




 DIO

Projekt: Data Stewardship

1. Deep Dive Workshop
IKTdZ 10. Ausschreibung

Förderung durch:

 Bundesministerium
Klimaschutz, Umwelt,
Energie, Mobilität,
Innovation und Technologie

 **FFG**
Forschung wirkt.

Welcome



First name basis?
It's friendlier this way



Sarah

Data Steward @ DIO

sarah.stryeck@dataintelligence.at

Lisa

Tech Lead @ DIO

lisa.nussbaumer@dataintelligence.at

Data Steward for IKTdZ10

Goal: assess and demonstrate the impact of data stewardship practices on research project outcomes and the quality of data management plans

Offer to IKTdZ10 projects:

- 3 cross-project Deep Dive Workshops
- max. 3 individual data consulting days / project (online or offline)
- 1 public Workshop

Agenda



Introduction of participants and projects

Block 1: Data Management

Coffee-Break

Exercise 1

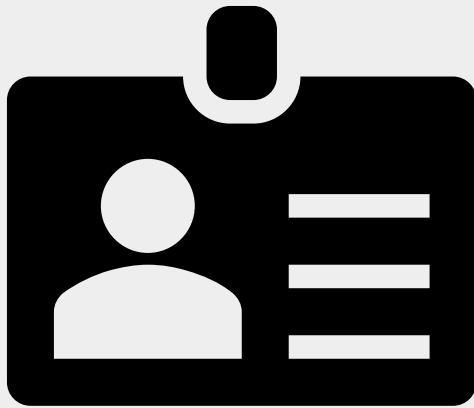
Block 2: Version Control

Coffee-Break

Block 3: Quality Assurance

Exercise 2

Wrap-up



Introduce yourself!

- Name
- Research Project
- Expectations

Block 1: Data Management

What is Data Management?

Data management is a concept for the usage of digital data...

- ... from generation...
- ... storage ...
- ... processing...
- ...archival...
- ...to deletion,
- ... as well as documentation of all steps.

Why do we need it?

- To handle large amounts of data
- To enable access to data for several individuals (e.g., collaborators)
- To ensure a single point of truth
- To ensure findability and back-up strategies
- To ensure data safety and data security
- To ensure compliance with legal regulations (e.g., storage and deletion of sensitive data)
- To ensure data quality and integrity
- To (re-)use data and generate added value (see dark data)

Documentation and FAIR principles

- Data documentation is crucial for ensuring transparency and reproducibility in research, allowing others to understand, validate, and build upon findings.

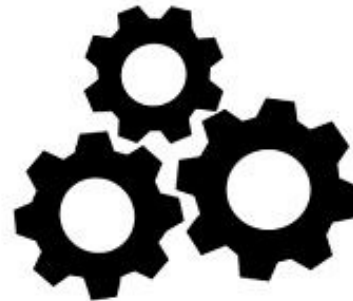
F_{indable}



A_{ccessible}



I_{nteroperable}



R_{eusable}



Findable

- **Attach a DOI** to your data. Many data platforms (e.g. Zenodo) make that really easy for you.
- Provide **rich machine-readable metadata**. If you upload your data to a good data platform, the most relevant metadata will be asked from you anyhow. So it's easy to do things right.



Dr. Heidi Seibold (How-to-FAIR?)

Accessible

FAIR is not the same as Open  the point is to provide the exact conditions of accessibility

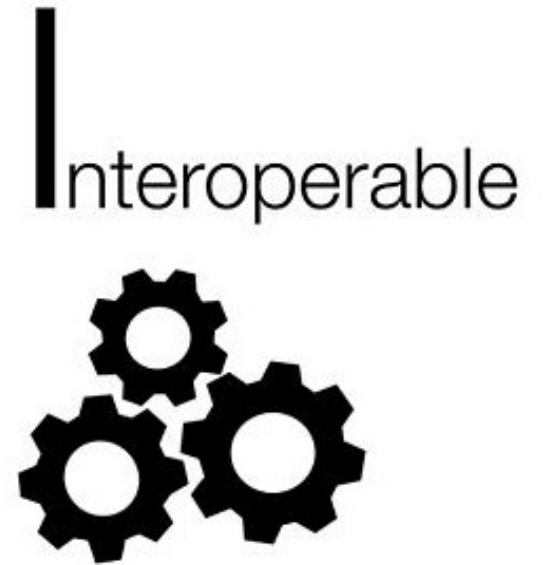
- Explain how someone can access your data. May that be via **accessing it through a data platform** or **through an application** that is evaluated by a data-use-and-access committee.



