Workshop for Sustainable Research
Data Management

Graduate School
"Collective decision-making"
Dr. Juliane Jacob
25.05.2021 10-12 and 1-3 pm
Overview

- Introduction
- Research Data, RDM and DMP
- FAIR principles
- Open Access and Policies
- Legal Rights and Licences
- Storage and publication of research data
- five exercise
- five Polls

Mood query
Who are you?

- Your research project?
- Your data?
- Your expectations of the workshop
  ....
Sustainable Research Data Management
What is research data?

- Research data is any information that has been collected, observed, generated or created to validate original research findings.
- Although usually digital, research data also includes non-digital formats such as laboratory notebooks and diaries.
Types of research data?

<table>
<thead>
<tr>
<th>METADATEN, z.B.:</th>
<th>PRIMÄRDATEN/ROHDATEN, z.B.:</th>
<th>SEKUNDÄRDATEN, z.B.:</th>
<th>FIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Name Autor</td>
<td>- Messdaten, Manuskripte, Fotografien, Aufnahmen, ...</td>
<td>- Aufbereitete Daten (Selektion, Korrektur, Aggregation, ...)</td>
<td>- Administrative Information, z.B. Mitarbeiter, Laufzeiten, Finanzierung von Projekten</td>
</tr>
<tr>
<td>- Marke / Modell Kamera</td>
<td>- Digitalisate</td>
<td>- Dokumentation (Methoden, Laborbücher, Versuchsprotokolle)</td>
<td>- Förderanträge</td>
</tr>
<tr>
<td>- Kalibration / Einstellungen</td>
<td></td>
<td>- Prozeduren (Algorithmen, Software)</td>
<td>- ...</td>
</tr>
<tr>
<td>- ...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What is RDM?

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.

Aspects of RDM:

- Order and structure
- Planning
- Storage and Backup
- Security and access rights
- Long-term archiving
- Publication of research data
- Legal Aspects
Do you ever have...

- described data?
- Re-used other data?
- Published data?
- not found your own data on your laptop?
- lost data?
- Not be able to understand data?
- Had problems with opening/reading data?
Why RDM?

Good RDM makes usage easier

- Find data faster (e.g. by name, structure)
- 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
“Sustainable” RDM

Open Data - benefits

- Makes data usable, instead of just using
- Freedom that goes beyond free of charge
- Allows for adaptation and updating of RD
- Enables collaborative work
- Strengthening communities
- Creates diversity of ideas
Why should I share my data at all?
Why should I share my data at all?

Slide by Brian Hole and Ubiquity Press
FAIR
Findable  Accessible  Interoperable  Reusable

Ah!

How do you open a .xboq file?

Here
Findable

Research data must be findable

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

- Permanent digital identifier, consisting of digits and/or alphanumeric characters, which is assigned to a digital object and refers directly to it.
- For research data:
  - DOI (Digital Object Identifier)
  - URN (Uniform Resource Name)
  - Handle
  - ORCID (People)
- 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.

http://dx.doi.org/10.5255/UKDA-SN-6969-1
Accessible

Research Data must be (long term) 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 resources have been deleted
Interoperable

RD must be 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

RD must be 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

- What tools do you and your colleagues use?
- How did you come in contact with data management plans?
- Are there any instructions from supervisors?
Exercise 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 (15min)
- Introduce your mind map briefly to the other seminar participants (5min)
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.
Elements of a DMP

1. Project description
   - General information on the project such as objectives, promoter and duration
   - Relevant policies

Source: https://www.uni-bielefeld.de/(en)/forschungsdaten/angebot/dmp/
Elements of a DMP

2. Existing types of data
   - Description of the existing data that can be reused for the project and their integration into the project.
   - Reasons, if no data is reused
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

- Information on the uniform regulation of data management within the project e.g. for:
  - Data storage
  - Data naming, folder structure
  - Synchronisation, Versioning
  - Collaborative working
  - Documentation
  - Creation of metadata
Elements of a DMP

