Research Data Management in a Nutshell
BERD@BW Data Literacy Snack, 26.5.2021

Irene Schumm, Lorena Steeb
Mannheim Univ. Library, Research Data Center

Except where otherwise noted, this work is licensed under a Creative Commons Attribution 4.0 International License
The BERD@BW project

• **Center for Business, Economic and Related Research Data**, founded by the University of Mannheim and ZEW

• Connecting research and infrastructure for a better Research Data Management (RDM) in Business, Economics and related areas

• Funded by the state Baden-Württemberg
Agenda for today

- What is research data management?
- Why engage in RDM?
- Elements of RDM
- The Research Data Management Plan (RDMP)
What is research data management?

RDM describes the
a) organization
b) storage
c) preservation, and
d) sharing

of data collected and used along the research data cycle:

Why engage in RDM?

Knowledge preservation: Data (especially digital data) is fragile, easily lost and often expensive to collect.
Why engage in RDM?

Knowledge preservation: Data (especially digital data) is fragile, easily lost and often expensive to collect.

- Improve collaborative work
- Validation and replication of (own) research findings
- Reusability of data: Enhanced visibility, faster retrieval, better understanding
Why engage in RDM?

Why engage in RDM?

Knowledge preservation: Data (especially digital data) is **fragile, easily lost and often expensive to collect**

Policies and requirements
- Research institutions
- Publishers
- Funders

Growing **public awareness and interest** in Open Data

Improve **collaborative work**

Validation and replication of (own) research findings

Reusability of data: Enhanced visibility, faster retrieval, better understanding
Why engage in RDM? – Policies of research institutions

Example: Code of Good Research Practice, Univ. of Mannheim (2014)

- "Studies shall be replicable. Therefore, publications shall include a complete and detailed description of the methods of data collection, the statistic analysis and the results in order to allow for verification of the results through replication."

- "After its publication, research data shall be passed on for further scholarly use provided that this does not violate any legal or contractual regulations."

- "The person responsible for the research project shall make sure that the original data set, on which publications, patents and/or follow-up works are based, is stored on durable and secure data storage devices for at least ten years after publication or patenting."

Source: https://www.uni-mannheim.de/media/Universitaet/Dokumente/Richtlinie_gute_wissenschaftliche_Praxis_en.pdf
Why engage in RDM? – Policies of publishers

Example: **Data and Code Availability Policy,** American Economic Association

- “It is the policy of the American Economic Association to publish papers only if the data and code used in the analysis are clearly and precisely documented and access to the data and code is non-exclusive to the authors.”
- “Authors of accepted papers that contain empirical work, simulations, or experimental work must provide, prior to acceptance, information about the data, programs, and other details of the computations sufficient to permit replication, as well as information about access to data and programs.”
- “Data and programs should be archived in the AEA Data and Code Repository.”

Source: https://www.aeaweb.org/journals/data/data-code-policy
## Why engage in RDM? – Policies of funders

<table>
<thead>
<tr>
<th></th>
<th>DFG</th>
<th>EU Horizon Europe (2021-2027)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>What?</strong></td>
<td><strong>Suggestion</strong>: Archive/publish data, materials, information, methods, source code and software used to obtain the research findings, as far as possible and reasonable; in the <strong>proposal</strong>: description of type, extent, documentation of data originating or used in the project, archiving plans and options for reuse.</td>
<td>In the <strong>proposal</strong>: „Applicants generating/collecting data and/or other research outputs [...] must provide maximum 1 page on how the data/research outputs will be managed in line with the FAIR principles“ → Month 6: <strong>DMP deliverable</strong></td>
</tr>
<tr>
<td><strong>Where?</strong></td>
<td>Archives/repositories in your own institution or other, well-established infrastructure</td>
<td>Not specified</td>
</tr>
<tr>
<td><strong>How long?</strong></td>
<td>10 years</td>
<td>Not specified</td>
</tr>
</tbody>
</table>
Elements of RDM
Searching and finding data

• Commonly used or official data sources (e.g. Datastream, Administrations, GSOEP, GESIS, ICPSR...)

