Dr. Juliane Jacob

Workshop for sustainable Research Data Management
Topics

- Introduction
- Research Data, Research Data Management, Data Management Plan
- FAIR principles
- Policies and Open Access
- Licences and (Legal Rights)
- Publication of research data
- Q&A
- Mood Query

- Who are you?
  - Your research project? Your data? Your expectations?
  - ...

11.01.2022  |  Juliane Jacob
Center for Sustainable Research Data Management

- Central „operational unit“ according to §93 HmbHG, assignment to VP 3
- Cooperation with RRZ and other data centers
- Our services
  - Consulting, training and information
  - Research Data Repository (UHH-RDR)
  - Research Information System (RIS)
  - Providing of Digital Object Identifier (DOI)
  - Tool and consulting for Data Management Plans (DMP)
  - Curation of (web-) applications / Heurist database toolbox
RDM and Good Scientific Practice

Quality assurance of scientific work!
Guideline for Safeguarding Good Scientific Practice
and how to deal with scientific misconduct.

- What are my tasks?
- What do I have to do personally?
- What am I responsible for?
What do we mean by research data?

All data and materials used in research.

Measurement data, laboratory values, audio, video, texts, surveys, images, objects from collections, samples, algorithms, software, 3D data, scans, simulation data...
## Types of Research Data

<table>
<thead>
<tr>
<th>Metadata</th>
<th>Primary/Raw Data</th>
<th>Secondary Data</th>
<th>Research information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name author</td>
<td>Measurement data, manuscripts, photographs, recordings, digital copies</td>
<td>prepared data (selection, correction, aggregation, ...)</td>
<td>administrative information, e.g.</td>
</tr>
<tr>
<td>Company/model/camera</td>
<td>...</td>
<td>Documentation (methods, laboratory books, test protocols)</td>
<td>employees, running times, financing of projects</td>
</tr>
<tr>
<td>Calibration/setting</td>
<td>...</td>
<td>Procedures (algorithms, software)</td>
<td>Funding applications</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>
What is Metadata?

Metadata is “data about data” and contains information about the context.

<table>
<thead>
<tr>
<th>Metadata type</th>
<th>Target</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive metadata</td>
<td>Identification and</td>
<td>Title, author, abstract, keywords ...</td>
</tr>
<tr>
<td></td>
<td>discoverability</td>
<td></td>
</tr>
<tr>
<td>Administrative metadata</td>
<td>Creation and storage</td>
<td>Rights management and access rights, storage...</td>
</tr>
<tr>
<td></td>
<td>Structure of the resource</td>
<td>Chapters, sections, structure, files...</td>
</tr>
<tr>
<td>Structural metadata</td>
<td>Technical characteristics</td>
<td>Scope, coding, technical structure, formats</td>
</tr>
<tr>
<td>Technical metadata</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Why is metadata important?

- Metadata ensure long-term data usage and interpretation
- Well-documented data can also be understood by people who were not involved in their collection
- Structured metadata can be machine read and processed, making searching quick, easy and reliable
Risk factors for research data

- The departure of those responsible
- Data and backups get "lost"
- Security issues
  - non-updated software versions
  - insufficiently secured accesses (passwords)
- Technological development
  - file formats
  - functionality
- Change of policies or services of the university
  - elimination of service offers
Do you ever have...

- not be able to understand (your) data?
- had problems with opening/reading data?
- not found your own data (on your laptop)?
- lost data?
- described and documented data?
- published data?
- Re-used other data?
- checked results of third-party scientific work?
Research Data in the Focus of Science Organisations

HRK  German Rectors’ Conference
The Voice of the Universities
Management of research data – a key strategic challenge for university management

Making information resources more usable
A position paper on the management of research data in the Helmholtz Association¹

DFG Guidelines on the Handling of Research Data
Research data is an essential foundation for scientific work. The diversity of this data reflects the wide range of different scientific disciplines, research interests and research methods. Research data might include measurement data, laboratory values, audiovisual information, texts, survey data, objects from

¹ Helmholtz Association of German Research Centres
Definition of Research Data Management

RDM is a term that describes the organisation and control of the work processes involved in generating and managing research data as efficiently as possible. In this way, RDM accompanies research from the initial planning stage through to archiving, subsequent use or deletion of the data.
Sustainable RDM

- Makes data usable, instead of just using
- Prevent data loss (Data is independent from people)
- Enable (semi-) automatic processing (Metadata)
- Long-term understanding of results
  (Documentation!)
- Using data in the long term
- Optimal use of resources (through potential reuse)
- Fulfil the requirements of money/third-party funding providers
F indable
A ccessible
I nteroperable
R eusable
Findable

- Description through metadata (machine readable)
- Visibility (MD registration in detection systems and catalogues)
- Citation capability with the aid of persistent identification (PIIDs)
What are Persistent Identifiers?

- PIDS have digits and/or alphanumeric characters, which are assigned to a digital object and refers directly to it.

- For research data:
  - DOI (Digital Object Identifier), URN (Uniform Resource Name), Handle, ORCID

- URLs are problematic because they do not refer to content, but to a location.

- Ideally, nothing changes in the PID when the location of the digital object changes.

- If the object is deleted, a landing page with the corresponding metadata is retained.

