Research Data Management: lessons learned - and still to learn

SWITCH Research Data Management (RDM) Workshop, 15. Dezember 2014

Dr. Matthias Töwe, ETH-Bibliothek, ETH Zürich
Overview

- Digital Curation Office
- Background
  - Survey
  - Research Data
  - Aims
- Services
  - ETH Data Archive
  - Workflow «Small Data» with docuteam packer
Digital Curation Office

- Team within **Customer Services Department**
  - 2.7 FTE (+1 per 2015)
  - plus 1.8 FTE in Library IT Services

- Initiated as a project in 2010/2011 after years of observation

- Productive operation since 2014 (some services earlier)

- Original approach rooted in digital **preservation** and first studies focused on library content (master files from digitization, but also e-journals)

- Today much more comprehensive understanding of services in **digital curation and data management** and focus is on **research data**
Vision for diverse data types

Data production and archiving

- Researchers
  - Measurement Calculation Interpretation
- Administration
  - Documents Records
- Library
  - Digitisation Licensing Deposit

(Re-) Use

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

Access according to producers’ choice

Selection

- Local storage? Deletion?
- Auditable filing? Deletion?
- Local storage? Deletion?

ETH Data Archive (Rosetta)

Hierarchical storage environment of ETH Zurich
Currently not within scope

- **Management of «active» research data** which is being worked on
- **«Big Data»** in the sense of large unstructured heaps of data which need to be processed regularly
- **Ensuring Post Cancellation Access and Preservation of licensed content** (e-journals and e-books)
  - Reason: **content is not unique to ETH Zurich** and can therefore be preserved in collaborative initiatives
  - ETH-Bibliothek is member of both **Portico and Global LOCKSS** Network
    - Membership is managed by colleagues from licensing
    - Selection of titles in LOCKSS is managed by colleagues from periodicals
Idealised life-cycle of research data

- Hypothesis / Research question
- Data capture or -collection
- Re-use
- Access and verification
- Analysis and interpretation
- Publication
- Synthesis

Digital Research Data
2. **Part** - How long a period do you or your research group have in mind for storing data?

---

**Survey and diagrams:** S. Scheid
Formats in use?

2. Part - Which formats do you and your research group work with?
2. **Part** - 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?

Survey and diagrams: S. Scheid
Aims

- Facilitate **re-use**
- Support the **publishing** of research data
- Offer **DOI-registration** as part of the service
- Keep published data **permanently available**
- Where possible and reasonable: plan active **preservation measures** (format migration)
- Facilitate the **verifiability** of results
- Support guidelines for **good scientific practice**
- Offer low-barrier solution(s) for **safeguarding data for limited periods** of time
- **Work closely with IT Services** on coordination and delimitation
- Support data management

- **Gain know-how** and establish ETH-Bibliothek as service provider in this context
Our services

- Advice on data management (*under construction*)
- ETH Data Archive (Ex Libris Rosetta)
  - Safeguarding data for limited periods of time (at least 10 years)
  - Permanent safeguarding
  - Preservation measures (format migration)
  - Mass processes and individual objects
- docuteam packer
  - Viewer and editor for locally created file structures with metadata
  - Preparation of data for submission to ETH Data Archive
Access Rights and Retention Periods

**Access Rights:**
- Open Access
- Restricted Access
  - User ID
  - User Group
  - IP Range
  - Embargo Period depending on a defined date or relative to a date field in the metadata

**Retention Periods:**
- Unlimited
- Limited, with «automatic» deletion
  - After a defined date or relative to a date field in the metadata
The (very) abstract view

« Well defined data sources »

ETH Data Archive (Rosetta)
- Ingest
- Manage
- Preserve
- Publish
- Deliver
Rosetta as part of the systems landscape

Producers
- Forschende
- Hochschularchiv
- Repositorien
- Bibliotheks-inhalte

Pre-Ingest
- Docuteam Packer / Feeder
- Submission Application

DOI-Registrierung

Consumers
- Wissensportal
- Archivdatenbank
- Repositorien

ETH Data Archive (Rosetta)

Manueller Upload

Storage

(Illustration: Franziska Geisser)
Methods for Deposit and Upload

- **Manually:**
  Web dialogue for upload and metadata capture

- **Semi-automatically:**
  Batch-Upload of files with existing metadata (CSV)

- **Automatically:**
  - Adapted Submission Application packages structured files with existing metadata in XML-format
  - Submission Application can also be implemented as interface with an existing source application
  - **Automatically after manual preparation** in docuteam packer (local viewer and editor for file-structure and metadata)
Submission Application

E-rara Submission App.

E-Collection Submission App.

