Presentation

Digital Curation at ETH-Bibliothek
Ex Libris Information Day Zurich, 29th August 2012

Author(s): Töwe, Matthias

Publication Date: 2012

Permanent Link:
https://doi.org/10.3929/ethz-a-007362271

Rights / License:
In Copyright - Non-Commercial Use Permitted
Digital Curation at ETH-Bibliothek: Preservation with Rosetta and More

Ex Libris Information Day
Zurich, 29th August 2012
Dr. Matthias Töwe
1. Background and initial objectives
2. Result of gap analysis
3. Aspects of local data management
4. Current status
5. Outlook
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

→ But: *Digital data can only be used in a defined technical environment, which usually remains stable for only a few years*
• 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 could range from granting access to known colleagues to fully Open Data, if desired

• 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)
• **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
## FUNCTIONAL LEVELS

<table>
<thead>
<tr>
<th>What?</th>
<th>Why?</th>
<th>Who?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Curation</td>
<td>Ensure intellectual re-usability</td>
<td>Data Producers</td>
</tr>
<tr>
<td>Content Preservation</td>
<td>Ensure technical re-usability</td>
<td>ETH-Bibliothek</td>
</tr>
<tr>
<td>Bitstream Preservation</td>
<td>Ensure technical stability</td>
<td>IT-Services ETH Zurich</td>
</tr>
</tbody>
</table>

Adapted after Jens Ludwig, Wissgrid
Data production and archiving

Researchers
- Measurement
- Calculation
- Interpretation

Administration
- Documents
- Records

Library
- Digitization
- Licensing
- Deposit

Selecting
- 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

(Re-)Use

Knowledge portal etc.
Stable reference (eg. DOI)
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
• Can we actually «raise awareness» with researchers?

• Do we bother researchers with this – or have they been waiting?

• 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

• Some researchers might not really need LTP, but something else

• There are researchers who do not want to be involved

➔ Best start with those who actually want a change
WHY ROSETTA?

- Concept in accord with OAIS reference model (Open Archival Information System, ISO 14721)
- Developed with partners from the library community
- Supports relevant functions in preservation
- Makes use of available tools from the wider LTP community
- Relies on 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
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
DATA MANAGEMENT AND OAIS?

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

Structuring and metadata-capture, DOI-generation

Ingest, DOI-registration

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

Delivery via DOI, if rights ok

User

Library

Rosetta

DATA MANAGEMENT ➔ PRESERVATION
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 second half of 2012)

✔ Gap analysis and decision on development

✔ Migration to version 3.0
WORK TO DO AND IN PROGRESS

✓ Commissioned developments Rosetta
  ✓ Specification
    o Development (Beta-version 09/2012)
    o Testing (09-10/2012)

✓ Development and testing for local data management
  ✓ Specification
    o Development (Several 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