PIDS ensure that a digital object can be permanently indexed, retrieved and thus quoted.
Accessible

- RD are accessible via their PIDs using standardised, open protocols (HTTP(S), FTP, SMTP)
- Where necessary, the protocol supports authentication and rights management
- Secure long-term archiving
- Metadata remain available even if data have been deleted
Interoperable

- Allow data exchange between researchers, organisations etc (adhering to format standards)
- Standardised, machine-readable metadata
- Controlled vocabularies, value ranges, thesauri and ontologies
- Standardised references to other MD/ RD (e.g. 'is new version of', 'is supplement to', 'relates to',
- Use of open, documented formats
Reusable

- Precise, standardised, detailed description
- Unambiguous, accessible user licenses
- Clear indication of origin (provenance information)
- Research data and metadata comply with the relevant standards for the field
Data Management Plan
Do you practise data management already?

- What tools do you and your colleagues use?
- How did you come in contact with data management plans?
- Are there any instructions from supervisors?
Data Management Plan – Definition

A data management plan (DMP) is a document describing the life cycle of data (in particular research data) from collection to archiving, including all measures taken to ensure that the data remains available, usable and traceable.
Breakout Session – DMP without assistance

- In the group, create a data management plan for a (fictitious) project
- Use the padlet
- Consider and note for the individual phases of your project, which steps and tools are necessary in data management (10min)
- Introduce your mind map briefly to the other workshop participants (5min)
Elements of a DMP
Elements of a DMP

1. Project description
   - General information on the project such as objectives, promoter and duration, relevant policies
Elements of a DMP

2. Existing types of data
   - Description of existing data that can be reused for the project and how their integration could look
   - Reasons, if no data is reused
Elements of a DMP

3. Types of data to be generated
   - Information on data types and formats
   - Estimated data volume
   - Information on the process of data generation and quality assurance (e.g. multiple sampling, validation, data peer review)
   - Relevance of reuse by others
Elements of a DMP

4. Data organisation
   - Details on consistent data organization within the project (e.g. for data storage, data naming, synchronization, versioning, and other collaboration workflow, documentation, creation of metadata)
Elements of a DMP

5. Administrative and legal aspects description
   - Funding and legal requirements
   - Copyright/Data Owner
   - Access and use
   - Data protection
   - Data security and backup
Elements of a DMP

6. Archiving, sharing, and publishing data
   - Which data will be shared?
   - How will data be shared?
   - Information on the planned interoperability with external discipline-specific data services.
Elements of a DMP

7. Responsibilities and duties
How are the responsibilities for data management defined and distributed within the project?
Elements of a DMP

8. Costs and Resources

- Reports on the costs and the personnel expenditure for maintaining the data management plan and running costs for data curation, production of metadata, archiving, etc.
Data management plans according to RDMO

- Optimisation of research data management before project start
- Guidelines for handling research data over the entire course of the project
- Planning the whereabouts of the research data after the end of the project
  - Publication
  - Long-term archiving
- Increasing the quality and efficiency of scientific work
- More than just a request from the funding agency
- In the ideal case, profit for the researcher and the institution
Breakout Session – DMP with RDMO

Research Data Management Organiser is a tool for creating data management plans with various templates [https://dmp.fdm.uni-hamburg.de/](https://dmp.fdm.uni-hamburg.de/)

- Create a DMP (for a project) in your breakout session
- Answer questions of the RDMO catalogue
- Add your group members to the project as members
- Distribute tasks
- Export your DMP
**Answers for DDLitLab**

In the following, we have summarized the information about the project as given by you and your collaborators.

### General

**Topic**

What is the main research question of the project?  
Please give some keywords describing the research question.

**Research field**

Which research field(s) does this project belong to?  
What is the main research question of the project?  
Please give some keywords describing the research question.

**Project schedule**

When does the project start?  
When does the project end?

### Options

Back to project overview

### Export

PDF  
Rich Text Format  
Open Office  
Microsoft Office  
HTML  
Markdown  
mediawiki  
LaTeX

**Navigation**

Please note that using the navigation will discard any unsaved input.

General  
→ Topic  
Research field  
Project schedule  
Project coordination  
Project partners  
Funding  
Other requirements I  
Other requirements II  
Content classification  
Technical classification  
Data usage  
Metadata and referencing  
Legal and ethics  
Storage and long-term preservation

11.01.2022  |  Juliane Jacob
# Comparison of requirements of Research Funders (Date 10/22/2021)

<table>
<thead>
<tr>
<th>Funder</th>
<th>Plan necessary?</th>
<th>Delivery with application?</th>
<th>Content</th>
<th>Updates?</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC Horizon Europe</td>
<td>DMP</td>
<td>DMP within the first 6 project month</td>
<td>Content of Horizon Europe Template</td>
<td>Updates, if significant changes and at the end of the project</td>
</tr>
<tr>
<td>DFG</td>
<td>Information on handling with RD</td>
<td>Yes</td>
<td>Content of the guidelines for dealing with RD</td>
<td>No</td>
</tr>
<tr>
<td>BMBF</td>
<td>DMP depends on the program</td>
<td>Yes, if necessary</td>
<td>Content depends on program</td>
<td>Depends on the program</td>
</tr>
<tr>
<td>VolkswagenStiftung</td>
<td>DMP</td>
<td>Yes</td>
<td>Content of the Science Europe Template</td>
<td>No</td>
</tr>
</tbody>
</table>
Further requirements: Policies...

Internal guidelines or guidelines that are formally established by the company and via its management.

- UHH: Statutes to ensure good scientific practice and to avoid scientific misconduct at the UHH
- UHH Open Access Policy for data and publications
Interdisciplinary Policies

”In the interest of transparency and to enable research to be referred to and reused by others, whenever possible researchers make the research data and principal materials on which a publication is based available in recognised archives and repositories in accordance with the FAIR principles accessible in recognized archives and repositories

”[...] the research data (generally raw data) on which they are based are generally archived in an accessible and identifiable manner for a period of ten years at the institution where the data were produced or in cross-location repositories.”

target group: Applicant, obligatory (no funding without policy or DMP)
Interdisciplinary Policies

ERC (2016): Guidelines on FAIR Data Management in Horizon 2020

“In general terms, your research data should be 'FAIR', that is findable, accessible, interoperable and re-usable.”