5. Administrative and legal aspects
   - Funding and legal requirements
   - Copyright / Owner of data rights
   - Access and use
   - Data protection - for sensitive and personal data
   - Data backup/ encryption and backups
Elements of a DMP

6. Archiving, data exchange and publication
- Which data types are published?
- Which repository/archive?
- Access options (open, restricted, etc.), license
- Time and financial framework of data preparation, provision or archiving
- Reasons if no publication is planned
- Information on the procedure after the end of the retention period
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

- Personnel costs (data manager, data processor, data curator)
- Infrastructure (equipment, software, etc.)
- Service costs (data, long-term archiving, publication, training)
Phases of Research Data Management

**Planning & Conception**
- Project proposal
- DMP
- RIS

**Data Acquisition**
- Conception (Model)
- Methods (documentation)
- Digitisation
- Selection
- Storage & backups

**Working with Research Data**
- Processing
- Evaluation
- Analysis & Visualisation
- Communication
- Working Environment and Workflows
- Policies

**Publication**
- Journals
- RIS
- Repositories
- Metadata
- Legal Questions

**Archiving**
- Long-term storage on RRZ
- Subject specific repositories
- Repositories of the Contributors
Phases of Research Data Management

- Project proposal
  - DMP
  - RIS

Planning & Conception

Data Acquisition
- Conception (Model)
- Methods (documentation)
- Digitisation
- Selection
- Storage & backups

Working with Research Data
- Processing
  - Evaluation
  - Analysis & Visualisation
  - Communication
  - Working Environment and Workflows
  - Policies

Publication
- Journals
- RIS
- Repositories
- Metadata
- Legal Questions

Archiving
- Long-term storage on RRZ
- Subject specific repositories
- Repositories of the Contributors

Reuse
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
Exercise - RDMO (20min)

Research Data Management Organiser is a tool for creating data management plans with various templates

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 in the template for Horizon 2020
### Requirements of Research Funding Institutions

**Tabelle 1: Vergleich der Anforderungen der Förderer bezüglich des Datenmanagementplans**  
*(Stand: 07.07.2020)*

<table>
<thead>
<tr>
<th>Förderer</th>
<th>Plan gefordert?</th>
<th>Abgabe bei Antrag?</th>
<th>Inhalt</th>
<th>Updates?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EC Horizon 2020</strong></td>
<td>Datenmanagement-plan</td>
<td>Nein, erster Plan innerhalb der ersten 6 Projektmonate</td>
<td>Inhalte des Horizon 2020 Template</td>
<td>Update, falls signifikante Änderungen auftreten sowie zum Projektende</td>
</tr>
<tr>
<td><strong>DFG</strong></td>
<td>Angaben zum Umgang mit Forschungsdaten</td>
<td>Ja</td>
<td>Inhalte der Leitlinie zum Umgang mit Forschungsdaten</td>
<td>Nein</td>
</tr>
<tr>
<td><strong>BMBF</strong></td>
<td>Plan manchmal erforderlich, abhängig vom Programm</td>
<td>Falls notwendig, ja</td>
<td>Inhalt hängt vom jeweiligen Programm ab; Bildungsforschung : Checkliste</td>
<td>Kommt auf das Programm an</td>
</tr>
<tr>
<td><strong>VolkswagenStiftung</strong></td>
<td>Datenmanagement-plan</td>
<td>Ja</td>
<td>Inhalte des Science Europe Templates</td>
<td>Nein</td>
</tr>
</tbody>
</table>
Policy

- Internal guidelines or guidelines that are formally established by the company and via its management.
Interdisciplinary Policies:

Deutsche Forschungsgemeinschaft

DFG: Code of Conduct (Sept. 2019)

“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:

European Commission

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.
“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:

University of Hamburg (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 ("goldener Weg"), 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 ("grüner Weg", "self-archiving").

