Presentation

Long-term preservation of digital content
digital curation

Author(s):
Töwe, Matthias

Publication Date:
2011

Permanent Link:
https://doi.org/10.3929/ethz-a-006691684

Rights / License:
In Copyright - Non-Commercial Use Permitted
Long-term preservation of digital content

ETH Zurich, ETH-Bibliothek
Zurich, 20 May 2011
Dr. Matthias Töwe
• **Data from digitization projects**
  • Selected documents from special collections, (older) theses, journals, images, rare books

• **Born digital content**
  • Up to now e.g. theses, selected lectures
  • Increasing in future: Research data and administrative records (ETH Archives), hosted Swiss journals

• **Licensed and acquired content**
  • E-Journals, e-books
DIGITIZATION PROJECTS

„Boutique digitization“
- Focus on ETH-Bibliothek’s special collections
- Outstanding documents, e.g. letters from Einstein
- Small to medium scale collections, e.g. protocols of ETH’s board

Mass digitization
- Materials from ETH-Bibliothek‘s vast image archive:
  - E-pics → www.e-pics.ethz.ch
- Content from Switzerland: ETH-Bibliothek is running the platforms for two nationwide programs
  - Retro.seals.ch (scholarly journals) → http://retro.seals.ch
  - E-rara.ch (rare books) → http://www.e-rara.ch
CHARACTERISTICS

• Digital copy does not replace the original
• Access is the main objective
• Safety copies of unique material
• Originals are preserved
  • Journals: At least one copy
  • Rare books
  • Images: photographs are most endangered
• Increasingly born digital content
• „Boutique digitization“:
  • „It might be cheaper to re-digitize material than to preserve the digital copy.“

• Mass digitization:
  • Considering loss / damage of single files: Re-digitization of preserved originals is an option
  • Considering the risk of loss of whole collections: Because of large quantities reproduction would be very costly and extremely time-consuming.
• **Example: Journals in retro.seals.ch**

  • Digital image master files (TIFF)
  • IDX-files comprising metadata and OCR-fulltext
  • XML-file describing each volume’s structure
  • JPEG-files as access copies
  • PDF-files for download (generally created on the fly, only publishers‘ original files for current journals are preserved)

→ *All can be re-created from the image masters, but by costly and time-consuming procedures*
Files are employed in heavily used and regularly maintained applications:

- Format obsolescence would have been recognized early
- No case of format obsolescence has come up

However, only bitstream preservation has been addressed systematically so far

There is...

- ...no logging of files‘ change history
- ...no systematic support for preservation planning
- ...no support for migration procedures
(Applications and all data are of course included in regular backup)

• **Journals in retro.seals.ch**
  - All relevant files written to off-site tape-archive

• **Rare books in e-rara.ch**
  - ZIP-archive capsules are formed containing both TIFF-images and XML files with structure and metadata
  - For the time being, data is delivered and taken care of by the library owning the original
  - ETH-Bibliothek’s data is written to off-site tape-archive; hierarchical storage management (HSM) in place
Underestimated risks:

- Unclear or changing responsibilities
  → *Loss of contextual information*

- Missing or insufficient documentation
  → *Can hardly be completed in retrospective*

- Unclear, haphazard, intransparent file structures
  → *Risk of redundancies and losses*

→ *These and further risks can make long-term preservation of any data questionable*
- ETH-Bibliothek has created a team dedicated to digital curation
  - All kinds of data
  - Current focus on research data
- It is currently implementing Ex Libris Rosetta to support long-term preservation
  - Support of configurable workflows for processing of files
  - Logging of all changes to files and their metadata
  - Monitoring of formats and upcoming risks of obsolescence
  - Preservation planning according to identified risks
• ETH Zurich‘s IT-Services are introducing a new storage infrastructure
  • Robust and highly scalable storage is a necessary but not sufficient prerequisite for any kind of long-term preservation
  • Concepts of hierarchical storage management can be applied more efficiently
  • Based on rules for pre-defined classes of data
  • Prerequisite for automation
In the medium to longer run:

1. Investment into indexing and structuring will get same attention as digitization itself
   - Image masters are not enough
2. Convergence between online applications and archival solutions
   - Two sides of the same medal
   - Much simpler workflows
3. Archival needs as criteria for format development and choice
4. Archival measures integrated earlier in the lifecycle
5. Consolidation of support functions, e.g. format monitoring
No application will „solve“ long-term preservation

- Software can support monitoring, management, processing and decision making
- Informed decisions need to be taken throughout any data’s lifecycle
- Institutional structures and processes must ensure the stability of data management
- Digital preservation must not be regarded as a niche for a specialized team
- It must be well rooted within the whole organization in order to identify risks already when data is created
• **Established rules still apply:**
  
  • Using data or having it used is the best way not to lose it  
  • Data not in use is at risk  
  • Data on physical offline media is hard to track and at risk of being „forgotten“ – even if the media are ok  
  • Prefer open, long lived standard formats  
  • Know and document  
    • What you have  
    • Where it is  
    • Who is responsible
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

Animation on Rosetta:

http://www.youtube.com/watch?v=I7cCD6Htaml