“as open as possible, as closed as necessary”

target group: Applicant, compulsory (no funding without policy, DMP at the beginning and in cause of changes during the project), very specific suggestions for implementation.
Disciplinary Policies: Good Laboratory Practice (GLP)

OECD (1997): OECD Principles of Good Laboratory Practice

“...ensure that after completion (including termination) of the study, the study plan, the final report, raw data and supporting material are archived.”

“Any change in the raw data should be made so as not to obscure the previous entry, should indicate the reason for change and should be dated and signed or initialled by the individual making the change.”

target group: Research laboratories; Required by law in many countries for non-clinical experimental testing of substances (REACH)
Disciplinary Policies

ZBW Open Access Policy (2020)

“The ZBW strongly recommends the publication of its own research results in digital, freely accessible form... Directly via OA media or in parallel or time-shifted via the local repository infrastructure.”

“Quality and reputation are important for the publication medium.”

“ZBW recommends that its authors use standardized, open licenses for their open access publications.”

Voluntary commitment, coordinated with research funders and demanded by them.
Journal Policies

E.g. Nature, Policy from 2017

“Supporting data must be made available to editors and peer reviewers at the time of submission for the purposes of evaluating the manuscript. All manuscripts reporting original research published in Nature journals must include a data availability statement. Data availability statements should provide a statement about the availability of the minimal dataset that would be necessary to interpret, replicate and build upon the methods or findings reported in the article. ... The preferred way to share large data sets is via public repositories. Details about how to share some specific data sets can be found in the sections below.”

Compulsory for everyone who wants to publish an article in the Nature Journals.
Institutional Policies of UHH

UHH: Statutes to ensure good scientific practice and to avoid scientific misconduct at the UHH

"Primary data as the basis for publications are to be stored on durable and secure media in the institution in which they were created for ten years, unless special regulations provide for a longer retention period."

DFG guidelines for non-DFG projects at UHH. Considers all other aspects. Support from an office, an ombuds committee and a committee of experts.
Open Access Policies of UHH

Publications

For scientific publications, the University of Hamburg recommends a direct open access publication ("gold OA"), either through a fee to be paid to the publisher or free of charge on the “Institutional Repository” of the University of Hamburg.“

"In the case of publications in publishers with costs, a non-exclusive right of use should be negotiated in order to enable an immediate second publication “green OA", "self-archiving").
Open Access Policies of UHH

Data

“The University of Hamburg is committed to the permanent preservation and reliable provision and utilization of publication data ("Open Access" in the narrower sense), research data ("Open Data") and teaching and learning data ("Open Educational Resources") (...). The University of Hamburg provides the services and infrastructures necessary to handle scientific data (...).”
Legal position on research data

(Peter Brettschneider, DOI 10.5281/zenodo.5243232)

**Data**
- If applicable, factory security / ancillary copyrights
- But: Not all research data is protected by copyright!
- Possibly data protection

**Database**
- Rights to the database work (original conception)
- Database manufacturer right (substantial investment)

**Data Storage Medium**
- Ownership

**Software**
- factory security
Legal obstacles vs. Open Data

1. General Data Protection Regulation (GDPR)
   - The respective state data protection laws and the GDPR (not the BDSG!) apply to universities and other public bodies in the federal states
   - only for personal data
   - Processing of personal data: "Prohibition with reservation of permission" Principles of purpose limitation and data minimization

GDPR: "‘personal data’ means any information relating to an identified or identifiable natural person (‘data subject’); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person."
Authorization for publication

- Consent has been given by the persons and meets the requirements (GDPR/DSGVO):
  - Voluntary consent
  - "informed consent"
  - The purpose and type of processing of the data must be clearly defined
- Publication is essential for the "presentation of research results on events in contemporary history" (§ 11 Abs. 3 HmbDSG)
Information form and declaration of consent

3 Beispielformulierungen für Aufklärungsformular und Einwilligungserklärung

3.1 Aufklärungsformular

„Das Forschungsvorhaben .......... hat zum Ziel ............ Ihre Aufgabe besteht darin, .......... Dabei können mitunter Situationen entstehen, die Sie als weniger angenehm empfinden...


Verantwortlich für die Analyse und Speicherung Ihrer Daten ist .........., Institut .........., Universität Hamburg.

Die Teilnahme an dem Forschungsvorhaben erfolgt auf gänzlich freiwilliger Basis. Sie können jederzeit und ohne Angabe von Gründen Ihr Einverständnis zur Teilnahme zurücknehmen, ohne dass Ihnen hieraus Nachteile entstehen.“

3.2 Einwilligungserklärung


11.01.2022 | Juliane Jacob
Legal obstacles vs. Open Data

2. Copyright

- Who does data belong to
  - There is no ownership of data
  - Only the form of the representation is worthy of copyright protection
  - Protection requirement for databases is a "substantial investment"
  - The author is the person who made this investment (employer (UHH) or researcher?)
  - "University professor privilege" (attention to third-party funders)
  - "Enriched data sets" (e.g. images, diagrams, texts): Copyright applies
Legal obstacles vs. Open Data