Dr. Heidi Seibold (How-to-FAIR?)

Interoperable

- Use **common data formats**. For tabular data that could for example be csv, for images jpeg. What's best in your community might be decided through a community standard.
- Use **words that others will understand** or define them. For example if the column names in your table are not self explanatory, explain them.
- Provide **context for your data**. Is it connected with other data or papers? You can also add your metadata to public knowledge graphs, e.g. Wikidata.



Dr. Heidi Seibold (How-to-FAIR?)

Re-usable

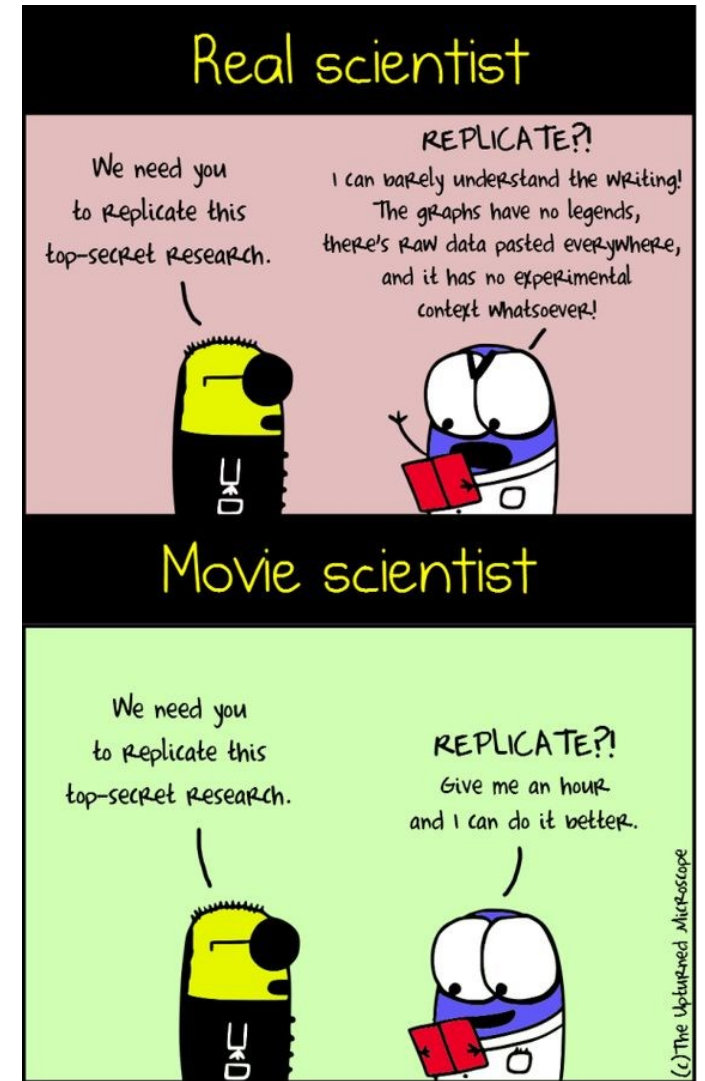
- Include rich **machine-readable metadata according to the community standards.**
- Attach a **license** to your data (license is part of the metadata) that makes it clear, what others can do with your data.



Dr. Heidi Seibold (How-to-FAIR?)

3 Essential Elements of Successful Research Data Management

- 1) Backups
- 2) Documentation
- 3) Future-Proofing Files



1) Backups

- Prevent data loss with the **3-2-1 rule**

3 copies of data:

- Local/working copy (Here)
- Other local/external copy or in a remote location (Near)
- Remote copy (Cloud, external server), geographically separate (Far)

2 geographically separate locations:

