Rosetta and Data Curation
in the context of research: first insights

Haifa, 11th September 2011
Andreas Kirstein
1. Introduction
2. Survey „Handling research data“ : first results
3. Status quo of data handling
4. Needs of researchers at ETH Zurich
5. Workflows and integration with Rosetta
6. Formats and other issues
7. Current and future work
ETH ZURICH – THE UNIVERSITY

ETH Zurich

• Swiss Federal Institute of Technology Zurich
• Founded in 1855
• One of the leading international universities for technology and the natural sciences
• More than 16,000 students from 80 countries, 3,500 among them are doctoral candidates
• More than 400 professors teach and conduct research in the areas of engineering, architecture, mathematics, natural sciences, system-oriented sciences, and management and social sciences
ETH Zurich infrastructure

- ETH-Bibliothek
  - Information, library and collection management for ETH Zurich
  - Main library and special libraries in departments
  - Special collections including ETH Archives
  - Library IT-Services dedicated to ETH-Bibliothek’s applications

- ETH Zurich IT Services
  - IT infrastructure management
  - IT service provider
  - Storage management

→ Aim: central services + local services managed centrally
ETH Zurich guideline on integrity of research

• Project managers must ensure that (“primary“) data is kept for as long as is appropriate for the discipline
• No common tool is available to support this

Irreplaceable data

• Unique observational data in long continuous timelines
• Other data which is to be used for comparative research

Published and / or referenced data

Administrative records from ETH Archives

Library materials (born digital theses, digitization masters)
Online survey with all professorships or research groups per department

<table>
<thead>
<tr>
<th>Department</th>
<th>abbr.</th>
<th>Number of professorships or research groups per department</th>
<th>Rate of return (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Environmental Sciences</td>
<td>D-UWIS</td>
<td>33</td>
<td>63</td>
</tr>
<tr>
<td>Agricultural and Food Sciences</td>
<td>D-AGRL</td>
<td>19</td>
<td>78</td>
</tr>
<tr>
<td>Humanities, Social and Political Sciences</td>
<td>D-GESS</td>
<td>23</td>
<td>83</td>
</tr>
<tr>
<td>Earth Sciences</td>
<td>D-ERDW</td>
<td>17</td>
<td>94</td>
</tr>
<tr>
<td>Chemistry and Applied Biosciences</td>
<td>D-CHAB</td>
<td>50</td>
<td>65</td>
</tr>
<tr>
<td>Civil, Environmental and Geomatic Engineering</td>
<td>D-BAUG</td>
<td>40</td>
<td>100</td>
</tr>
<tr>
<td>Biosystems Science and Engineering</td>
<td>D-BSSE</td>
<td>10</td>
<td>80</td>
</tr>
<tr>
<td>Mathematics</td>
<td>D-MATH</td>
<td>15</td>
<td>87</td>
</tr>
<tr>
<td>Materials Science</td>
<td>D-MATL</td>
<td>12</td>
<td>100</td>
</tr>
<tr>
<td>Management, Technology and Economics</td>
<td>D-MTEC</td>
<td>20</td>
<td>89</td>
</tr>
<tr>
<td>Mechanical and Process Engineering</td>
<td>D-MAVT</td>
<td>in progress</td>
<td></td>
</tr>
<tr>
<td>Computer Science</td>
<td>D-INFK</td>
<td>in progress</td>
<td></td>
</tr>
<tr>
<td>Architecture</td>
<td>D-ARCH</td>
<td>in progress</td>
<td></td>
</tr>
<tr>
<td>Biology</td>
<td>D-BIOL</td>
<td>in progress</td>
<td></td>
</tr>
<tr>
<td>Information Technology and Electrical Engineering</td>
<td>D-ITET</td>
<td>in progress</td>
<td></td>
</tr>
<tr>
<td>Physics</td>
<td>D-PHYS</td>
<td>in progress</td>
<td></td>
</tr>
</tbody>
</table>

Questionnaire consists two parts

**First part:** how research data are handled generally, questions are based on a survey carried out by the working group "Research data" of the Leibniz community

**Second part:** specific questions about research data, questions based on the paper „Conducting a Data Interview“ from Witt & Carlson, Purdue University Libraries, 2010
With which types of data do researchers in your discipline work?

