TECHNISCHE UNIVERSITÄT DRESDEN

DRESDEN concept

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

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

30.09.2021

ZIH
Center for Information Services & High Performance Computing

Why are we here today

DFG Guidelines for handling research data

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

Deutsche Forschungsgemeinschaft DFG

TECHNISCHE UNIVERSITÄT DRESDEN

TUD Guidelines for handling research data

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

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

---

EU Open Data Pilot

- "as open as possible, as closed as necessary"

European Commission

Motivation of such guidelines

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

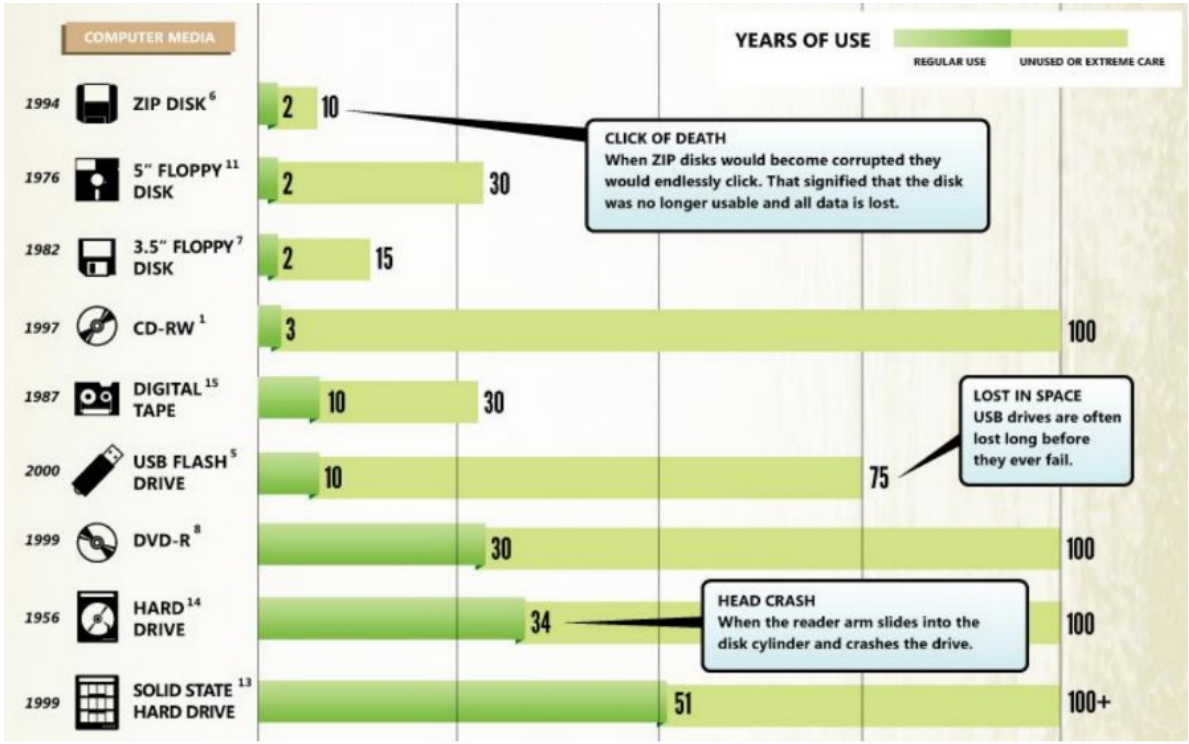
Problems with keeping data

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

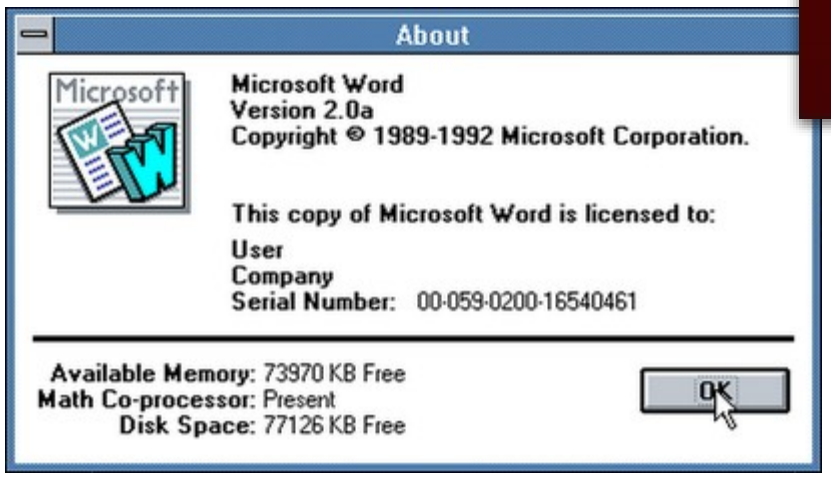
Hardware degradation



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 Serices (CC-BY)

Problems with keeping data

Gradual loss of knowledge about data

Source: https://www.splunk.com/pdfs/dark-data/the-state-of-dark-data-report.pdf (2019)