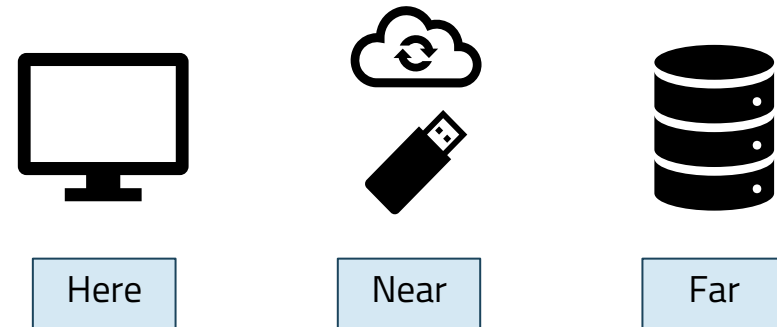
- Here (local)
- Near (remote)

More than **1** type of storage device:

- Local disk
- External drive
- Cloud storage

- Automated, regular Backups

Example:



2) Document, Document, Document

- **ORGANIZATIONAL SYSTEM**

- Standardized file naming conventions
- Folder hierarchy

- **WORKFLOW**

Documenting the workflow used in the research process:

Details on data cleaning, processing, analysis, and visualization techniques

- **DATA AND CODE**

- Units/measurement scales, variables
- Assumptions
- What do acronyms stand for?



Example: **NLP**

- 1) Technology/Computing: **N**atural **L**anguage **P**rogramming
- 2) Healthcare/Medicine: **N**euro-**L**inguistic **P**rogramming

Types of Documentation

- **README:** is a text file typically found in a project-related folder or alongside a dataset, providing an overview of the contents and structure of the folder or dataset.
- **Data Dictionary:** functions as a reference guide and defines and describes the various elements, variables, and fields within a dataset.
- **Protocol:** is a document that outlines the procedures or experimental design used in research projects.
- **Lab Notebooks:** serve as primary records documenting the research process.
- **Data Management Plan:** is a document that outlines the strategies and processes that will be used to manage and maintain a dataset over its lifecycle.

(Collaborative) Documentation Tools

Open-source:

- [DMPTool](#) / [DMPonline](#): assist in creating data management plans (DMPs) required by funding agencies
- [RDMO](#): supports research projects in planning, implementing, and managing all aspects of research data management
- [Open Science Framework \(OSF\)](#): supports the entire research lifecycle and allows to manage, collaborate, and share research data, documents, and workflows.
- [Invenio](#): platform developed by CERN to manage digital assets. It provides solutions for repositories, document management, and digital libraries

Commercial, limited open-source version:

- [GitBook](#): host and build documentation using Markdown and Git

3) Future-proofing Files for Long-term Use

To avoid digital obsolescence, it's crucial to save your data in formats that are **open, lossless, and unencrypted**.

Type of Data	Preferred Data Format
Text	Unformatted text files (*.txt, *.asc, *.c, *.h, *.cpp, *.m, *.py, *.r, etc.) encoded in ASCII, UTF-8, or UTF-16, XML (.xml), HTML (.html), PDF/A (.pdf)
Spreadsheets	Comma-separated values file (.csv), OpenDocument Spreadsheet (.ods)
Images	JPEG Image Encoding family (.jpeg, .jpg), TIFF (.tiff, .tif), Portable Network Graphics (.png), Scalable Vectors Graphics (.svg)
Video	Material Exchange Format (.mxf)
Audio	Material Exchange Format (.mxf), FLAC (.flac), WAV (*.wav, uncompressed, pulse-code modulated)
Geospatial	Geography Markup Language (.gml), Keyhole Markup Language (.kml), ESRI Shapefile (.shp, .shx, .dbf), Geo-referenced TIFF (.tif, .tiff, .gtiff)
Numerical	NetCDF (.nc), HDF5 (.hdf5), CSV (.csv), JSON (.json)

Sources: [ETH Zürich](#), [4TU.ResearchData](#)

Data Management Policies and Acts

RDM Policy at Universities

e.g., RDM Framework policy at TU Graz defines roles and responsibilities for researchers, rectorate, faculties etc.

In addition: development of faculty-specific implementations



Framework Policy for Research Data Management at Graz University of Technology

Graz University of Technology (henceforth "TU Graz") is committed to the highest standards of research excellence and to maximising the academic and societal impact of its research and teaching. TU Graz recognizes and affirms the fundamental importance of professional and responsible research data management for maintaining the quality and integrity of research.

