Rosetta at ETH Zurich –
routes into the digital archive

IGeLU
7th Conference
Zurich, 11th to 13th September 2012
Dr. Matthias Töwe
OUTLINE

1. ETH Zurich and ETH-Bibliothek
2. Background and objectives
3. Integration of Rosetta
4. Approach for local data management
5. Ongoing work
ETH Zurich (2011)

- Swiss Federal Institute of Technology Zurich
- Founded in 1855
- University for technology and the natural sciences
- More than 17‘000 students from 80 countries, 3‘700 among them are doctoral candidates
- More than 450 professors in engineering, architecture, mathematics, natural sciences, system-oriented sciences, and management and social sciences
• **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](http://www.nebis.ch))
  • Consortium of Swiss Academic Libraries ([http://lib.consortium.ch](http://lib.consortium.ch))
  • Swiss electronic library e-lib.ch ([http://www.e-lib.ch](http://www.e-lib.ch))
  • National digitization projects with more than 5.5 mio. pages:
    • [http://www.e-rara.ch](http://www.e-rara.ch) (rare books)
    • [http://retro.seals.ch](http://retro.seals.ch) (Swiss journals)
OBSERVATIONS SINCE (AT LEAST) 2002

- Library had turned digital to a degree where ensuring long-term preservation and access became major issues.
- «Sooner or later we will need an OAIS-compliant system to support long-term preservation»
- No product in the market
- No chance to do full development ourselves
- Since then: increasing pressure
  - Digital deposits on the horizon for ETH Archives
  - Research data gaining attention within ETH Zurich
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** 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
OBJECTIVES

• Facilitate **re-use of data**
• **Accountability**, re-assessment during a **limited period**
• **Citability** of data referred to in publications
  → **DOI-registration** already operational
• **Re-use** by own group, by **granting access** to known colleagues or as fully **Open Data**

• **Preserve** the scientific and historic **heritage of ETH Zurich**
• **Administrative records** from ETH Archives
• **Library materials** (born digital theses, digitization masters; **excluding** licensed / acquired e-journals and e-books)
WHY ROSETTA?

• Concept in accord with OAIS reference model (Open Archival Information System, ISO 14721)

• Developed with partners from the library community → including side effects

• **Supports** relevant functions in preservation

• Makes use of **available tools** from the wider LTP community

• Includes a **format library** which interacts with PRONOM
WHY ROSETTA?

• Opportunity for a development partnership...

• ...based on an «out of the box» product

• **Scalability** in both object numbers and data volume

• Operation on **virtual servers** corresponding to ETH’s IT strategy

• **Integration** with existing system environment at ETH-Bibliothek using available know-how

• Strategic preference for **vendor supported** products
Data production and archiving

Researchers
Measurement, Calculation, Interpretation

Administration
Documents, Records

Library
Digitization, Licensing, Deposit

(Re-)Use
Knowledge portal et al.
Stable reference (eg. DOI)

Selection
Fileserver, application etc.
Fileserver, SAP etc.
Fileserver, Online-platform

Access according to producers’ choice

Digital long-term archive and management of permanently relevant data (Rosetta)
Hierarchical storage environment of ETH Zurich
OPTIONS FOR SUBMISSION

• Manual:
  Web dialog for upload and metadata entry by staff users

• Semi-automatic:
  Batch-upload of files with existing metadata in CSV-spreadsheet

• Automated:
  • Submission application packages structured batches of files with existing metadata in XML-format and submits them
  • Submission application may also interface directly with existing source applications
THE «SIMPLE» CASE: LIBRARY MATERIALS

Library materials:

- Born digitals, digitization masters
- **Unique** to ETH’s heritage
- **Dark archive**, online-access on dedicated platform
- **Metadata and files** are available **in a structured way** from well-defined sources
- **Example use case:** ETH E-Collection (institutional repository)
ROSETTA’S ROLE: USE CASE «E-COLLECTION»

MD or fulltext query

Link to «external resource»

Files

Submission Application

MD

SIP

Primo

Rosetta

MD

NEBIS (Aleph)

Full View of Record
Record 1 out of 1
Display format: Standard format Catalog card Citation MARC tags

FMT

LDR

DD5

008 11
0247 |a 10.3921/ethz-a-006480645 |2 doi
040 |a Su20k9 NEBIS ETH-086
245 |a Work-Life Conflict und Gesundheit (in der Zusammenhang) Populationen |c von Michaela Katharina Knecht-Meier
FUNCTIONAL LEVELS

What? Why? Who?