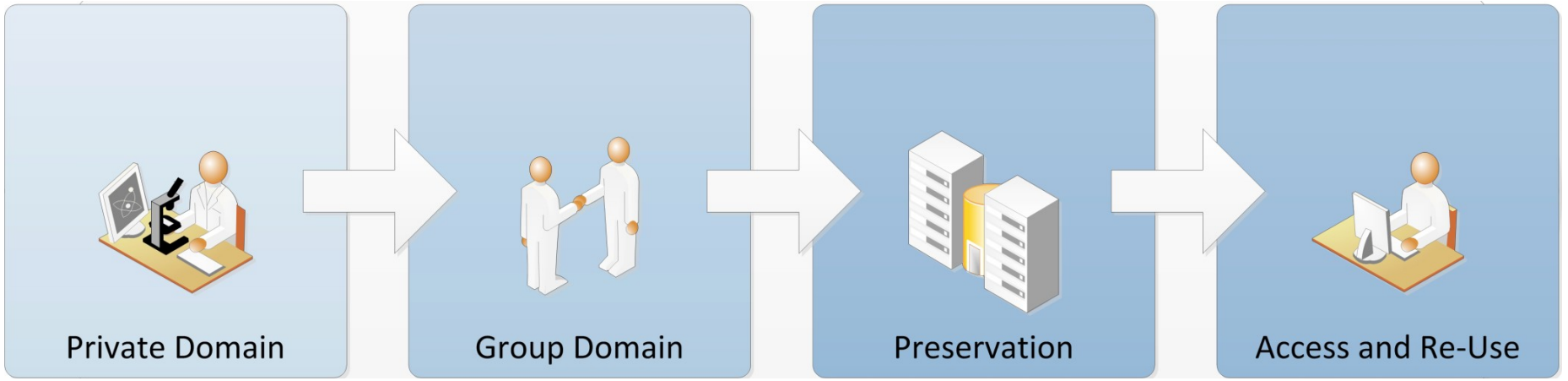
Importance of proper hand over

Your data without documentation

Designed by Freepik

Probably noone ever will use it

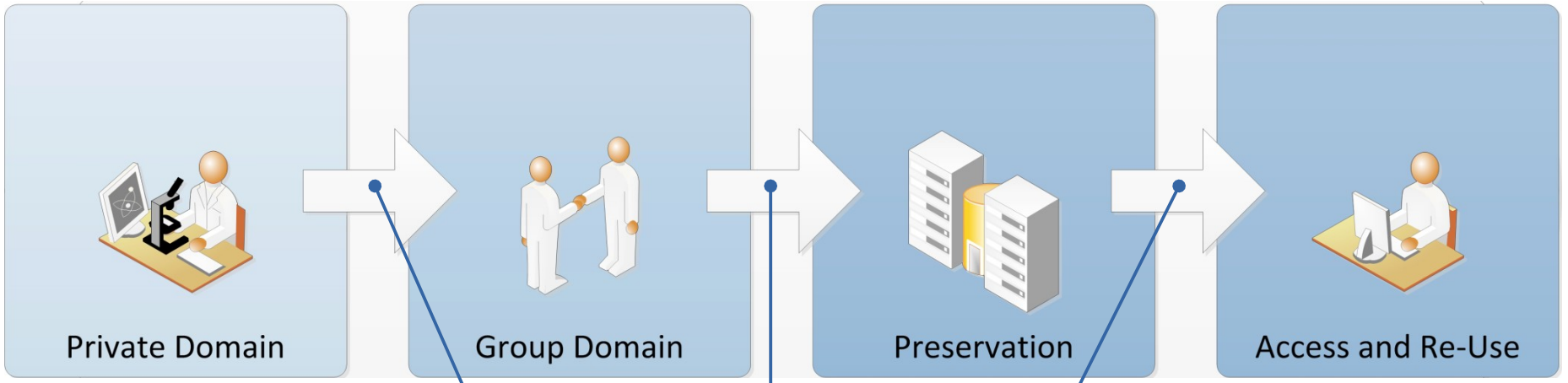
Where a hand over happens



Source: Projekt Radieschen https://doi.org/10.2312/RADIESCHEN_005

Effort of data curation