Preamble

This framework policy is motivated by the belief that good research data management (henceforth RDM for short) cultivates:

1. Best practice for ensuring that scientific arguments are reproducible and re-usable by researchers, society and industry in the long term.
2. Responsible performance, verification, evaluation and re-use of research through adequate documentation, preservation and availability of research data according to interoperable standards.
3. Better exposure of the work of researchers at Graz University of Technology, leading to affirmation of the quality of the research process as a whole.
4. Responsible managing of research data in accordance with the FAIR (Findable, Accessible, Interoperable and Reusable) principles¹, including the safe storage of personal data or protection of intellectual property developed by scientists across TU Graz.
5. Improved practices for meeting the demands of funders and publishers with respect to research data management and sharing.

In extension of, and alignment with, existing policies on Open Access², Intellectual Property³, and Research Integrity⁴, TU Graz hence adopts the following Framework Policy for Research Data Management at Graz University of Technology. The policy serves as an overarching description of rights and responsibilities across TU Graz as a whole, and is to be complemented by faculty-specific implementation strategies which take account of particular disciplinary requirements. These faculty specific strategies will be based on a template defined by the TU Graz RDM Policy Working Group and the Rectorate. These processes will be guided by the Digital TU Graz project – Chancenfeld Forschung team.

This is an aspirational policy. Implementation will take some years, and will depend on the availability of resources.

Data Management Policies and Acts

European Data Governance Act

A European Data Governance Act, which is fully in line with EU values and principles, will bring significant benefits to EU citizens and companies.

A key pillar of the [European strategy for data](#), the [Data Governance Act](#) seeks to increase trust in data sharing, strengthen mechanisms to increase data availability and overcome technical obstacles to the reuse of data.

The Data Governance Act will also support the set-up and development of common European data spaces in strategic domains, involving both private and public players, in sectors such as health, environment, energy, agriculture, mobility, finance, manufacturing, public administration and skills.

The Data Governance entered into force on 23 June 2022 and, following a 15-month grace period, is applicable since September 2023.

Data Management Policies and Acts

The European Data Act

What is the European Data Act?

The European Data Act makes more data available for use, and sets up rules on who can use and access what data for which purposes across all economic sectors in the EU.


According to Article 1, Subject matter and scope (proposal 23.2.2022):

1. This Regulation lays down harmonised rules on making data generated by the use of a product or related service available to the user of that product or service, on the making data available by data holders to data recipients, and on the making data available by data holders to public sector bodies or Union institutions, agencies or bodies, where there is an exceptional need, for the performance of a task carried out in the public interest:

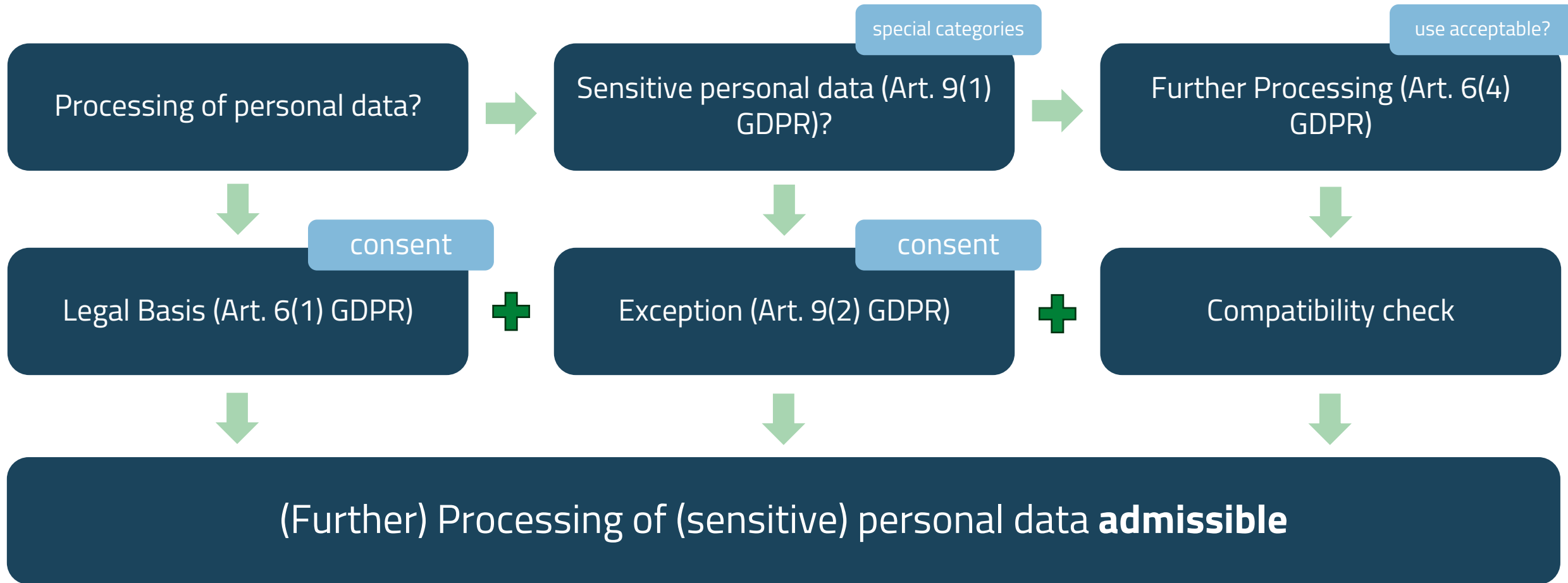
2. This Regulation applies to:

- (a) manufacturers of products and suppliers of related services placed on the market in the Union and the users of such products or services;
- (b) data holders that make data available to data recipients in the Union;
- (c) data recipients in the Union to whom data are made available;
- (d) public sector bodies and Union institutions, agencies or bodies that request data holders to make data available where there is an exceptional need to that data for the performance of a task carried out in the public interest and the data holders that provide those data in response to such request;
- (e) providers of data processing services offering such services to customers in the Union.

What is sensitive/confidential data?

- Data that requires a **high level of protection**, e.g.: names, medical records, proprietary information
 - Sensitive information can have many forms (text, image etc.)
 - Four broad categories:
 - a. PII (Personally Identifiable Information): data that can be used to identify an individual
 - b. PHI (Personal Health Information): health data, genetic data, biometric data about an individual
 - c. Financial information, e.g.: credit card numbers
 - d. Intellectual property, e.g.: patents
-  EU regulation: **GDPR** defines sensitive data and the responsibilities of processing this kind of data

Processing of Personal Data acc. to GDPR



Source: Wiedemann, Nils Torben, KI-Training nach dem neuen Datenrecht: Neue Chancen für KI made in EU?. Webinar. https://www.pairs-projekt.de/de/berichte/webinar-ki-eu-regulierung?utm_medium=email&_hsmi=79056579&_hsenc=p2ANqtz-8eiSaZ26Y8pzE-2h2jIZ_-_WbVnL7SqV5gutTBARVlolu6KyWHydBXWEsu_OCguTGEVYScKYm53UdUDrHg_-IGELqS6bbOnrRqUpi5711JsoKPcY&utm_content=79056579&utm_source=hs_email

Strategies to handle sensitive data

- Pseudoanonymization
- Anonymization
- Aggregation
- Differential Privacy
- Homomorphic encryption

Examples for Data Restriction Measures	
Remove direct identifiers	Can be stored separately from analytical datasets, Use pseudonyms or codes
Remove fine detail on geo-information	
Limit demographic information as much possible	
Code to remove detail	Rounding [age in 5-year intervals], Top/bottom coding (e.g., top income is > €100,000)
Statistical methods	Adding noise, suppressing some information (e.g., highly identifying cases)



When is data sufficiently anonymized?



It is ...

...complex

- **Legal** requirements are abstract, rarely provide concrete numbers

Example: National law (GER), §52 of the Metering Point Operation Act

- Anonymization can be achieved by combining data from at **least five** connection users
- Pseudonymization can be done using alphanumeric designations of the location for measuring, extracting, or consuming energy

- **Data processors:** find balance between protecting privacy and preserving data utility
- **Individuals/data subjects:** anonymization should provide a reasonable guarantee that their personal information will remain private