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 Rights and Open Access
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

   DSGVO: "‘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."
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 (GDPR/DSGVO):
   - Voluntary consent
   - "informed consent"
   - The purpose and type of processing of the data must be clearly defined

2. 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 daraus Nachteile entstehen."

3.2 Einwilligungsformular

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 (exception 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 transferred (also verbally) 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
Contact Person Legal Rights

- UHH data protection: datenschutz@uni-hamburg.de
- Multimedia Kontor Hamburg https://www.hh-datenschutz.de/
 Licenses

- 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
Creative Commons Lizenzen

only CC-BY und CC0 are conform to the FAIR-principles
Lunch break
Research Data Acquisition
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
# Types of research data acquisition

<table>
<thead>
<tr>
<th>Type of acquisition</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collection / recording of own data</td>
<td>full control over type and scope</td>
<td>combined with considerable costs and risks</td>
</tr>
<tr>
<td>Reuse of existing data</td>
<td>Costs / effort of the survey is mainly borne by others</td>
<td>Scope and format of the data is specified</td>
</tr>
<tr>
<td></td>
<td>Potentially greater scope (meta studies)</td>
<td>possibly considerable training effort</td>
</tr>
<tr>
<td></td>
<td>Standardized methods</td>
<td></td>
</tr>
</tbody>
</table>


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
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
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?)

„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)“
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?)
### Reuse of data

**Exercise (Excel table data quality)**

<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>
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

- = 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.*

- The Internet Media Type / MIME (Multipurpose Internet Mail Extension) ensures that the correct application is assigned to the file formats in the www
# MIME-Types and File-extensions

<table>
<thead>
<tr>
<th>MIME-Typ</th>
<th>Dateiendung(en)</th>
<th>Bedeutung</th>
</tr>
</thead>
<tbody>
<tr>
<td>application/acad</td>
<td>*.dwg</td>
<td>AutoCAD-Dateien (nach NCSA)</td>
</tr>
<tr>
<td>application/applefile</td>
<td></td>
<td>AppleFile-Dateien</td>
</tr>
<tr>
<td>application/astound</td>
<td>*.asd *.asn</td>
<td>Astound-Dateien</td>
</tr>
<tr>
<td>application/dsptype</td>
<td>*.tsp</td>
<td>TSP-Dateien</td>
</tr>
<tr>
<td>application/dxf</td>
<td>*.dxf</td>
<td>AutoCAD-Dateien (nach CERN)</td>
</tr>
<tr>
<td>application/force-download</td>
<td></td>
<td>Registrierungsdateien</td>
</tr>
<tr>
<td>application/futuresplash</td>
<td>*.spl</td>
<td>Flash Futuresplash-Dateien</td>
</tr>
<tr>
<td>application/gzip</td>
<td>*.gz</td>
<td>GNU Zip-Dateien</td>
</tr>
<tr>
<td>application/javascript</td>
<td>*.js</td>
<td>serverseitige JavaScript-Dateien</td>
</tr>
<tr>
<td>application/json</td>
<td>*.json</td>
<td>enthält einen String in JavaScript-Objekt-Notation</td>
</tr>
<tr>
<td>application/listenup</td>
<td>*.pltk</td>
<td>Listenup-Dateien</td>
</tr>
<tr>
<td>application/mac-binhex40</td>
<td>*.hqx</td>
<td>Macintosh Binärdateien</td>
</tr>
<tr>
<td>application/mbedlet</td>
<td>*.mbd</td>
<td>Mbedlet-Dateien</td>
</tr>
<tr>
<td>application/mif</td>
<td>*.mif</td>
<td>FrameMaker Interchange Format Dateien</td>
</tr>
<tr>
<td>application/msexcel</td>
<td>*.xls *.xla</td>
<td>Microsoft Excel Dateien</td>
</tr>
<tr>
<td>application/mshelp</td>
<td>*.hlp *.chm</td>
<td>Microsoft Windows Hilfe Dateien</td>
</tr>
<tr>
<td>application/mspowerpoint</td>
<td>*.ppt *.ppz *.pps *.pot</td>
<td>Microsoft Powerpoint Dateien</td>
</tr>
<tr>
<td>application/msword</td>
<td>*.doc *.dot</td>
<td>Microsoft Word Dateien</td>
</tr>
<tr>
<td>application/octet-stream</td>
<td>*.bin *.file *.com *.class *.ini</td>
<td>Nicht näher spezifizierte Daten, z.B. ausführbare Dateien</td>
</tr>
<tr>
<td>application/oda</td>
<td>*.oda</td>
<td>Oda-Dateien</td>
</tr>
</tbody>
</table>
Proprietary file formats