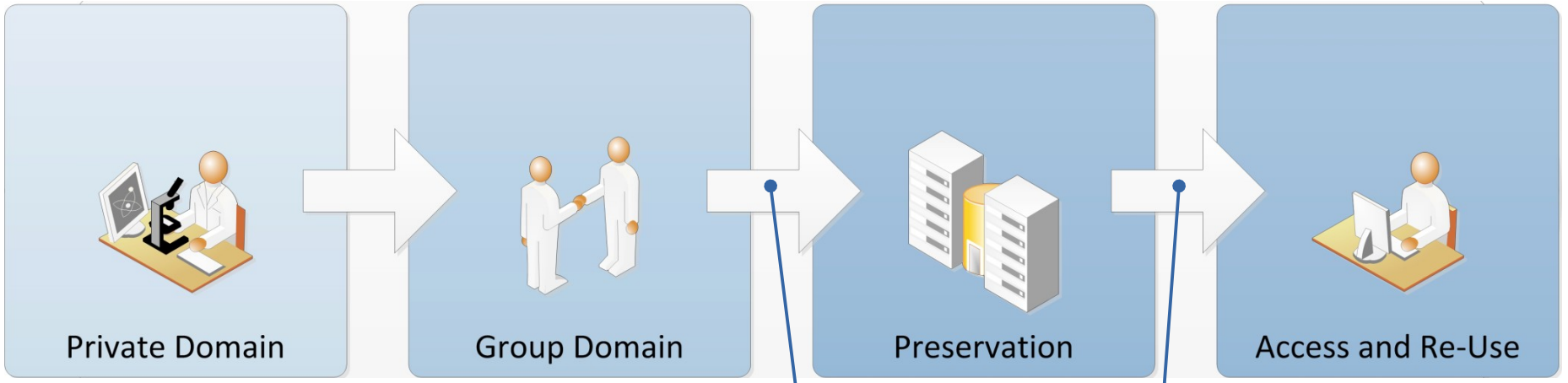
Data curation



Source: Projekt Radieschen https://doi.org/10.2312/RADIESCHEN_005

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

Data curation



Source: Projekt Radieschen https://doi.org/10.2312/RADIESCHEN_005

- How to preserve and make available
- Legal requirements