2. Copyright - Usage Rights

- The author is initially entitled to all copyrights/ancillary copyrights (except software)
- A database that meets the criterion of substantial investment is protected against “substantial parts” of the data contained herein being removed (for example copied) from third parties without the permission of third parties
- joint copyrights with co-authors
- (Exclusive) rights of use can be (verbally) transferred by contract (employer, publisher)
- "Second publication right", preprints
- "Limitation provisions" (§§ 44a–63a UrhG) e.g. right of quotation
- Rights to databases expire 15 years after their publication or production (§ 87d UrhG)
Legal obstacles vs. Open Data

Data protection (security-relevant data, species protection, protection of cultural assets)

- The principle of data economy applies
- ethical aspects / sensitive data: if in doubt, consult the ethics committee
Anonymization vs. Pseudonymization

- Through anonymization, personal data can be turned into data without personal reference
- Pseudonymisation: reference to the person whose data is affected can be restored (e.g. using a separately stored key)
- By deleting the key, anonymization can take place
- Pseudonymised data are still personal data
Authorization for publication

1. Consent has been given by the persons and meets the requirements (General Data Protection Regulation GDPR/DSGVO):
   - Voluntary consent
   - "informed consent"
   - The purpose and type of processing of the data must be clearly

2. Publication is essential for the "presentation of research results on events in contemporary history" (§ 11 Abs. 3 HmbDSG)
Contact Person Legal Rights

- UHH data protection: datenschutz@uni-hamburg.de
- Multimedia Kontor Hamburg https://www.hh-datenschutz.de/
What is a license?

Nutzungsbedingungen

3. Rechtliche Rahmenbedingungen

Datengebende räumen GESIS mit der Übermittlung das einfache Nutzungsrecht an den in SowiDataNet | datorium eingestellten digitalen Datensätzen und Dokumenten ein, insbesondere

1. das Recht diese Datensätze und Dokumente systematisch zu archivieren und für den Zweck der langfristigen digitalen Sicherung und der Veröffentlichung zu verändern. Dabei kann GESIS alle zweckdienlichen technischen Mittel, Formate und Methoden anwenden,
2. sowie das Recht, Datensätze und Dokumente im Rahmen des Online-Angebots von GESIS auch zum Download, dauerhaft öffentlich zugänglich zu machen.

Die Urheberrechte der Datengebenden bleiben dabei gewahrt, so können Daten und Dokumente bspw. weiterhin auch an andere Institutionen zur Archivierung oder Veröffentlichung übergeben werden.


GESIS behält sich vor, die Aufnahme von Daten und sonstigen Materialien abzulehnen oder bereits aufgenommene aus dem Angebot zu entfernen, sollten Zweifel an der rechtlichen Zulässigkeit einer Archivierung bestehen.
What is a license?

- Licenses regulate the conditions for the re-use of published data
- So-called open licenses are available to everyone for free use and only need to be linked
- The prerequisite is that you own the rights
- Selection of the license depending on the type of data:
  - e.g. Creative Commons (CC) - licenses for articles, monographs, images etc.
  - Open Database License (ODbL) for databases, or CC from version 4
  - General Public License (GNU) for software
- If no license is granted, the stricter copyright applies
What is a license?

**Commercial Licenses:**
- Permits use for a fee
- Economic interests, e.g. publishers
  - Examples: Data delivery by a newspaper publisher for a fee

**Free Licenses:**
- Allow free use
- Maximize reuse and visibility
- Examples: CC licenses, ODC licenses
Not licensed material

• Non-licensed material is not freely available!
• Ask the author to what extent the work can be reused
• Search for licensed material
Basic vs. Exclusive Use Rights

**Basic Use Rights**

- Entitled to use something in the permitted way
- Use by others not excluded
- Example: many image databases, CC licenses

**Exclusive Use Rights**

- Use to the exclusion of all others
- Exclusive right, one-time payment
- Example: Transfer of exclusive rights to a publisher
Example Pixabay

• Until 2018 mostly freely available images
• Switched to Pixabay license in January 2019
• Lead to considerable legal uncertainty
  • Images uploaded before 2019: CC0
  • Images uploaded since 2019: Pixabay License
  • However, the Pixabay license is always shown, pay attention to the upload date
  • it is unclear what "distribution" means
• Recommendation: always name the source and respective license, even if the license does not require it
Properties of “good” Licenses

• Standardized license text
• No Transfer of Exclusive Rights
• Distributed internationally and compatible with a large number of jurisdictions
• Machine readable

Recommendation

https://creativecommons.org/
Creative Commons Licenses

- Attribution
- Attribution-Share Alike
- Attribution-Non Commercial
- Attribution-Non Commercial-Share Alike
- Attribution-No Derivs
- Attribution-Non Commercial-No Derivs

Only CC-BY und CC0 are conform to the FAIR-principles!
As open as possible, as closed as necessary.
**Which license fits best?**

<table>
<thead>
<tr>
<th>Dürfen Bearbeitungen Ihres Werkes geteilt werden?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ja</td>
</tr>
<tr>
<td>Nur unter der gleichen Lizenz</td>
</tr>
<tr>
<td>nein</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Darf Ihr Werk kommerziell genutzt werden?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ja</td>
</tr>
<tr>
<td>nein</td>
</tr>
</tbody>
</table>

- **Ja**: CC BY
- **nein**: CC BY NC

**CC BY SA**

**CC BY ND**

**CC BY NC SA**

**CC BY NC ND**

Brettschneider (2020)
Creative Commons Licenses

Title: “Furggelen afterglow”
Author: “Lukas Schlagenhauf”—with a link to their profile page
Source: “Furggelen afterglow”—with a link to the original photo on Flickr
License: “CC BY-ND 2.0”—with a link to the license deed

“Furggelen afterglow” by Lukas Schlagenhauf is licensed under CC BY-ND 2.0.
**Public Domain**