Data Curation
Ensure intellectual re-usability
Data Producers

Content Preservation
Ensure technical re-usability
ETH-Bibliothek

Bitstream Preservation
Ensure technical stability
IT-Services ETH Zurich

Adapted after Jens Ludwig, Wissgrid
CONCERNS WITH RESEARCH DATA

- **Keep it simple** – at least for researchers
- Digital curation cannot «improve» data retroactively: «garbage in – garbage out»
- **Researchers need to contribute** (e.g. structure, metadata, documentation)
- **Can we actually «raise awareness»** with researchers? → begin with those with high level of awareness and interest
- **Who decides** about data when the producer is no longer available?
- **Public access** must not be a prerequisite for preservation
- Researchers should be provided with a convincing **added value service which makes their lives easier**
Researchers’ requirements only partly concern functions within the OAIS-frame

High degree of **flexibility required in Pre-Ingest or even earlier**

**Consequences for the role of Rosetta** (or any LTP-system):

- Little use in further increasing complexity of LTP-system with new functionality which has little to do with LTP
- Where available, transfer data from existing upstream applications
- Achieve **flexibility in local data management** as needed, not in central application
LOCAL DATA MANAGEMENT

- **SIP Package Handler** (Docupack)
- **Viewer and editor** for structure and metadata
- **Data and metadata** remain on each research group’s local storage as long as desired
- **Creates Submission Information Package (SIP) of selected data** for ingest into Rosetta when archiving is initiated
- **Origin in archival sector** no coincidence: requirements are similar
RESEARCH DATA MANAGEMENT → PRESERVATION

User

Structuring and metadata-capture, DOI-generation

Point of time X: selection for archiving and «Submit» incl. access rights in archive

Assigned storage (local PC, share); permissions via file system functionality

Library

Ingest, DOI-registration

Rosetta

Delivery via DOI, if rights ok
WORKFLOW FOR ETH ARCHIVES

1. Offer of deposit
2. Appraisal ok?
3. Appraisal approved
4. Ingest, DOI-registration
5. Appraisal, selection, description embargo period?
6. DOI-generation
7. Full MD and structure
8. Deliver full MD and structure
9. Assigned storage, temporary access for depositing unit
10. Archival Information System
11. MD
12. Primo
13. Eth Archives
14. Library
15. Rosetta
16. Organisational unit of ETH
ROSETTA: COMMISSIONED DEVELOPMENTS

- Collections API
  - Create/delete collection via submission of an externally produced METS-file
  - Update collection metadata and location of an existing collection
  - Get collection by ID or path

- Retention Period Functionality
  - Manage Retention Policies based on date or period
  - Assign Retention Period through task or via METS
  - Reports on IEs to be deleted and job for deletion

- Collection Viewer
  - Display of collections’ context (tree structure)
  - «Explorer-like» navigation
WORK SO FAR

✓ DOI-registration by ETH Zurich as member of DataCite operational

✓ Full survey of research groups (Profs.) at ETH Zurich

✓ Identification of pilot partners and their requirements

✓ Check and update inventory of data hosted by the library

✓ Submission application for institutional repository
  o (Productive operation planned to start 09/2012)

✓ Gap analysis and decision on development

✓ Migration to version 3.0, now 3.0.1
WORK TO DO AND IN PROGRESS

✓ Commissioned developments Rosetta
  ✓ Specification
    o Development (Beta-version 10/2012)
    o Testing (10-11/2012)

✓ Development and testing for local data management
  ✓ Specification
    o Development (beta versions between 07-10/2012)
    o Testing (Continuously between 07-10/2012)

✓ Ongoing work with respect to service provision (2012)
  o Definition of service levels
  o Analysis of formats in pilot groups
  o Definition of white list for supported formats
  o Drafting of a data policy to support researchers
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