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

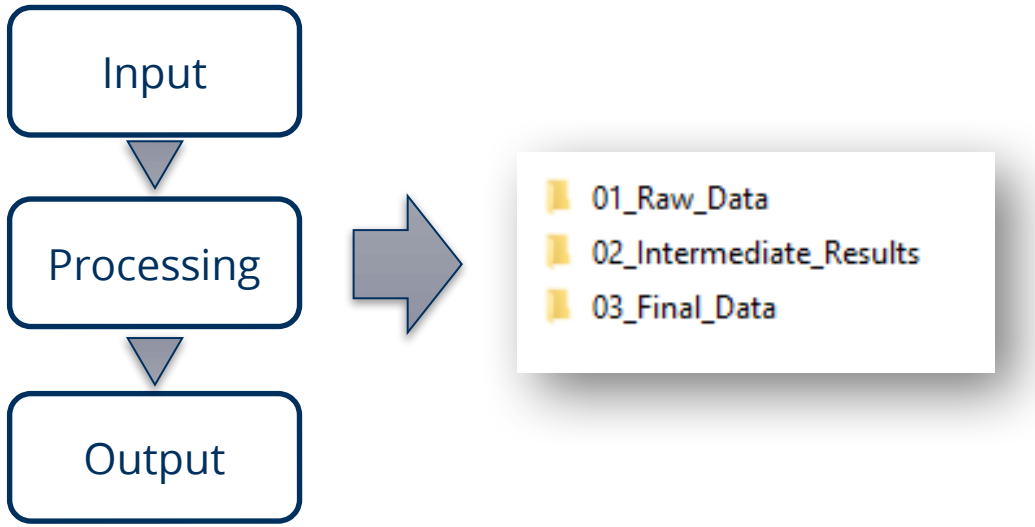
Focus on decreasing the amount of ROT data:

- **R**edundant, **O**bsolete or **T**rivial
- e.g. temporary data, doublettes, test data

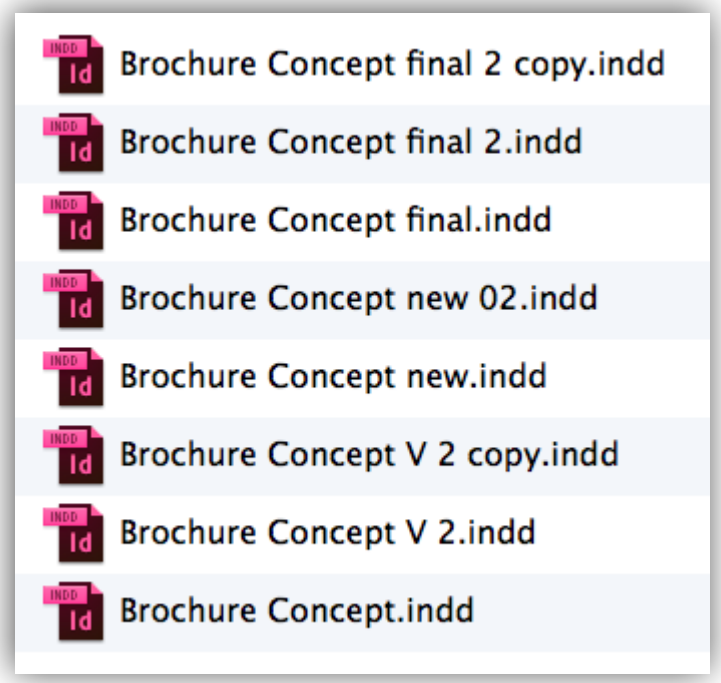
ROT data causes avoidable costs: storage, confusion, time

