existing source systems

→ *Reliable storage infrastructure is necessary, but not sufficient*

POSSIBLE ROUTES INTO THE DATA ARCHIVE



WHAT HAPPENS INSIDE?

- **Upon Ingest:**
  - **Validation** including virus check, fixity check (checksums), format validation, extraction of technical metadata from the file itself
  - **Packaging** of the object and its representations together with all metadata into an *Archival Information Package* (AIP)
- **In permanent storage**:
  - For any use, object is copied to a working area, upload of changed object creates **new version**
  - Based on technical metadata, **periodic checks** e.g. for fixity can be scheduled
  - **File formats** can be monitored against an international knowledge base of formats for risk assessment

WHAT HAPPENS INSIDE?

- **If a format is identified as at risk:**
  - **Testing** of external tools for migration
  - **Validation** of results: are significant properties preserved?
  - **Execution** on the objects in question
  - **New versions are created, the original is preserved**
- **This is no magic device**: **there must be tools** available which can be employed
- Therefore **formats should be accepted restrictively:**
  - If they can be **properly identified and technical metadata** can be extracted: long-term preservation can be supported
  - If not: only **preservation *as is*** can be ensured. Still new tools might appear.

CURRENT STEPS

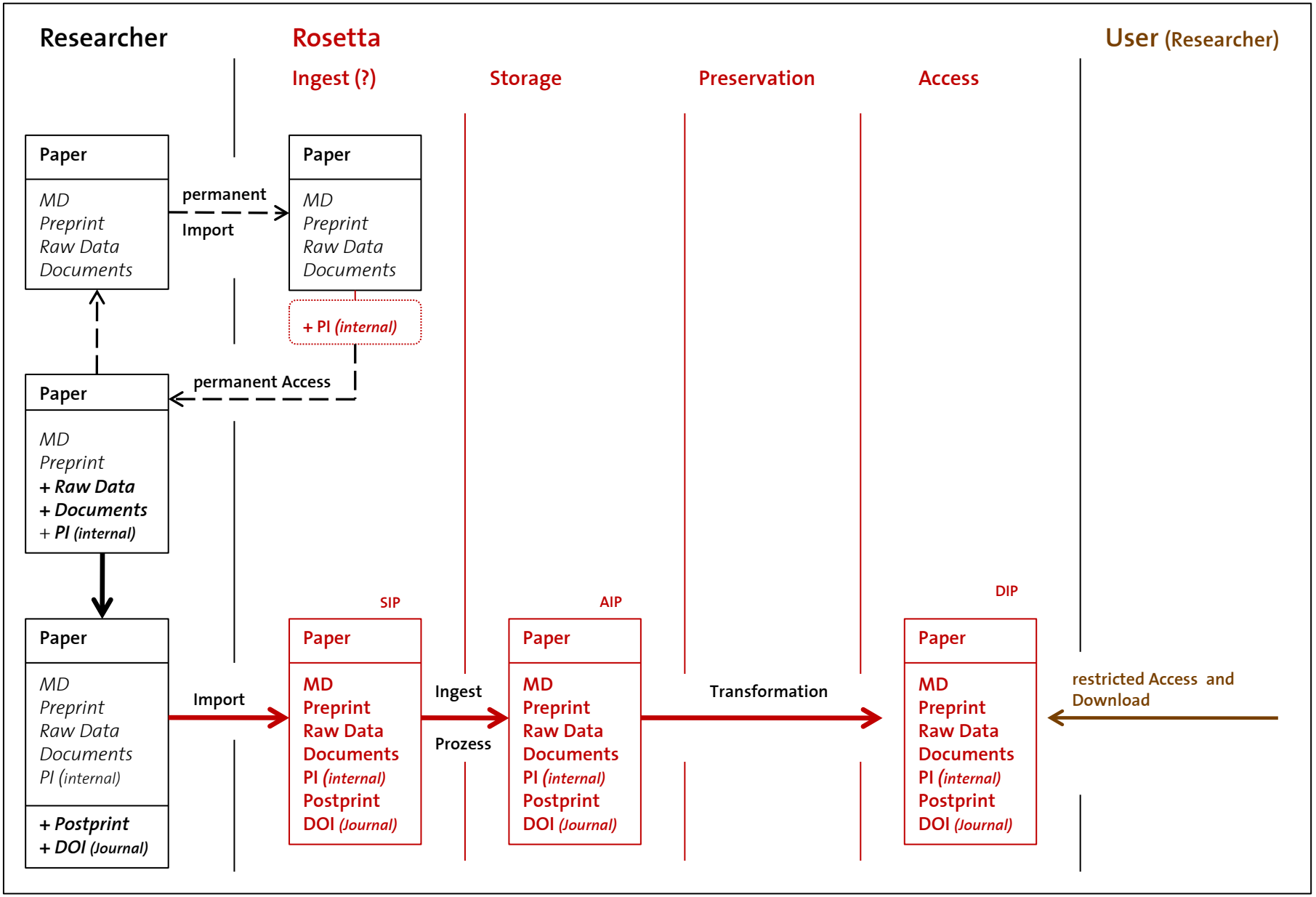
- **Pilot phase with four to five pilot partners**
- **Different requirements**
- **Proposed workflows**
  - Manual data management around a publication ready manuscript
  - (Group) datasets as raw material for following publications
  - Automatic import from an existing data management application (e.g.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

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

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

Example workflow (manual, publication oriented)

MD: Metadata; **PI**: Persistent Identifier; **SIP**: Submission Information Package; **AIP**: Archival Information Package; **DIP**: Dissemination Information Package



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)

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

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

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
- **Written agreement** between data producer and data archive on formats, procedures and access rights
- Management of **active data not treated in current project**

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 which kind of purpose
    → *affects **what** we want to preserve in which form*
- **No «rocket science», but an ongoing task with complex dependencies and a lot of work behind**
- **There is no absolute safety** against willful attacks: On the server level, manipulations are possible, but they won't go unnoticed

RECOMMENDATIONS

Independently from decision to use data archive or not and from retention period:

- «Know what you have»
- Agree on **internal policy on data handling for the group** and enforce it
- **Estimate who** might want to re-use data **and when («designated community»)**
- Make sure that **documentation and metadata are sufficient for the purpose**
- **Encourage the use of uniform file naming conventions**
- **Use pre-defined folder structure** agreed by the group
- **Avoid unclear redundancies**
- **Clearly mark versions**
- **Don't keep** data on unmanaged offline media
- **Use broadly accepted formats**, preferably openly documented ones
- **Convert proprietary formats** from vendors into exchange formats accepted in your community
- **Keep key information** in human readable form

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