D-ERDW

- Original software development
- Description of rock samples
- Specific research results (non-standard)
- Publications
- GIS information
- Geospatial data
- Geochemical data
- Geological Database
- Geotechnical Database
- Results from numerical models
- Raw seismic data
- Computer simulations results
- Algorithms
- Field data
- Digital images, video
- Measurement data
- Digital geological maps

number of times mentioned, n = 20
With which types of data do researchers in your discipline work?

D-MATL

- Program code for analysis
- Algorithms
- Publications
- Text
- Lab reports
- Crystal structure models
- Molecular models
- Digital images
- Movies/videos
- Simulation data
- Measurement data

number of times mentioned, n = 17
In your discipline, are there metadata standards for describing research data?

D-ERDW: 29% (n = 20) Yes, 48% (n = 20) No, 24% (n = 20) Don’t know

D-MATL: 29% (n = 17) Yes, 65% (n = 17) No, 6% (n = 17) Don’t know
What is the main type of research data worked on by your research group?

**D-ERDW**

- Other: 0% (n = 20)
- Data collected as part of international research projects: 50% (n = 20)
- Data collected as part of Swiss cooperation projects: 25% (n = 20)
- Data from long-term observations: 25% (n = 20)
- Data from individual projects: 0% (n = 20)

**D-MATL**

- Other: 0% (n = 17)
- Data collected as part of international research projects: 0% (n = 17)
- Data collected as part of Swiss cooperation projects: 50% (n = 17)
- Data from long-term observations: 0% (n = 17)
- Data from individual projects: 100% (n = 17)
SURVEY „HANDLING RESEARCH DATA“: FIRST RESULTS

Which formats do you and your research group work with?

**D-ERDW**
- Other
- PDF
- SPSS, SPLUS
- JPG, TIFF, GIF
- Word, Excel, MS Office

**D-MATL**
- Other
- PDF
- SPSS, SPLUS
- JPG, TIFF, GIF
- Word, Excel, MS Office
- LATEX

**Other:** GIS, ACCESS, FILEMAKER, POSTGRES, ILLUSTRATOR, PHOTOSHOP, InDesign, DREAMWEAVER, MATLAB, FILEMAKER PRO, ASCII....

**Other:** ADOBE ILLUSTRATOR + PHOTOSHOP, MATLAB, ASCII, VMD, NAMD, MATHEMATICA....

11th Sept. 2011
A. Kirstein
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?

**D-ERDW**

- Raw data: 60% (n = 20)
- Statistically evaluated data: 15% (n = 20)
- Aggregate data: 30% (n = 20)
- Standardised data: 15% (n = 20)
- Not interested: 15% (n = 20)
- Other: 15% (n = 20)

**D-MATL**

- Raw data: 75% (n = 17)
- Statistically evaluated data: 45% (n = 17)
- Aggregate data: 45% (n = 17)
- Standardised data: 15% (n = 17)
- Not interested: 15% (n = 17)
- Other: 15% (n = 17)

→ Pilot projects are focused on managing and archiving research data published in papers (status preprint) or data sets.
How long a period do you or your research group have in mind for storing data?

**D-ERDW**

- According to legal requirements
- Don’t know
- Will be specified for each project application
- Long-term archiving, because the research data is part of a long-term survey (>10 years)
- Good scientific practice (10 years)

**D-MATL**

- According to legal requirements
- Don’t know
- Will be specified for each project application
- Long-term archiving, because the research data is part of a long-term survey (>10 years)
- Good scientific practice (10 years)
Who is generally the owner of the research data, or who manages it?

**D-ERDW**
- Other
- Bachelor’s or Master’s students
- Doctoral student
- Research group leaders
- Research project leaders
- Individual scientists themselves

**D-MATL**
- Other
- Bachelor’s or Master’s students
- Doctoral student
- Research group leaders
- Research project leaders
- Individual scientists themselves

% (n = 20)
% (n = 17)
In your opinion, should there be a ‘data management policy’ at ETH?