- It is not possible to waive all copyrights in Germany
- Which is possible:
  - Unconditional license or waiver of law
  - “no rights reserved”
  - CC0
- Recommended for data that is not protected by copyright
  - Acknowledge that data is in the public domain
  - There are no usage rights to be granted (such as with CC-BY)
  - Public Domain Mark (no license)

→ Good scientific practice requires citing data (attributing names)
Summary

- Licenses make it easier to pass on usage rights
- Generally:
  - Use licenses designed for the subject matter of the license (no software license for text or data)
- Appropriate licenses for data:
  - Creative Commons (from version 4.0 for databases)
  - Open Data Commons (less known)
- Overview of possible licenses:
  - [https://opendefinition.org/licenses/](https://opendefinition.org/licenses/)
Breakout Session

- [https://chooser-beta.creativecommons.org/](https://chooser-beta.creativecommons.org/)
Reuse

- Six standard license agreements
- Not all licenses can be combined when re-using/editing multiple works
- The general rule is that the more restrictive license is passed on

https://creativecommons.org/about/cclicenses/
Poll

Is the combination possible? Yes/No

1. [CC BY] + [CC BY-SA] = ?

2. [CC BY-SA] + [CC BY-NC-SA] = ?

3. [CC BY] + [CC BY-ND] = ?
Result Poll

Explanation

1. cc BY + cc BY SA

2. cc BY SA + cc BY NC SA

3. cc BY + cc BY ND
### Possible combinations

<table>
<thead>
<tr>
<th></th>
<th>PUBLIC DOMAIN</th>
<th>PUBLIC DOMAIN</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PUBLIC DOMAIN</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>×</td>
<td></td>
</tr>
<tr>
<td></td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>×</td>
<td></td>
</tr>
<tr>
<td></td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>×</td>
<td>×</td>
<td></td>
</tr>
<tr>
<td></td>
<td>✔️</td>
<td>✔️</td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td></td>
</tr>
</tbody>
</table>
The life cycle of data: Which data should I keep?

Decision should depend on potential reuse:

1. Further publication → referenced (=processed) data with additional documentation
2. Verification → referenced data including analysis steps
3. Further analyses → all original data plus software used to collect it
4. Learning and Teaching → samples of original data and compiled data including analysis steps

→ Research data must remain analysable and interpretable
Which data should I keep?

Research data must remain analysable and interpretable

- Preservation of presentation and analysis environments
- Documentation of annotation and analysis processes
What can be removed?

- Empty data records / tables
- Duplicates
- Incomplete data records
- Versions (which?)

Brecht, Bertoldq | 1 Brecht, Bertolt
Brecht, Bertold. | 1 Brecht, Bertolt
Brecht, Bertold | 105 Brecht, Bertolt
Brecht, bertold | 1 Brecht, Bertolt
Brecht, Bertholdt | 32 Brecht, Bertolt
Brecht, Berthold | 14 Brecht, Bertolt
Brecht, Bert | 1 Brecht, Bertolt
Brecht Berthold | 1 Brecht, Bertolt
Bertolt, Brecht | 1 Brecht, Bertolt
Bertold, Brecht | 2 Brecht, Bertolt
Brecher, Gusatv | 1 Brecher, Gustav
Borowsky, Alexander | 10 Borovsky, Alexander
Bondy, Francois | 3 Bondy, François
Boeres, Emil | 6 Boeres, Emile
Bodanzky, Arthur | 5 Bodanzky, Artur
Blunck, Friedrich | 2 Blunck, Hans Friedrich
Blech, Leo, Kreisler, Fritz | 1 Blech, Leo
Leo Blech | 3 Blech, Leo
Blech. Leo | 2 Blech, Leo
Blech, Leo. | 1 Blech, Leo
Blech, Foto | 1 Blech, Leo
Blech Leo | 1 Blech, Leo
Challenges – Data cleanup

- Redundancy (duplication of content / data structures / code)
- Gaps (incomplete records)
- Inconsistencies in the data (e.g. transliteration, coding, abbreviations)
- Lack of transparency (documentation / interpretability of values)
- Unclear origin (expertise of the data creator unclear / missing versioning)
„Poor data quality can be worse than missing data because it can waste resources and lead to faulty ideas and solutions, or at minimum challenges trust in the results and implications drawn from the data. Improvement in data quality can thus have significant benefits.”

(Curating for Quality P5)
### Breakout-Session

<table>
<thead>
<tr>
<th>Provenance / Documentation</th>
<th>Rate the data quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the data about?</td>
<td>Are the data interpretable?</td>
</tr>
<tr>
<td>Where does the data come from?</td>
<td>Is the data record complete?</td>
</tr>
<tr>
<td>When were they collected?</td>
<td>Are the data consistent?</td>
</tr>
<tr>
<td>In what format is the data?</td>
<td>Were any gaps marked?</td>
</tr>
<tr>
<td>Have changes been versioned?</td>
<td></td>
</tr>
</tbody>
</table>
Challenges

- Redundancy (duplication of content / data structures / code)
- Gaps (incomplete records)
- Inconsistencies in the data (e.g. transliteration, coding, abbreviations)
- Lack of transparency (documentation / interpretability of values)
- Unclear origin (expertise of the data creator unclear / missing versioning)
Re-use of data: requirements

- The quality of the data must be sufficient for the research project
- Provenance and rights must be clarified
- Completeness must be sufficient for the project goal
- Interpretability (documentation available?)
Data Documentation

- Context: project history, intention / objective, hypotheses ...
- Methods: sampling, circumstances of the survey, technical framework ...
- Data structures, relationships between objects
- Value ranges, quality criteria, validity
- Changes (!) In the course of the project, versioning
- Legal questions / access