- are mostly dependent on
  - specific software
  - corresponding software licenses
  - specific platforms / operating systems
  - are at risk from ‘software obsolesence’
  - 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>Recommendation</th>
<th>Avoid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tabellendaten</td>
<td>CSV, TSV, SPSS portable</td>
<td>Excel</td>
</tr>
<tr>
<td>Text</td>
<td>TXT, ODT, HTML, RTF, PDF/A only if layouts is important</td>
<td>Word, PowerPoint</td>
</tr>
<tr>
<td>Multimedia</td>
<td>Container: MP4, Ogg Codec: Theora, Dirac, FLAC</td>
<td>QuickTime, H264</td>
</tr>
<tr>
<td>Bild</td>
<td>TIFF, JPEG2000, PNG</td>
<td>GIF, JPG</td>
</tr>
<tr>
<td>structures 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)
<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</td>
<td>01</td>
<td>01</td>
<td>Fabaceae</td>
<td>Anthyllis vulneraria L., Wundklee, woundwort</td>
<td>0001-Anthyll</td>
<td>wundwort</td>
</tr>
<tr>
<td>2</td>
<td>Bredemann &amp; Nieser</td>
<td>01</td>
<td>02</td>
<td>Fabaceae</td>
<td>Astragalus boeticus L., Kaffee-Traganth, swedish coffee</td>
<td>(</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Bredemann &amp; Nieser</td>
<td>01</td>
<td>03</td>
<td>Fabaceae</td>
<td>Astragalus falcatus Lam., Sicheltragant, russian milk vetch</td>
<td>(</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Bredemann &amp; Nieser</td>
<td>01</td>
<td>04</td>
<td>Fabaceae</td>
<td>Astragalus glycyphyllos L., Süßer Tragant, sweet milk vetch</td>
<td>(</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Bredemann &amp; Nieser</td>
<td>01</td>
<td>05</td>
<td>Fabaceae</td>
<td>Astragalus hamosus L., Angelstragel, yellow milk vetch</td>
<td>(</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Bredemann &amp; Nieser</td>
<td>01</td>
<td>05a</td>
<td>Fabaceae</td>
<td>Cajanus cajan (L.) Millsp., Taubenerbse, pigeonpea</td>
<td>(</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Bredemann &amp; Nieser</td>
<td>01</td>
<td>06</td>
<td>Fabaceae</td>
<td>Senna tora (L.) Roxb., Gemüse-Kassie, sicklepod</td>
<td>(</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Bredemann &amp; Nieser</td>
<td>01</td>
<td>07</td>
<td>Fabaceae</td>
<td>Cicer arietinum L., Kichererbse, chickpea</td>
<td>(</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Bredemann &amp; Nieser</td>
<td>01</td>
<td>08</td>
<td>Fabaceae</td>
<td>Coronilla scorpioides (L.) W.D.J.Koch, Skorpionspeitsche</td>
<td>(</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Bredemann &amp; Nieser</td>
<td>01</td>
<td>09</td>
<td>Fabaceae</td>
<td>Securigera varia (L.) Lassen, Bunte Kronwicke, Crown-vetch</td>
<td>(</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Bredemann &amp; Nieser</td>
<td>01</td>
<td>10</td>
<td>Fabaceae</td>
<td>Cyamopsis tetragonoloba (L.) Taub., Guar, Guar, 0011-Cyar</td>
<td>(</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Bredemann &amp; Nieser</td>
<td>01</td>
<td>10a</td>
<td>Fabaceae</td>
<td>Macrotyloma uniflorum (Lam.) Verdc. var. uniflorum, Pety's rue, Pety whin</td>
<td>(</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Bredemann &amp; Nieser</td>
<td>01</td>
<td>11</td>
<td>Fabaceae</td>
<td>Galega officinalis L., Echte Geisraute, goat's-rue, 0013-Beere</td>
<td>(</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Bredemann &amp; Nieser</td>
<td>01</td>
<td>12</td>
<td>Fabaceae</td>
<td>Genista anglica L., Englischer Ginster, petty whin, 0014-Deutscher Ginster</td>
<td>(</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Bredemann &amp; Nieser</td>
<td>01</td>
<td>13</td>
<td>Fabaceae</td>
<td>Genista germanica L., Deutscher Ginster, german greenweed</td>
<td>(</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Bredemann &amp; Nieser</td>
<td>01</td>
<td>14</td>
<td>Fabaceae</td>
<td>Genista sagittalis L., Flügelginster, winged greenweed</td>
<td>(</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Bredemann &amp; Nieser</td>
<td>01</td>
<td>15</td>
<td>Fabaceae</td>
<td>Genista tinctoria L., Färber-Ginster, dyer's greenweed</td>
<td>(</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Bredemann &amp; Nieser</td>
<td>01</td>
<td>16</td>
<td>Fabaceae</td>
<td>Glycine max (L.) Merr., &quot;Sojabohne, Soja&quot;, &quot;soybean, Soy&quot;</td>
<td>(</td>
<td></td>
</tr>
</tbody>
</table>
<?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-4376" label="The 4376 NGMCP document">