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

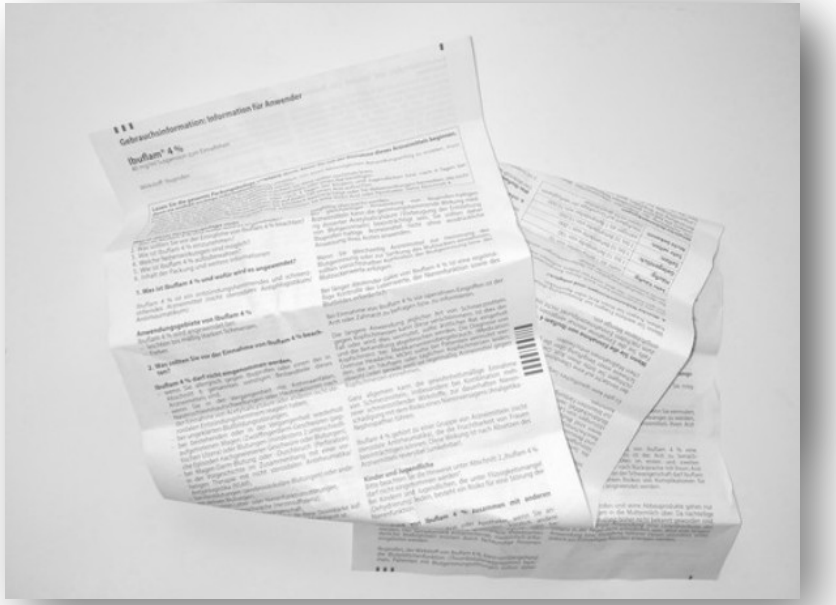


Image: Brett Jordan on https://www.unsplash.com/

Data documentation – **Metadata**

Data

Designed by Freepik

- "raw" data is not self-explanatory

Metadata

Image: Brett Jordan on https://www.unsplash.com/

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

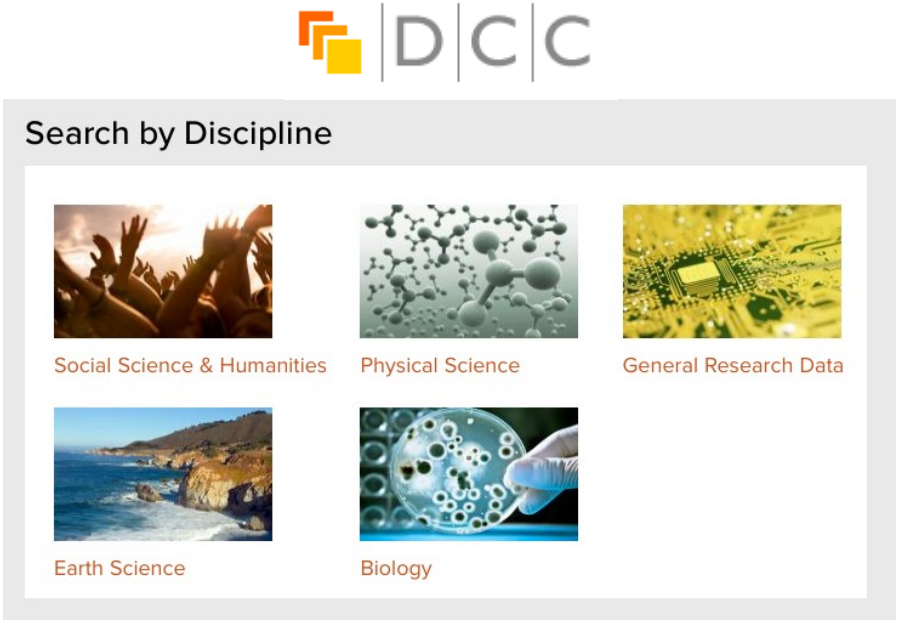
Data documentation – **Metadata standards**

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://rd-alliance.github.io/metadata-directory/standards/



Data documentation – **Adding (textual) documentation**

Data

Designed by Freepik

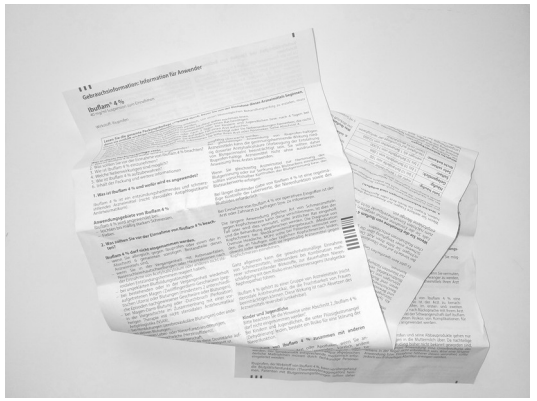
+

Metadata

Image: Brett Jordan on https://www.unsplash.com/

+

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

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

open access

Persistent identifier

Preserve and keep available – **Types of repositories**

Preserve and keep available – **Types of repositories**

Intended „Customer“

One or few selected institutions	Open for all institutions
✓ Support contact in-house — Public visibility	— Some-times unpleasant terms of use