Docuteam Feeder

Rosetta SIP

(Illustration: Franziska Geisser)
Submission Application

- Packages files from a source system
- Processes source metadata (mapping to Dublin Core)
- Generates the directory structure required by Rosetta (> Collections)
- Generates a Rosetta Ingest-METS

(Illustration: Franziska Geisser)
Publishing und Delivery

- **Discovery** is performed outside of Rosetta:
  - OPAC / Primo
  - Archival database
  - Other Content Management System

- **Publishing** via OAI-PMH, SRU Request or Web Service

- **Delivery** from Rosetta:
  - depending on defined Access Rights
  - out of the box or additional viewers (plug-ins)
## Access via Knowledge portal (Primo)

### Ergebnisse 1 - 10 von 93

<table>
<thead>
<tr>
<th>#</th>
<th>Title</th>
<th>Authors</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Electrochemistry Data (NMC_94wt_0bar)</td>
<td>Martin Ebner</td>
<td>2013</td>
</tr>
<tr>
<td>2</td>
<td>Electrochemistry Data (NMC_90wt_0bar)</td>
<td>Martin Ebner</td>
<td>2013</td>
</tr>
<tr>
<td>3</td>
<td>Binary Data (NMC_94wt_300bar)</td>
<td>Martin Ebner</td>
<td>2013</td>
</tr>
</tbody>
</table>
Link into ETH Data Archive (Rosetta)
«View» in Rosetta: Images
Typical customer requests

«I want to submit a manuscript for an article. The editor demands that raw data is deposited in a repository. What can I do?»

«Up to now we have been archiving data to our doctoral theses on CD-ROM and we would like to find a better solution. How can we do this? -- Oh, and our professor is retiring in n months.»

«We want to link from an article research data which we would like to be publicly available. We have analysed the data with our own methods and now other groups might want to look into them with their methods. How can we do this?»
Productive Use Cases

- **Dark Archive for ETH E-Collection** (ETH-Bibliothek):
  - Periodic ingest into ETH Data Archive

- **Data cited within a publication**:
  - Data is openly available (Example from D-ITET: http://dx.doi.org/10.5905/ethz-iis-1)

- **«Professors Archives»** (Dark Archive):
  - Retention of data from CD-ROMs from doctoral theses for 10 or 15 years

- **«Software Disclosure»** (ETH transfer (Technology Transfer Office)):
  - Code-packages of software which was developed at ETH
  - ETH transfer publishes instructions for upload on its web pages
    (https://www.ethz.ch/content/associates/intranet/en/research-and-technology-transfer/inventions-patents-licenses/computer-programs/software-disclosure.html)
Use cases under preparation

- Research data archive of research group, prepared with docuteam packer
- Workflow for deposit and assessment for University Archives of ETH (docuteam packer, CMI Star)
- «Dark Archive» of Master files from digitisation (e-rara, e-manuscripta, e-periodica, e-pics)
- ETH-internal customers:
  - Archives of Contemporary History (AfZ)
  - Institute for the History and Theory of Architecture (gta)
Workflow «Small Data» with docuteam packer

Selection and documentation of context
Researchers

Structuring and DOI-generation*
Metadata capture

Filesystem (networked, local)
Selection for archiving
docuteam packer

Curation and usability
Library

Bitstream preservation
IT-Services

Point of time X: «Submit» a SIP
docuteam feeder

ETH Data Archive

Delivery via DOI
Access via Knowledge Portal

Server
Storage
Network
What is *docuteam packer*?

For users

- Viewer and editor for local preparation of archival packages for transfer to ETH Data Archive
  - Create and edit folder structure, as it should be reflected in ETH Data Archive
  - Enter and edit metadata
  - DOI-creation (Digital Object Identifier; to be registered by ETH Data Archive)
  - Assign access rights and retention periods to be enforced by ETH Data Archive

In the background

- Create a **Submission Information Package (SIP)** or **Archival Information Package (AIP)** of metadata + structure (**METS-format**, Metadata Encoding and Transmission Standard) **and data**
The GUI and its elements (1)

- Tree view of folders and files
- Statistics per element
- Technical metadata
- Events
- Event details
The GUI and its elements (2)

- Tree view of folders and files
- Descriptive metadata
- Preview functions
Example Use Cases

Research groups

- **Data belonging to a manuscript** are collected, submitted to the long term archive and made accessible via DOI for reviewers and readers.

- **Research group has a structured filing** without metadata; it should be edited and submitted into the long term archive.

- **PhD students of a group** are presented with a filing structure they should follow when managing their data.

... Administrative staff within ETH

- Delivers structured data to ETH Zurich’s university archives...
- ...archives’ staff appraises and selects content and adds metadata.
What docuteam packer is not!

- No comprehensive data management solution
- No records management solution
- No collaboration platform
- No data repository
- No long term archive - but a tool to prepare for and submit to archive
- No solution for local rights management
- Not tied to use with Rosetta as the only long term archive

➢ Consider alternative approaches where these are more appropriate
➢ Be careful with using the tool without submitting to an archive
Role within the overall concept

- There will be no obligation to use docuteam packer
- Other routes exist for ingesting content into ETH Data Archive
  - Web-Upload to Rosetta
  - Automatic batch processes based on existing metadata (e.g. exported from an application as METS-XML or CSV plus respective objects.
- Decide on best route to use after discussion with customer
- First customers going live now, including University Archives of ETH
Perspective on Swiss level

- Programme «Scientific Information» of the Swiss University Conference (SUC)
  - Proposal Data Lifecycle Management
    (new proposal under preparation for February 2015)
  - Partners:
    EPFL, ETHZ (SIS), HEG, Unis Basel, Genève, Lausanne, Zürich, SWITCH
Questions?

www.library.ethz.ch/Digital-Curation

data-archive@library.ethz.ch

Dr. Matthias Töwe
Head Digital Curation
ETH-Bibliothek
Rämistrasse 101
8092 Zurich
044 632 60 32
matthias.toewe@library.ethz.ch