Christian Löschen
Center for Information Services and High Performance Computing (ZIH)

- Lessons in Open Science –
Preservation and publication of research data

30.09.2021
Why are we here today

**DFG Guidelines** for handling research data

- “… research data should be archived”
- “… should be made available as soon as possible”

**TUD Guidelines** for handling research data

- “… research data should be archived/published in a data repository”

**EU Open Data Pilot**

- “as open as possible, as closed as necessary”

for further information please see [German only](https://tu-dresden.de/tu-dresden/qualitaetsmanagement/ressourcen/dateien/wisprax/Leitlinien-fuer-den-Umgang-mit-Forschungsdaten-an-der-TU-Dresden.pdf?lang=deo)
Motivation of such guidelines

- Data has a value
- Transparency and trust
- Exploiting the full potential of data once generated
Problems with keeping data

Physical data loss

Sources: Sharyn Morrow @ https://www.flickr.com/photos/sharynmorrow/3717359715 (CC BY-NC-ND 2.0)
Christiaan Colen @ https://www.flickr.com/photos/christiaancolen/20012126873 (CC BY-SA 2.0)
Source: Dave Hill http://flickr.com/photos/dmh650/4031607067
Problems with keeping data

Hardware degradation

Source: http://www.crashplan.com/medialifespan/
Problems with keeping data

Software/Format extinction

Source: https://winworldpc.com/product/microsoft-office/2x

Adobe Flash

Source: https://www.tapestodigital.com (CC BY-NC-ND 2.0)

*Source: https://www.adobe.com/content/dam/acom/images/shared/product_mnemonics/128x128/flash-player-128x128.png
Problems with keeping data

Gradual loss of knowledge about data

Source: Christine Malinowski, MIT Libraries Data Management Services (CC-BY)
Problems with keeping data

Gradual loss of knowledge about data

60% of respondents report that half or more of their organization’s data is dark.

Importance of proper hand over

Your data **without** documentation

![Image of hands holding pills](Designed by Freepik)

Probably noone ever will use it
Where a hand over happens

Source: Projekt Radieschen [https://doi.org/10.2312/RADIESCHEN_005]
Data curation

- Select relevant data
- Restructure and rename if necessary
- Add data documentation

Source: Projekt Radieschen [https://doi.org/10.2312/RADIESCHEN_005]
Data curation

- How to **preserve** and make **available**
- **Legal** requirements

Source: Projekt Radieschen [https://doi.org/10.2312/RADIESCHEN_005](https://doi.org/10.2312/RADIESCHEN_005)
Data curation efforts on domain transfer

- Select relevant data
- Restructure and rename
- Add data documentation
- Preserve and make available
- Legal requirements
Select relevant data

Data hygiene – Clean up unnecessary data without any value
Focus on decreasing the amount of ROT data:

- **Redundant, Obsolete or Trivial**
- e.g. temporary data, doubletes, test data

ROT data causes **avoidable costs**: storage, confusion, time
Restructure and rename

Do the folders represent the expected data structure?
Restructure and rename
Data documentation – for a solid hand over

Your data with documentation

Image: Brett Jordan on https://www.unsplash.com/
Data documentation – **Metadata**

**Data**

- "raw" data is not self-explanatory

**Metadata**

- Data about Data
- Data + Metadata = Information
- Necessary to find and to structure data

*Designed by Freepik*

Image: Brett Jordan on [https://www.unsplash.com/](https://www.unsplash.com/)
Data documentation – **Metadata standards**

Benefits

- Assures a complete, specific and standardised description of your research object
- Enables the organization and classification with similar data records

Finding standards for a particular discipline:

- [http://www.dcc.ac.uk/resources/metadata-standards](http://www.dcc.ac.uk/resources/metadata-standards)
Data documentation – **Adding (textual) documentation**

- **Data**
- **Metadata**
- **Documentation**

- **Data + Metadata + Documentation = Knowledge**
- **Objective:** full **understanding** and **verifiability** of data and research
  - Description of the origin, processing, use and restrictions of the data
- **Enables sharing of data**

Image: Brett Jordan on https://www.unsplash.com/
Preserve and keep available – **Using a data repository**

**Repository:** *Document server* for archiving and/or publishing digital data

- Stores digital assets (i.e. data)
- Allows retrieval of the data
- Is searchable
- Sometimes processes for quality assurance
- Provides standardized interfaces

Preserve data for > 10 years, maintain readability and interpretability

Make data unrestricted accessible to the public

Persiste nt identifie r
Preserve and keep available – **Types of repositories**

<table>
<thead>
<tr>
<th>Research Field</th>
<th>One or few disciplines</th>
<th>Open for all disciplines</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>⚡ Support and processes for discipline-specific data</td>
<td>✅ Open for all disciplines (the long tail)</td>
</tr>
<tr>
<td></td>
<td>⚡ Broad awareness in its particular community</td>
<td>⚡ Often no special support for discipline-specific data</td>
</tr>
<tr>
<td></td>
<td>⚠ Not available for all disciplines</td>
<td></td>
</tr>
</tbody>
</table>
Preserve and keep available – **Types of repositories**

<table>
<thead>
<tr>
<th>Intended „Customer“</th>
<th>One or few selected institutions</th>
<th>Open for all institutions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="https://via.placeholder.com/150" alt="Support contact in-house" /></td>
<td><img src="https://via.placeholder.com/150" alt="Some-times unpleasant terms of use" /></td>
</tr>
<tr>
<td></td>
<td><img src="https://via.placeholder.com/150" alt="Public visibility" /></td>
<td></td>
</tr>
</tbody>
</table>
Preserve and keep available – **Repositories – some examples**