**E-ERDW**

- **Comment**
  - Yes: [30% (n = 20)]
  - No: [15% (n = 20)]

**D-MATL**

- **Comment**
  - Yes: [50% (n = 17)]
  - No: [20% (n = 17)]

**Some comments:**
- To some extent, yes, but often these policies get out of control, generating more work than they actually save. So, I’d call them “guidelines” and strongly.
- In principle yes, but needs to be flexible enough to not put off any data collector/producer, and to satisfy wide variety of scenarios.
Results from the survey and from interviews

- „None“: Data on a file system and/or on offline media – only group leader can retrieve anything manually
- **Managed on- and offline storage** including conversion to open formats (e.g. doc to rtf or txt) and periodic migration to new media
- **Supported applications** on group level
  Capture data when produced, support handling, analysis and visualisation, but not long term preservation in a narrower sense

→ *There is awareness that data needs to be taken care of*

→ *Preservation must not be mixed up with initiatives for Open Data*
Researchers...

• want to keep **full control** at least of who accesses their data - even though they might theoretically be in favour of Open Data
• need to **re-arrange and select** data prior to ingest, add **documentation** and **legal documents**
• need to **edit metadata and add data** to ongoing series, e.g. annually
• are **interested in support for preservation and quality control** (checklists, feedback on metadata…)
• need to keep certain data for **limited periods** (e.g. 10 to 12 years)
• see **archiving needs** often related to data and materials used for **publications** and want to **persistently reference** them
Rosetta supports important functions with a **clear focus on long-term preservation** according to OAIS.

Rosetta can only support preservation when **adequate staffing and an active preservation management** are in place.

An **active international community** is needed to collect, manage and share information and knowledge, e.g. on formats.

**Flexibility** in data management partly **contradicts** the requirements of a **stable preservation environment**:

Where should be the interface between data management and preservation?
Data production and archiving

Project proposal with plan for data management and archiving

Research project

Measurement Calculation Interpretation + Documentation and Metadata

Prior data management of „active“ data (Group, Inst., Dept.)

Research data management and archival system (Rosetta)

Hierarchical storage environment of ETH Zurich

(Re-)Use

Knowledge portal et al.

Stable reference (e.g. DOI)

Access according to producer’s decision
POSSIBLE WORKFLOWS AND ROSETTA

Data production and handling for current analysis

Pre-ingest, e.g. structuring, re-arranging, selecting

Long-term preservation according to OAIS

ExLibris Rosetta

Manually

(Semi-)automatically

20 11th Sept. 2011 A. Kirstein
POSSIBLE EXTENSION OF ROSETTA’S SCOPE?

Data production and handling for current analysis

- Pre-ingest, e.g. structuring, re-arranging, selecting

Manually

(Semi-)automatically

Long-term preservation according to OAIS

ExLibris Rosetta
Serious risks need to be addressed during data production

- Unclear or frequently changing responsibilities
  - Loss of meta-information
- Missing or incomplete documentation
  - Loss of contextual information
- Haphazard directory and file structures
  - Versioning issues, uncontrolled redundancies

→ Research is particularly prone to these risks due to its dynamic development and the high mobility of staff
Meaningful re-use of research data will rely heavily on contextual information and structural relations.

Exhaustive documentation is required.

There is a need to appraise, select and re-arrange objects prior to ingest and later in time.

Treatment and ingest of research data might have more in common with challenges in administrative archives than with those in typical library collections.

Trying to keep this in mind for synergies in future development.
A (potentially large) number of „exotic“ formats needs to be dealt with, but there might be hope:

- For analyzed data, requirements from data exchange with colleagues or from publishers for submissions can stimulate standardization
- Many formats from self programmed software are in fact ASCII-files. The challenge here is not mainly technical, but careful documentation is required to enable meaningful re-use.
- Future agreements on standards are expected within disciplines due to their specific needs for exchange, preservation and re-use

→ Till then, use of open standards or exchange formats needs to be further advocated
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 administrative records for ETH archives
- Work on specification of **submission application** for library materials (institutional repository)
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
• **Convice the university‘s board** to grant ongoing funding as part of their risk management
Questions?

Remarks?

Andreas Kirstein
Head Media and IT Services
Vice Director
ETH-Bibliothek
Rämistrasse 101
8092 Zurich
+41 44 632 26 74
andreas.kirstein@library.ethz.ch
http://www.library.ethz.ch