• ... and the „long tail“
  – re3data: search engine for data repositories → find data repository (institutional, discipline-specific...) and continue data search there
  – Zenodo: multidisciplinary dataset repository
  – Google Dataset Search: multidisciplinary dataset search engine
Legal issues to watch out for: re-using data, proprietary data

- In general: mind copyright issues and licences
- Proprietary data (e.g. from commercial providers): license agreements, sometimes regulations not very transparent
- Data from data repositories/research data centers: do often come with a license
- Data from the web: copyright situation and license often unclear

→ Exemptions from copyright for researchers (e.g. for TDM)
• When processing personal data, data protection regulations have to be taken into account (EU: GDPR + national/state legislation)
• Exemptions/permissions for research – under conditions, e.g.:
  – Informed consent or protects public interest
  – Immediate pseudonymization, anonymization as soon as possible
• Involve data protection official and/or ethics committee as needed
• Also use due diligence for confidential data

See also: Data Protection Guide by RatSWD; Video tutorial: Umgang mit personenbezogenen Forschungsdaten - Rechtliche Grundlagen, Methoden und Hilfsmittel by Frauke Ziedorn et al.
Data protection, confidentiality, ethics

- When processing personal data, data protection regulations have to be taken into account (EU: GDPR + national/state legislation)
- Exemptions/permissions for research – under conditions, e.g.:
  - Informed consent or protects public interest
  - Immediate data anonymization or pseudonymization as soon as possible
- Involve data protection official and/or ethics committee as needed
- Also use due diligence for confidential data

Data Literacy Snack

Privacy Law Basics for Research Data

June 9, 1pm

More info: https://www.berd-bw.de/snacks
Documentation: data and code

**Project/study level information** (e.g. ReadMe)
guidance through the materials (data and code files, ...) needed to understand and reproduce analysis results

**Data level documentation**
(e.g. codebooks, data dictionaries, DDI metadata standard)

**Code level documentation**

→ More: Slides from Data Literacy Snack „*Reproducible Data Analysis 101*“ by Juli Tkotz
File management

File management

• File naming → establish conventions
  – Top level identifying information (e.g. project name or acronym, study title)
  – Include date (YYYYMMDD) and/or version (v01, v02)
  – If you use numbering, use leading zeroes
  – Not too long
  – more: guide from Stanford Univ. Libraries

• Code versioning: GitHub
Information security in RDM

- Encryption (e.g. Bitlocker, Cryptomator, encrypted zip file)
- Institutional file servers (with backup routines)
- Local storage device (as additional backup)
- Keep operating system and software updated
- Password protection
- Cloud storage (as backup)
- Lock up physical devices when not in use
And as the data gets colder...
What to publish/preserve?

Reference to: https://www.ru.nl/rdm/archiving-data/what-data-should-archived/
Licenses for sharing research data and code

Research data

Code

GNU General Public License (GPL)

MIT License

More: https://choosealicense.com/

→ Mind restrictions if you re-use data or code

Image sources and licenses:
CC logo by https://creativecommons.org/, CC BY 4.0
GPL logo by Free Software Foundation, Inc., CC BY-ND 4.0
MIT logo, public domain
File formats for preservation

→ Non-proprietary, open standard, uncompressed, unencrypted

Plain text
- txt
- doc

Tabular/structured data
- csv
- xml
- xls

Images
- tif, tiff
- ind

Video
- mp4
- wmv

Audio
- wav

Formatted text
- PDF/A
- doc, ppt, PDF
Where to publish/archive?

- Institutional data repository (e.g. MADATA) or research data center
- Journal data archives (e.g. American Political Science Review)
- Discipline-specific repository (e.g. GESIS)
- Multi-disciplinary repository (e.g. Zenodo)
- Data journal (e.g. Research Data Journal for the Humanities and Social Sciences)
- Code repository (e.g. GitHub)

Image source and license: by re3data, CC BY 4.0
Typical meta data in a data repository (here: MADATA)

<table>
<thead>
<tr>
<th>Item Type</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Benutzerumfrage der Universitätsbibliothek Mannheim 2016 - Fragebogen und Antwortdaten</td>
</tr>
<tr>
<td>Alternative Title</td>
<td>Survey of the Mannheim University Library 2016 - questionnaire and results</td>
</tr>
<tr>
<td>Date</td>
<td>20 April 2017</td>
</tr>
<tr>
<td>Creator</td>
<td>Auberer, Benjamin ; Kaiser, Jessica ; Leichtweiß, Angela</td>
</tr>
<tr>
<td>Divisions</td>
<td>Zentrale Einrichtungen - University Library</td>
</tr>
</tbody>
</table>

**DDC Classification:**
020 Library and information sciences

**Abstract:**

**Related publications to the data**
Auberer Benjamin und Klein Annette und Kaiser Jessica
Abschlussbericht zur Umfrage an der Universitätsbibliothek Mannheim 2016

**Data files (with license + accessibility info)**

**Digital Object Identifier → persistent identification**
https://doi.org/10.7801/1

**URI:**
https://madata.bib.uni-mannheim.de

**Availability (Controlled) → Delivery**

**File**

<table>
<thead>
<tr>
<th>Filename</th>
<th>Infos</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td></td>
</tr>
<tr>
<td>Filename:</td>
<td>Benutzerumfrage_2016_USBib Mannheim Fragebogen_deutsch.pdf</td>
</tr>
</tbody>
</table>

**PUBLIC DOMAIN**
A Research Data Management Plan (RDMP) ...

... is a „living document“ which keeps record of what research data is created and what happens to that data during and after the project.