→ Publications do not replace the documentation of the data
Technical Documentation

- Names, designations for variables and their values
- Explanations for codes / classification schemes
- Explanation of the terminology used, possibly definitions
- Coding of missing values / reasons for missing values
- Derived data, algorithms used, weightings
Saving data: file formats

- Is the way, in which information is digitally stored
- File extensions (e.g. .txt, .docx, .jpg) provide information on file formats and thus the data they contain

Win and Mac systems do not show file extensions by default. The display can be activated via the system settings.
Proprietary file formats

- are mostly dependent on
  - specific software
  - corresponding software licenses
  - specific platforms / operating systems
  - are at risk from ‘software obsolescence’
  - rapid technological development
  - Compatibility between versions not guaranteed
  - Market dependency
  - Conversion / export to open formats mostly lossy
File formats - open formats

- For long-term archiving, files
- be unencrypted
- not be compressed
- not proprietary / patented
- use an open / documented standard
# Table file formats

<table>
<thead>
<tr>
<th>File Formats</th>
<th>Recommended</th>
<th>Avoid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tabular data</td>
<td>Csv, tsv, spss portable</td>
<td>Excel</td>
</tr>
<tr>
<td>Text</td>
<td>Text</td>
<td>TXT, ODT, HTML, RTF, PDF/A only if layouts is important</td>
</tr>
<tr>
<td>Multimedia</td>
<td>Container: MP4, Ogg Codec: Theora, Dirac, FLAC</td>
<td>QuickTime, H264</td>
</tr>
<tr>
<td>Pictures</td>
<td>TIFF, JPEG2000, PNG</td>
<td>GIF, JPG</td>
</tr>
<tr>
<td>Structured data</td>
<td>XML, RDF, JSON</td>
<td>RDBMS</td>
</tr>
</tbody>
</table>
Data formats - recommendations

- Multimedia recordings
  - High quality
  - No lossy compression

- Texts
  - Unicode based coding
  - Open and documented formats
  - De facto standards (UTF-8, XML, TEI, JSON-LD, RDF ...)

- General recommendations
  - No proprietary or exotic formats
  - High quality
  - Compression: choose common formats (.zip, .tar.gz, .tar.bz2)
Example CSV

<table>
<thead>
<tr>
<th>ID</th>
<th>Sammlungsname</th>
<th>Band</th>
<th>Röhrchen</th>
<th>Familie</th>
<th>Name lateinisch und Autor</th>
<th>Name deutsch</th>
<th>Name englisch</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bredemann &amp; Nieser, 01, 01</td>
<td>Fabaceae</td>
<td>Anthyllis vulneraria L.</td>
<td>Wundklee, woundwort</td>
<td>0001-Anthyll:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Bredemann &amp; Nieser, 01, 02</td>
<td>Fabaceae</td>
<td>Astragalus boeticus L.</td>
<td>Kaffee-Traganth, swedish coffee</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Bredemann &amp; Nieser, 01, 03</td>
<td>Fabaceae</td>
<td>Astragalus falcatus Lam.</td>
<td>Sicheltraganth, russian milk vetch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Bredemann &amp; Nieser, 01, 04</td>
<td>Fabaceae</td>
<td>Astragalus glycyphyllos L.</td>
<td>Süßer Traganth, sweet milk vetch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Bredemann &amp; Nieser, 01, 05</td>
<td>Fabaceae</td>
<td>Astragalus hamosus L.</td>
<td>Angelstragel, yellow milk vetch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Bredemann &amp; Nieser, 01, 05a</td>
<td>Fabaceae</td>
<td>Cajanus cajan (L.) Millsp.</td>
<td>Taubenerbs, pigeonpea, 0006-7, Bredemann &amp; Nieser, 01, 06</td>
<td>Fabaceae, Senna tora (L.) Roxb., Gemüse-Kassie, sicklepod, 0007-Sen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Bredemann &amp; Nieser, 01, 07</td>
<td>Fabaceae</td>
<td>Cicer arietinum L.</td>
<td>Kichererbs, chickpea, 0008-Cicer-arie</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Bredemann &amp; Nieser, 01, 08</td>
<td>Fabaceae</td>
<td>Corinilla scrophoides (L.) W.D.J.Koch, Skorpionspiech</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Bredemann &amp; Nieser, 01, 09</td>
<td>Fabaceae</td>
<td>Securigera varia (L.) Lassen, Bunte Kronwicke, Crown-vetch</td>
<td>0010-Twic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Bredemann &amp; Nieser, 01, 10</td>
<td>Fabaceae</td>
<td>Cyamopsis tetragonoloba (L.) Taub., Guar, Guar</td>
<td>0011-Cyar</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Bredemann &amp; Nieser, 01, 10a</td>
<td>Fabaceae</td>
<td>Macrotyloma uniflorum (Lam.) Verdc. var. uniflorum, P.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Bredemann &amp; Nieser, 01, 13</td>
<td>Fabaceae</td>
<td>Genista sagittalis L., Flügelginster, winged greenweed</td>
<td>17, Bredemann &amp; Nieser, 01, 15</td>
<td>Fabaceae, Genista tinctoria L., Färber-Ginster, dyer's greenweed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Bredemann &amp; Nieser, 01, 16</td>
<td>Fabaceae</td>
<td>Glycine max (L.) Merr., &quot;Sojabohne, Soja&quot;, &quot;soybean, Soy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