<!-- Strukturinformationen fuer MyCoRe-Interna -->
<structure />

<!-- Beschreibungsdaten zum Dokument -->
<metadata>

<!-- 01 - Titel -->
<titles class="MCRMetaLangText">
   <title>Aṣṭavarga</title>
</titles>

<!-- 02 - Material -->
<materials class="MCRMetaClassification">
   <material classid="ngmcp-class-00000011" categid="XX" />
</materials>

<!-- 03 - Language -->
<languages class="MCRMetaClassification">
   <language classid="ngmcp-class-00000012" categid="S" />
</languages>

<!-- 04 - Script -->
<scripts class="MCRMetaClassification">
   <script classid="ngmcp-class-00000013" categid="D" />
</scripts>
</mycoreobject>
<!-- 06 - Author -->
<authors class="MCRMetaLinkID">
  <author xlink:type="locator" xlink:href="ngmcp-author-00000002"
    xlink:label="author" xlink:title=""/>
</authors>

<!-- 07 - Subject -->
<subjects class="MCRMetaClassification">
  <subject classid="ngmcp-class-0000014" categid="J"/>
</subjects>
[...]
</xml>
Example JSON

```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

PNG: 1007 Byte

JPEG: 4273 Byte

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-12.php
L-13-cop-16120182.php
L-13-cop-20122018.php
L-13-cop-mittag.php
L-13.php
L-13-todelete.php
L-6-new.php
L-6.php
L-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
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
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?
Services of RRZ

Software
https://www.rrz.uni-hamburg.de/services/software/alphabetisch.html

Serviceportal
https://rrz-serviceportal.uni-hamburg.de/

UHH-Share/Disk
Share folders and documents

Sharepoint
files, office applications, lists

gitlab
Source code / texts, wikis, task lists
Most common passwords (lastfm)
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 (http s : //)
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

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.
All too late?!
Data cleansing
Tools for data cleansing

- OpenRefine (former GoogleRefine)
- Python Pandas
- Knime
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)
Data publication

Modes of Peer Review:

- Blind Peer Review
- Open Identities
- Open Reports

Questions:

- Who said that?
- Where is my review?
- Published next to your article
- Hmm, missing a reference
- Ah, that's more clear
What is metadata?

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

- Technology used in data generation (e.g. hardware and software)
- Administrative information (project, participants, institution, etc.)
- Relationships with other data (publications, people, projects, etc.)
## Metadata types

<table>
<thead>
<tr>
<th>Metadata type</th>
<th>Target</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive metadata</td>
<td>Identification and discoverability</td>
<td>Title, author, abstract, keywords ...</td>
</tr>
<tr>
<td>Administrative metadata</td>
<td>Creation and storage</td>
<td>Rights management and access rights, storage...</td>
</tr>
<tr>
<td>Structural metadata</td>
<td>Structure of the resource</td>
<td>Chapters, sections, structure, files...</td>
</tr>
<tr>
<td>Technical metadata (&quot;technical metadata&quot;)</td>
<td>technical characteristics</td>
<td>Scope, coding, technical structure, formats</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