... contains aspects like:

• Who is responsible for the processes of research data management?
• What data will be used/produced (type, format, size/amount, source)? How will it be processed? How will it be documented?
• Could legal or ethical problems occur in collecting, analyzing and archiving/publishing the data?
• How and where shall the data be stored during and after the project duration? How shall regular back-ups be handled?
• Is it planned to publish the (meta-)data, and if yes, under which conditions?
• How much does it cost? (Who will cover it?)
## RDMP guidance

### SCIENCE EUROPE: Practical Guide to the international alignment of research data management

#### 5 DATA SHARING AND LONG-TERM PRESERVATION

<table>
<thead>
<tr>
<th><strong>Guidance for Researchers</strong></th>
<th><strong>Sufficiently Addressed</strong></th>
<th><strong>Insufficiently Addressed</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>5a</strong></td>
<td>Sufficiently Addressed The DMP...</td>
<td>Insufficiently Addressed The DMP...</td>
</tr>
<tr>
<td>How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?</td>
<td>- Clearly describes how the data and/or metadata will be made discoverable and shared.</td>
<td>- Provides little or no details on how and when data will be shared, or the explanation is not adequate or technically viable.</td>
</tr>
<tr>
<td>- Explain how the data will be discoverable and shared (for example by deposit in a trustworthy data repository, indexed in a catalogue, use of a secure data service, direct handling of data requests, or use of another mechanism).</td>
<td>- Specifies when data will be shared and under which licence.</td>
<td></td>
</tr>
<tr>
<td>- Outline the plan for data preservation and give information on how long the data will be retained.</td>
<td>- Includes the name of the repository, data catalogue, or registry where data will or could be shared.</td>
<td></td>
</tr>
<tr>
<td>- Explain when the data will be made available. Indicate the expected timely release. Explain whether exclusive use of the data will be claimed and if so, why and for how long. Indicate whether data sharing will be postponed or restricted for example to publish, protect intellectual property, or seek patents.</td>
<td>- Includes information on how long the data will be retained and gives precision on its timely release.</td>
<td></td>
</tr>
<tr>
<td>- Indicate who will be able to use the data. If it is necessary to restrict access to certain communities or to apply a data sharing agreement, explain how and why. Explain what action will be taken to overcome or to minimise restrictions.</td>
<td>- Clearly explains, if applicable, why data sharing is limited or not possible, and who can access the data under which conditions (for example, only members of certain communities or via a sharing agreement).</td>
<td></td>
</tr>
</tbody>
</table>

Source and license: by [Science Europe](https://www.science-europe.eu), CC BY 4.0
RDMP tools

- **RDMO – Research Data Management Organizer** of Leibniz-Institut für Astrophysik Potsdam & KIT (you can use the entity of forschungsdaten.info): different templates, like DFG, Horizon 2020, DMPonline, DMPTool
- **DMPonline** of the Digital Curation Center (UK): different templates, like Horizon 2020, UK-funders
- **DMPTool** of the University of California Curation Center (USA): different templates, like Horizon 2020, US-funders
- **ARGOS** from the EU-funded project OpenAIRE: template for Horizon 2020 projects
- **Data Stewardship Wizard**: of the Czech Technical University: different templates, e.g. Science Europe Template
THANK YOU

Follow us on Twitter: @berd_bw

Register to our BERD newsletter
Sources for digging deeper (1)

• General
  – Data Management Guide of MIT Libraries
  – Data Management Services from Stanford Univ. Libraries
  – Research Data Management by Radboud Univ.
  – Data Management Basics 1: Introduction to data management and sharing by UK Data Services
  – forschungsdaten.info (mainly in German)

• Documentation
  – ICPSR Guide to Codebooks
  – Template ReadMe for Social Science replication packages
  – Document your data by UK Data Services

• IT security
  – Basistipps zur Informationssicherheit by the Federal Office for Information Security, Germany
  – Personenbezogene Forschungsdaten - Kapitel 4: Schutz vor Datenmissbrauch by Leibniz Univ. Hannover
Sources for digging deeper (2)

• Licenses
  – Data: Creative Commons
  – Software: https://choosealicense.com/ and https://choosealicense.com/licenses/

• Data Protection
  – Umgang mit personenbezogenen Forschungsdaten – Rechtliche Grundlagen, Methoden und Hilfsmittel by Leibniz Univ. Hannover
  – Data Management Basics 2: Ethical and legal issues in data sharing by UK Data Services

• Data Management Plans
  – Practical Guide to the International Alignment of Research Data Management by Science Europe