Coffee Break

until 10:45

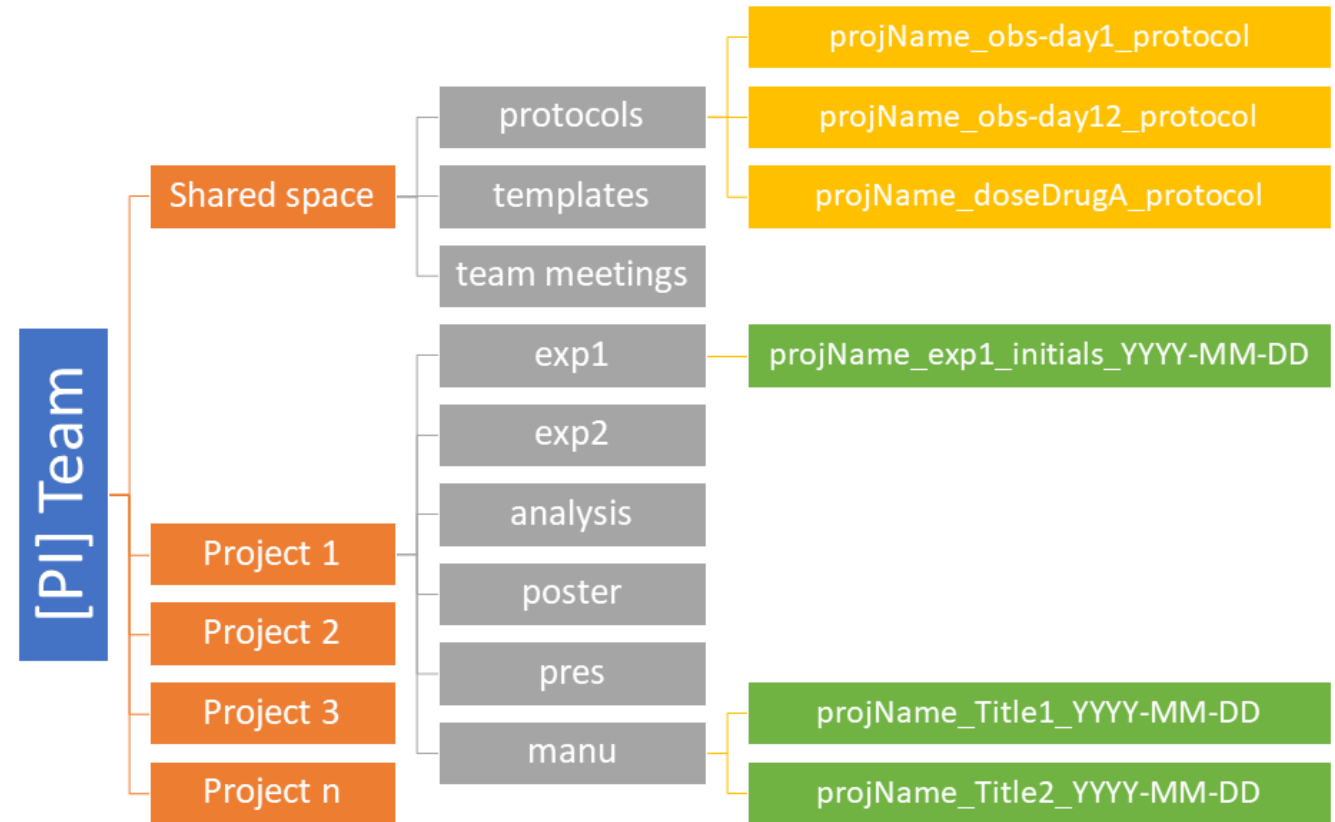
Exercise 1

Review of DMPs

Block 2: Version Control

Step 1: Systematic Organization

- **Customized Organization:** No universal system exists for data organization; the best system varies for each researcher and their data.
- **Identify Natural Groupings:** Determine natural data groupings like project, analysis type, or date for structuring.
- **Logical Folder Arrangement:** Organize folders logically to match how you'd search for content, prioritizing some groupings at higher levels.
- **Consistency is Key:** Ensure consistent application of the chosen system, habitually placing data where it logically belongs for easy retrieval later.



Briney KA, Coates H, Goben A (2020) Foundational Practices of Research Data Management. Research Ideas and Outcomes 6: e56508. <https://doi.org/10.3897/rio.6.e56508>

Step 2: File Naming

Imagine yourself nearing a deadline – which file names would facilitate your workflow?