Overview of discipline-specific metadata standards

Keywords

- Data Documentation Initiative (social sciences)
- GESIS; Datorium or SSOAR (Social Science Open Access Repository)
- Dublin Core (bibliographic data format)
- Marc21 (bibliographic data format)

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

- Two breakout sessions
- Each group has an unknown object
- Describe the metadata (10 min)
- Take notes in the padlet
- Each group then briefly presents the results
What are Repositories?

“A repository (Latin repositorium, 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
Growth of Data Repositories

Cambridge Structural Database
Repositories

Discipline-specific repositories are usually the best place for your data:

https://datorium.gesis.org
The UHH research data repository

... is used by all members of the University of Hamburg to archive and publish (closed) 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
- You decide whether your entry is open, restricted or closed access
- All file types possible
- Upload up to 50GB via the web interface
- At least 10 years retention period
- Each entry receives a DOI (including versions)

https://www.fdr.uni-hamburg.de/
Repositories with DINI certificate

- standardized assessment of (document) servers and services provided
- Quality control for document and publication services
- Criteria a.o.
  - Visibility of the entire offer
  - Supervision of the authors
  - Security, authenticity and integrity of the technical system
  - Long-term availability and findability of the documents archived there
- In Germany currently ~ 50 DINI-certified repositories (only document repositories with open access)
Trusted repositories: Certification (CoreTrustSeal)

- Criteria for DSA (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
Data in the repositories is not a matter of course

- The catalog does not provide sufficient explanations
- This is why metadata is essential
- Detailed information on the landing pages and in additional documents
- Landing pages may no longer be accessible in the long term
- Data producers are no longer available for inquiries

*Data should be reusable!*
Search of corresponding repositories

- **Re3data.org** (Karlsruhe Institute of Technology)
- **OpenDOAR** UK
- **Repository Finder** Search of repositories which correspond to Enabling FAIR Data Projekt Principles in Re3Data
Search of corresponding *data* in repositories

- DataCite.org NGO / DOI provider
- EU Open Data Portal
- B2FIND Recherche in EUDAT FD Sammlungen
- Research Data Australia
- ....
Exercise Repositories

- Scroll through re3data.de
- Find a repository for your data
- Note your choice in the padlet
How we can 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)
Useful links

- Center for RDM https://www.fdm.uni-hamburg.de/
- Research Date Repository https://fdr.uni-hamburg.de
- DMP Online http://dmp.fdm.uni-hamburg.de/
- www.forschungsdaten.org
- www.forschungsdaten.info
References - Reading recommendations

- Ludwig/Enke, 2013, „Leitfaden zum Forschungsdatenmanagement“
- Gesis.org: Anwendung der Datenschutzgesetzgebung auf die empirische Sozialforschung (z.B. Watteler & Kinder-Kurlanda 2015),
- Sicherheitspraktiken in Organisationen (z.B. Poller et al. 2017), und
Evaluation
VIELEN DANK!