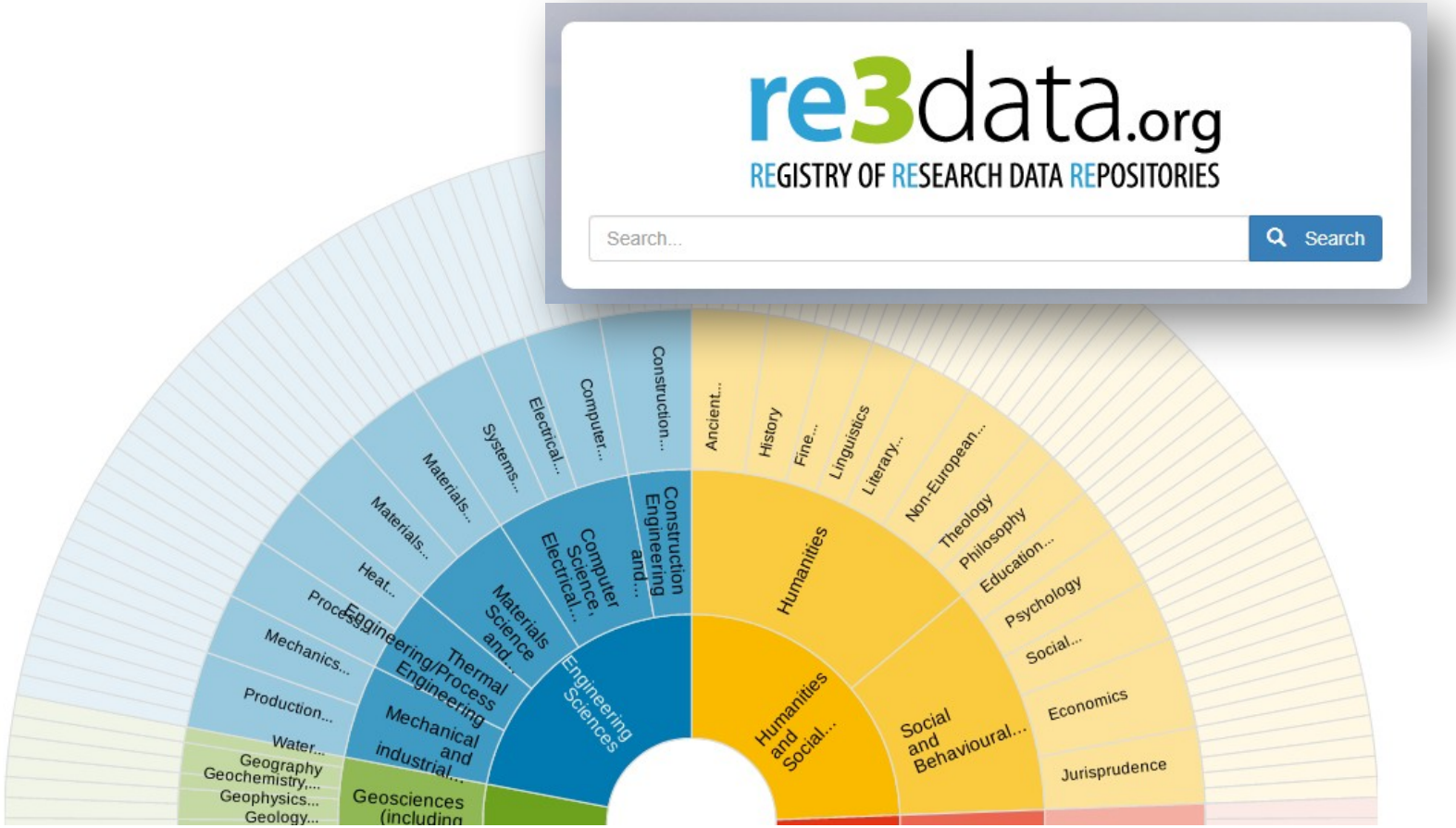
Preserve and keep available – **Repositories – some examples**