11.01.2022  | Juliane Jacob
Example XML

```xml
<?xml version="1.0" encoding="utf-8"?>
<mycoreobject xmlns:xsi = "http://www.w3.org/2001/XMLSchema-instance"
    xmlns:xlink = "http://www.w3.org/1999/xlink"
    xsi:noNamespaceSchemaLocation = "datamodel-document.xsd"
    ID="ngmcp-document-1376" label="The 1376 NGMP document">
    <!-- Structure information for MyCoRe Interna -->
    <structure />
    <!-- Description data for the document -->
    <metadata>
        <!-- Title -->
        <titles class="MCRMetaLangText">
            <title>A&7779;&amp;7789;avarga</title>
        </titles>
        <!-- Material -->
        <materials class="MCRMetaClassification">
            <material classid="ngmcp-class-00000011" categid="XX" />
        </materials>
        <!-- Language -->
        <languages class="MCRMetaClassification">
            <language classid="ngmcp-class-00000012" categid="S" />
        </languages>
        <!-- Script -->
        <scripts class="MCRMetaClassification">
            <script classid="ngmcp-class-00000013" categid="D" />
        </scripts>
    </metadata>
</mycoreobject>
```
Example JSON

```
{
  "id": 50283,
  "catalogno": "ZMH 22258",
  "collection": "entomology",
  "created-at": "2018-08-16T13:35:45.955Z",
  "updated-at": "2018-09-21T06:04:51.480Z",
  "details": {
    "tax-class": "Insecta",
    "tax-order": "Coleoptera",
    "odesc-name": "Tefflus, meyerlei, Fabricius 1801",
    "tax-domain": "Holometabola",
    "tax-family": "Carabidae",
    "geo-locality": "Côte d'Ivoire, Dimbroko",
    "geo-continent": "Africa",
    "zmh-catalog-no": "ZMH 22258",
    "identification-stage": "Imago",
    "identification-specimen": "pinned"
  },
  "pix": [
    "meyerlei-tefflus-zmh-22258-50mm-1x2-5.6f.jpg",
    "meyerlei-tefflus-zmh-22258-label.jpg"
  ]
}
```
Lossless versus lossy image compression

JPEG: 4273 Byte  PNG: 1007 Byte

All too late?

→ Data cleanup
Tools for data cleanup

- Open Refine (former Google Refine)
- Python Pandas
- Knime
Donate to *Programming Historian* today!

Cleaning Data with OpenRefine

Seth van Hooland, Ruben Verborgh, and Max De Wilde

This tutorial focuses on how scholars can diagnose and act upon the accuracy of data.

- Peer-reviewed
- CC-BY 4.0
- Support PH

**EDITED BY**

Adam Crymble

**REVIEWED BY**

Patrick Burns
Nora McGregor

**PUBLISHED | 2013-08-05**

**MODIFIED | 2021-10-27**

**TESTED | 2021-09-10**

**DIFFICULTY | Medium**

https://doi.org/10.46431

To get started, follow the directions in the "Setup" tab to download data to your computer and follow any installation instructions.

**Prerequisites**
Research Data Acquisition – collect data

1. Define methods and models for data processing
2. If necessary, localize existing data
3. Plan data management (formats, storage locations, etc.)
4. Clarify legal aspects of use and publication
5. Create / collect data
6. Create metadata / documentation
Focus on data quality

- Value ranges / value lists help with data quality
- Controlled vocabularies and ontologies or thesauri
- Predefined value ranges, format specifications (e.g. date YYYY-MM-DD)
- Validation (when entering / importing)
- Meaningful names
- Review / Peer Review / 4-eyes principle
Data organization

- The amount of data grows gradually during the course of the project
- Good organization helps save time and nerves
- Clarify conventions and procedures for
  - Storage location (network drives, shares)
  - Synchronization
  - Name
  - Versioning
  - Lifespan
  - Backup
Recommendations data organization

- Find out in depth about file formats & software
- Try to stick to standards and popular formats
- Structure and document your data and files in such a way that they are understandable even without your knowledge
- Use central backup storage systems and versioning for data protection
- Choose tools based on your needs
- Delete unnecessary data
Data organization - versioning

With file and folder names

- Suffix with serial number (V1, V2 ..., final, copy ...)
- Linking of uniform components in a defined order
- Avoidance of special characters and spaces
- If possible, mark every change
- Decide how many / which versions you want to keep
  - 20180301-drillcore-sibiria-abc-original
  - 20180301-microfilm1997-xyz-005
Stay Consistent

- L-12-cop-22092018.php
- L-12-cop-27102018.php
- L-12-cop-28092018.php
- L-12-cop-30092018.php
- L-13-cop-20122018.php
- L-13-todelete.phpL-6-new.php
- L-6.phpL-7-copy.php
Data organization - versioning

Many applications offer integrated versioning

- Activate versioning in Office programs, image processing, etc.
- Check persistence across sessions
- However, make sure to make regular backups
  - automated (TSM backup of the RRZ)
  - on servers
  - on external storage media
Version control software / text-based formats

- Free version control systems from the field of software development offer the possibility of parallel work and controlled consolidation through versioning
- Suitable for all text-based file formats
- `git` is currently de facto standard, distributed work
- `gitlab` is an open source server solution for software projects, which is available at Universität Hamburg
What does Gitlab offer?