<table>
<thead>
<tr>
<th>Research Field</th>
<th>One or few selected institutions</th>
<th>Open for all institutions</th>
</tr>
</thead>
</table>
| **One or few disciplines** | RODARE  
https://rodare.hzdr.de  
(Rossendorf Data Repository – HZDR) | PANGAEA – [https://www.pangaea.de/](https://www.pangaea.de/)  
(Earth & Environmental Science) |
(Climate and Earth System Research) |
| **Open for all disciplines** | OpARA  
https://opara.zih.tu-dresden.de  
(TU Dresden) | ZENODO – [https://zenodo.org/](https://zenodo.org/)  
DRYAD – [https://datadryad.org/](https://datadryad.org/)  
figshare - [https://figshare.com/](https://figshare.com/) |
Preserve and keep available – **Find your matching repository**

on [https://www.re3data.org/](https://www.re3data.org/)
Preserve and keep available – **Data formats**

Some file formats might cause **problems** in future:

- Software becomes **unavailable**
- Extinction of a file format
- File format not openly documented (aka „proprietary format“)
- Compression
- Encryption

Source: [https://winworldpc.com/product/microsoft-office/2x](https://winworldpc.com/product/microsoft-office/2x)
Preserve and keep available – **Sustainable data formats**

Whenever possible

- use open file formats
- or file formats widely used by research community

Consider migrating data into such a format, in addition to keeping a copy in the original format.

Note: In some cases, migrating data to another format might cause data/metadata loss.

Conversion of problematic data types:

https://documentation.library.ethz.ch/display/RC/Archivtaugliche+Dateiformate#ArchivtauglicheDateiformate-EmpfohleneKonvertierungsmethoden
### Sustainable data formats

<table>
<thead>
<tr>
<th>Data type</th>
<th>Recommended</th>
<th>Limited</th>
<th>Inappropriate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td>PDF/A, TXT, Markdown, Source Code</td>
<td>PDF, RTF, LaTeX, Open Document Formats</td>
<td>DOC, PPT</td>
</tr>
<tr>
<td>Tabular data</td>
<td>CSV, XML</td>
<td>Open Document Formats</td>
<td>XLS</td>
</tr>
<tr>
<td>Raster graphics</td>
<td>TIFF*, PNG, JPEG2000*, DNG</td>
<td>TIFF**, GIF, BMP, JPEG, JPEG2000**</td>
<td></td>
</tr>
<tr>
<td>Vector graphics</td>
<td>SVG</td>
<td></td>
<td>EPS, PSD</td>
</tr>
<tr>
<td>Audio</td>
<td>WAV</td>
<td>MP3, MP4</td>
<td></td>
</tr>
<tr>
<td>Video</td>
<td>FFV1 Codev in MKV</td>
<td>MPEG-2, MP4, MOV, AVI, MJ2</td>
<td>WMV</td>
</tr>
<tr>
<td>Raw</td>
<td>Binary data</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*) uncompressed or lossless compressed  
**) compressed or lossy compressed  
Source: https://documentation.library.ethz.ch/display/RC/Archivtaugliche+Dateiformate
Legal requirements – **For publishing data/Open Access**

**Open Access**

- Free access to scientific literature and materials on the Internet
- Free: **unrestricted access**, not only free-of-charge

**Open Data**

- Grant everyone unrestricted access to (research) data and **allow further use** (under certain conditions)

**Licensing**

- **Required** for Open Access
- Grant a **free license** that clearly **defines and permits** the use, redistribution and modification of copyrighted works.
Legal requirements – **Which license to choose?**

- Recommended for research data: Creative Commons 4.0
- Selectable license modules:
  - **Attribution** of the author
  - **Share-Alike** – Share modifications only under the same license
  - **Non-Derivates** – No sharing of modifications of the original
  - **Non-Commercial** – Forbid commercial use

- More Info: [https://creativecommons.org/choose](https://creativecommons.org/choose)
- Mind licensing conditions of journals! They often require a least restrictive license! (i.e. CC-BY)
- Applying NC-module might lead to unintended results
Legal requirements – **Personal data protection**

**GDPR** - EU General Data Protection Regulation (*German: „DSGVO“*)
- Applies when working with **personal data**, e.g. of study participants

**Get help**
- Data Protection Officer (TUD)
  [https://tu-dresden.de/tu-dresden/organisation/gremien-und-beauftragte/beauftragte/datenschutzbeauftragter](https://tu-dresden.de/tu-dresden/organisation/gremien-und-beauftragte/beauftragte/datenschutzbeauftragter)
- Unabhängige Treuhandstelle (Faculty Of Medicine)
  [https://tu-dresden.de/med/mf/forschung/services-fuer-forschende/unabhaengige-treuhandstelle](https://tu-dresden.de/med/mf/forschung/services-fuer-forschende/unabhaengige-treuhandstelle)
- Anwendungsbereich DSGVO – Interactive Virtual Assistant

**Possible measures**: anonymization, pseudonymization, informed consent
Legal requirements – Possible restrictions when publishing data

- Intellectual property
- Confidential material
Takeaways

1) Select and shape your relevant data
   - Data hygiene

2) Describe your data properly
   - By (standardized) metadata
   - And documentation

3) Use an appropriate repository
   - Preferred a disciplinary one (for publication)
   - Or use your institutional repository

4) Clarify legal matters

5) Document your data in time