1	File001.docx	
2	Experiment1_RawData_2023-11-15.csv	
3	SimulationCode_V2_EnhancedFunctionality.py	
4	Data_final.txt	
5	ClimateChange_TemperatureReadings_2015-2020_MonthlyAverages.xlsx	
6	NewSimulationCode.py	

Strategies for File Naming

- File names should be human- AND machine-readable
- Keep an untouched copy of the original file or raw data that won't be overwritten
- File naming convention:
 - 3 key pieces of information
 - Start of file name: most important sorting information (project name, experiment ID, location, date etc.)
 - Example: `Experiment1_RawData_2023-11-15.csv`
- Keep track of file versions:
 - **Basic:** captured in the file name, e.g.
 - `file_v02`
 - version date, ISO 8601 YYYYMMDD
 - append the version type, `_raw`, `_processed`, `_merged`
 - **Intermediate:** platform with version control built-in, e.g. GoogleDrive, Dropbox
 - **Advanced:** version control software, e.g. [git](#), [Mercurial](#), [Subversion](#)

Step 3: Write Conventions Down

- The DMP is recommended as a **living** document that describes all of the conventions decided on
 - Ease of reference for quick access to information
 - Assistance in remembering established guidelines and standards
 - Ensuring all project partners comprehend and meet expectations
- Consistent data conventions among collaborators can be a huge time-saver for everyone involved



Coffee Break

until 12:00

Block 3: Quality Assurance

Why Data Quality?

- Consulting firm Gartner said in 2021 that bad data quality costs organizations an average of \$12.9 million per year. Another figure that's still often cited is a calculation by IBM that the annual cost of data quality issues in the U.S. amounted to \$3.1 trillion in 2016.
- In research, findings based on wrong / flawed data cannot be trusted.

What is Quality?

- **Crosby:** Quality is conformance to requirements.
- **Juran:** Quality is fitness for use.
- **Deming:** Good quality means a predictable degree of uniformity and dependability with a quality standard suited to the customer.
- Quality is the degree to which performance meets expectations.
- **American Society for Quality:** Quality denotes an excellence in goods and services, especially to the degree they conform to requirements and satisfy customers.
- ...

Quality Assurance – Example from Pharma

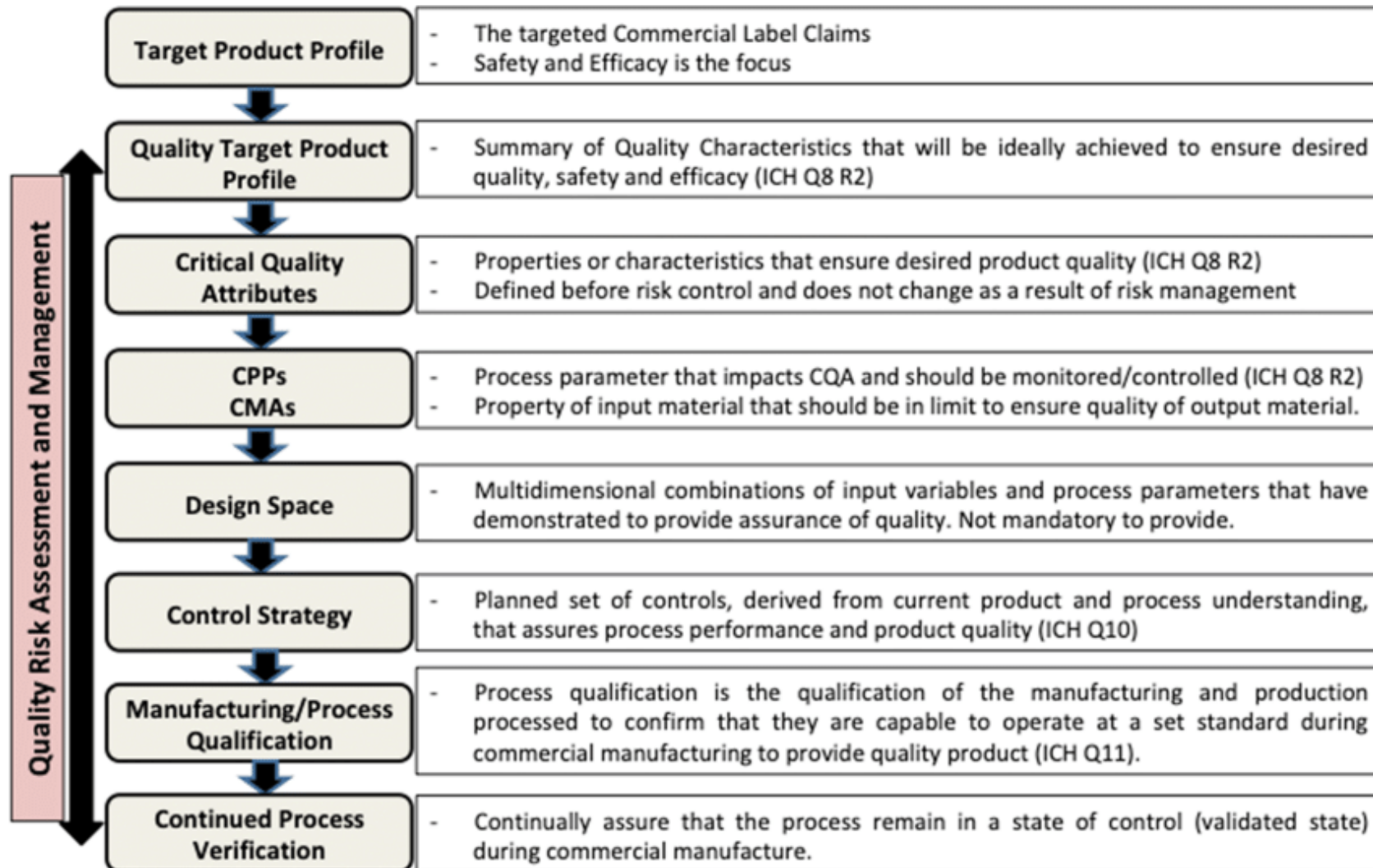
ALCOA+ is a set of principles that ensures data integrity in the life sciences sector. It was introduced by the main regulatory agency (FDA – the US Food and Drug Administration).