- Intuitive responsive web interface for git servers
- File versioning at line level
- Integrated 'ticket tool', **kanban**
- Wiki
- Automatic code translation (CI)
- Integration of communication channels
Most common passwords (reddit.com)
Password security

- Do not reuse important passwords
  - have passwords generated automatically
  - think of a system to create variants of a password
  - Use a password manager

- Secure passwords included
  - no umlauts or special characters
  - no terms from dictionaries
  - lots of letters / numbers / characters

- Pay attention to how your password is transmitted / saved
  - 2-factor authentication, encryption
  - Pay attention to a secure connection when transmitting (https://)
Hazards

- SPAM, phishing, trojan emails
- Sale on the Darknet (especially streaming services)
- Shopping & redirection to packing station...
- Place fraudulent advertisements on sales platforms (eBay, Amazon)
- Loss of your data / your work
Security user accounts

- Check the security of your accounts
- At https://haveibeenpwned.com/ you can find out whether your data is openly circulating on the network
- Change duplicates of the login / password combination of other accounts
- Skimp on personal information, personal details on security issues
- Use different email aliases for different accounts
- Separate professional and private use
Stay up to date

**Tr0ub4dor & 3**
- ~28 bits of entropy
- Difficult to guess: Easy
- Difficulty to remember: Hard

**correct horse battery staple**
- ~44 bits of entropy
- Difficult to guess: Hard
- You've already memorized it

Through 20 years of effort, we've successfully trained everyone to use passwords that are hard for humans to remember, but easy for computers to guess.
Data Publication

Modes of Peer Review:

- **Blind Peer Review**
  - WHO SAID THAT?
  - WHERE IS MY REVIEW?

- **Open Identities**
  - PUBLISHED NEXT TO YOUR REPORT

- **Open Reports**
  - OH, THAT'S MORE CLEAR
  - MAH, MISSING A REFERENCE

**Open Data**
Metadata Exercise „Draw an owl”

- Writes detailed instructions on how to paint an owl (7 min)
- We appoint a leader who reads his/her instruction manual
- Everyone draws an owl
- Compare your pictures
Breakout-Session – Metadata

- Choose one of the two objects
- Describe the metadata in the padlet (10 min)
- Each group presents the results
Overview of discipline-specific metadata standards

https://www.dcc.ac.uk/guidance/standards/metadata
http://rd-alliance.github.io/metadata-directory/subjects/

- Data Documentation Initiative (social sciences)
- GESIS; Datorium or SSOAR (Social Science Open Access Repository)
- Dublin Core (bibliographic data format)
- Marc21 (bibliographic data format)
Why don't we use all the great tools out there? Keyword sustainability

- Online provider
- How long will the service last?
- Are there any data retention obligations?
- Are there (in the future) costs?
- What happens with my data?
- Are there any rules of succession or substitution?
- Are proprietary file formats used?
- Who has access / rights to my data?
- Can I delete data?
How we can sustainably use data in the future:

- Data can be found in a catalogue
- Data must be in a format that can still be opened
- Formats to be described in a format registry
- all data and documentation of the archiving actions are summarized in one information package
- Data must be comprehensible (close gaps between the expected knowledge of future users and the necessary knowledge)
What are Repositories?

“A repository (Latin repositorem, warehouse‘) is a managed place for storing organized documents that are publicly accessible or accessible to a limited group of users. In an archive (Latin: Archivum, filing cabinet ‘), however, only historical documents are managed.”

"Digital research data repositories are information infrastructures that store and organize digital research data as permanently as possible ... in order to ensure that the data can be found and accessed ...“

Quelle: Pampel (2014), Vortrag „Ausgewählte Aspekte digitaler Informationsversorgung“, HU Berlin
UHH research data repository - [www.fdr.uni-hamburg.de](http://www.fdr.uni-hamburg.de/)

... is used by all members of the Universität Hamburg to archive and publish data sets and documents from all areas of research.

- Registration with B identifier (Shibboleth)
- S3 storage connection (5 petabytes)
- Self-Service Portal
- CoreTrustSeal certification pending
- Access right: *open, embargoed, restricted or closed access*
- All file types possible
- At least 10 years retention period
- Each entry receives a DOI (including versions)
Trusted repositories: Certification (CoreTrustSeal)

- Criteria for CoreTrustSeal or Data Seal of Approval (self assessment)
- Data can be found on the internet
- Data is available (clear rights and licenses)
- Data can be reused (format)
- Data is reliable / trustworthy
- Data has PIDs so it can be referenced
https://www.re3data.org/
Basic requirements for Repositories

- Metadata Standards
- Licenses
- Access rights
- High quality data
- DOI
- Certification: CoreTrustSeal
Breakout-Session Repository

- [https://www.re3data.org/](https://www.re3data.org/)
- Search for a repository of your research field
- What conditions must be met?
- Discuss in your group
Outlook

Possible further lectures:

- Security issues
- Data cleansing with different tools (Open Refine, Python Pandas)
- Requirements of the research funders
- Advice on and communication of best practices (e.g. FAIR data, metadata, ...)
- Advice on technical aspects
- Advice on legal and ethical aspects
- Introduction to Heurist, RDMO or FDR
Useful links

- Web: www.fdm.uni-hamburg.de
- Research Data Repository: www.fdr.uni-hamburg.de
- Research Information System: www.fis.uni-hamburg.de
- Open Access Portal: www.oa.uni-hamburg.de
- Data Management Plan: www.dmp.fdm.uni-hamburg.de/
- FUNDus! Sammlungsportal www.fundus.uni-hamburg.de/
- www.forschungsdaten.org
- www.forschungsdaten.info
References - Reading recommendations

- Ludwig/Enke, 2013, „Leitfaden zum Forschungsdatenmanagement“
- EU-Guide
  
  https://www.scienceeurope.org/media/jezkhnoo/se_rd_m_practical_guide_final.pdf
Q&A
Thank you for your attention!

Contact
Center for Sustainable Research Data Management
Monetastraße 4, 20146 Hamburg
Mail:
fororschungsdaten@uni-hamburg.de
Juliane.jacob@uni-hamburg.de