Intended „Customer“

Research Field	One or few selected institutions	Open for all institutions
One or few disciplines	RODARE https://rodare.hzdr.de (Rossendorf Data Repository – HZDR)	PANGAEA – https://www.pangaea.de/ (Earth & Environmental Science) CERA – https://cera-www.dkrz.de/ (Climate and Earth System Research)
Open for all disciplines	OpARA https://opara.zih.tu-dresden.de (TU Dresden)	ZENODO – https://zenodo.org/ DRYAD – https://datadryad.org/ figshare - https://figshare.com/

Preserve and keep available – **Find your matching repository**

on https://www.re3data.org/



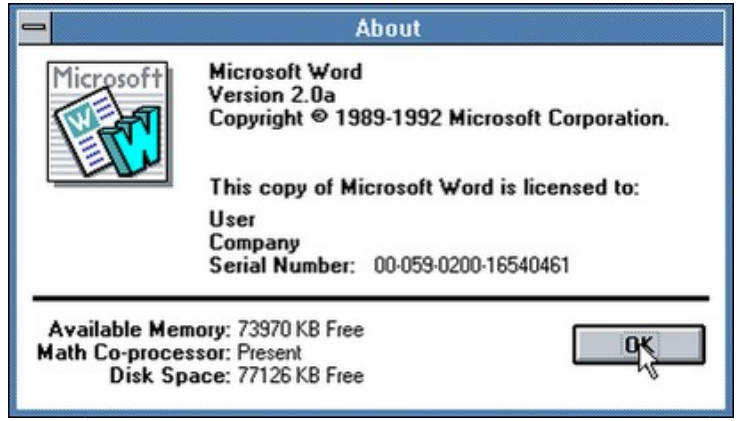
Filter

- Subjects
- Content Types
- Countries
- AID systems
- API
- Certificates
- Data access
- Data access restrictions
- Database access
- Database access restrictions
- Database licenses
- Data licenses
- Data upload
- Data upload restrictions
- Enhanced publication
- Institution responsibility type
- Institution type
- Keywords
- Metadata standards
- PID systems
- Provider types
- Quality management
- Repository languages
- Software
- Syndications
- Repository types
- Versioning

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

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#Archivtaugliche Dateiformate-EmpfohleneKonvertierungsmethoden

Preserve and keep available – **Sustainable data formats**

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

*) 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:

ution of the author

-Alike – Share modifications only under the same license

Derivates – No sharing of modifications of the original

Commercial – Forbid commercial use

- More Info: https://creativecommons.org/choose
- Mind licensing conditions of journals! They often require a least restrictive license! (i.e. CC-BY)
- Appyling NC-module might lead to unintended results

Legal requirements – **Personal data protection**

GDPR - EU General Data Protection Regulation *(German: „DSGVO“)*

- Applies when workign 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
- Unabhängige Treuhandstelle (Faculty Of Medicine)
  https://tu-dresden.de/med/mf/forschung/services-fuer-forschende/unabhaengige-treuhandstelle
- Anwendungsbereich DSGVO – Interactive Virtual Assistent
  https://wiki.bib.uni-mannheim.de/xerte/play.php?template_id=191

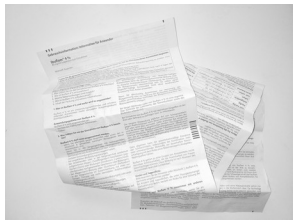
Possible measures: anonymization, pseudonymization, informed consent

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



re3data