- Attributable
- Legible
- Contemporaneous
- Original
- Accurate
- Complete
- Consistent
- Enduring
- Available

Quality Assurance – Example from Pharma



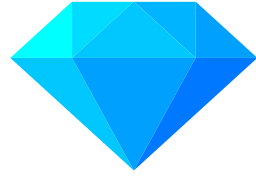
Quality Assurance – FDA recommendation



Exercise 2

OpenRefine

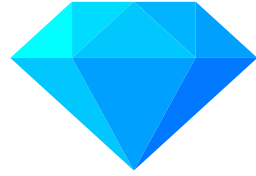
OpenRefine



OpenRefine is a powerful free, open source tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data.

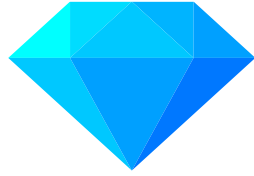
<https://openrefine.org/>

OpenRefine



- Work alone or in groups
- Go to : <https://openrefine.org/download>
- Install on your laptop (should also work without admin rights)
- Use one of your .csv files, or the FIFA example dataset on GitHub
- Create a new project

OpenRefine



- Use the **facet function** to explore your data (arrow, facet)
- Use the **transform function** to change your data (e.g., to blank, to upper case)
- Use the **sorting function** to arrange your data
- Use the **view function** to display only relevant columns
-

Check this tutorial: <http://datacarpentry.org/OpenRefine-ecology-lesson/index.html>

Contact us for for individualized data consulting and DPM support



Sarah Stryeck

Data Steward @ DIO

sarah.stryeck@dataintelligence.at



Lisa Nußbaumer

Tech Lead @ DIO

lisa.nussbaumer@dataintelligence.at

Thank you!

Feedback:
<https://forms.office.com/e/RJa00Fngkg>

Feedback - 1. Deep Dive Workshop
Data Stewardship IKTdZ



Back-up Slides

Recommended Components of a DMP

- Project title, duration, and research hypotheses
- Reuse of existing data
- Data to be collected:
 - Data types and formats
 - Expected storage space requirements (as precise as possible)
 - Methods of data collection
 - Hardware and software used
- Backup
- Folder structure and naming conventions
- Documentation and metadata
- Data exchange/access (within the project or with third parties)
- Legal Aspects
 - Data protection
 - Copyright
 - Usage rights
 - Licensing
- Data publication and reuse
- Storage and (long-term) archiving
- Responsibilities
- Costs/Budget