HPC.NRW

Introduction to RDM

RDM in HPC - Challenge or Chance?

Online | 26.02.2024

DAS KOMPETENZNETZWERK FÜR HOCHLEISTUNGSRECHNEN.

Introduction to RDM

Marcel Nellesen (RWTH Aachen University)

RDM in HPC - Challenge or Chance?

RDM Basics

HPC.NRW

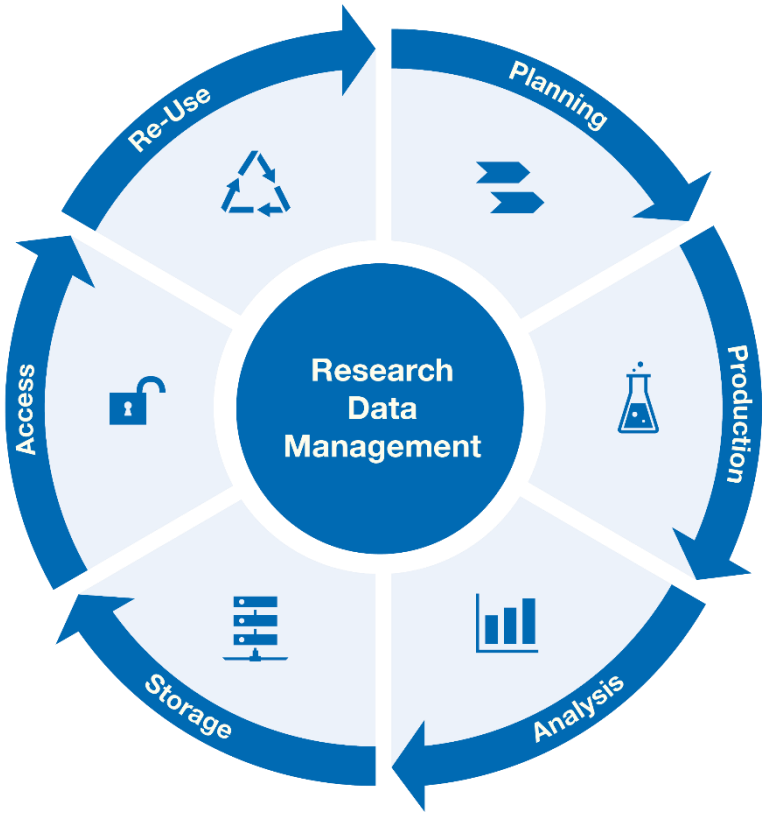
- When should I start with RDM?
  - Short Answer: NOW!!!
  - Long Answer:
    Ideally before you start with the project. However it is never too late.

RDM Basics

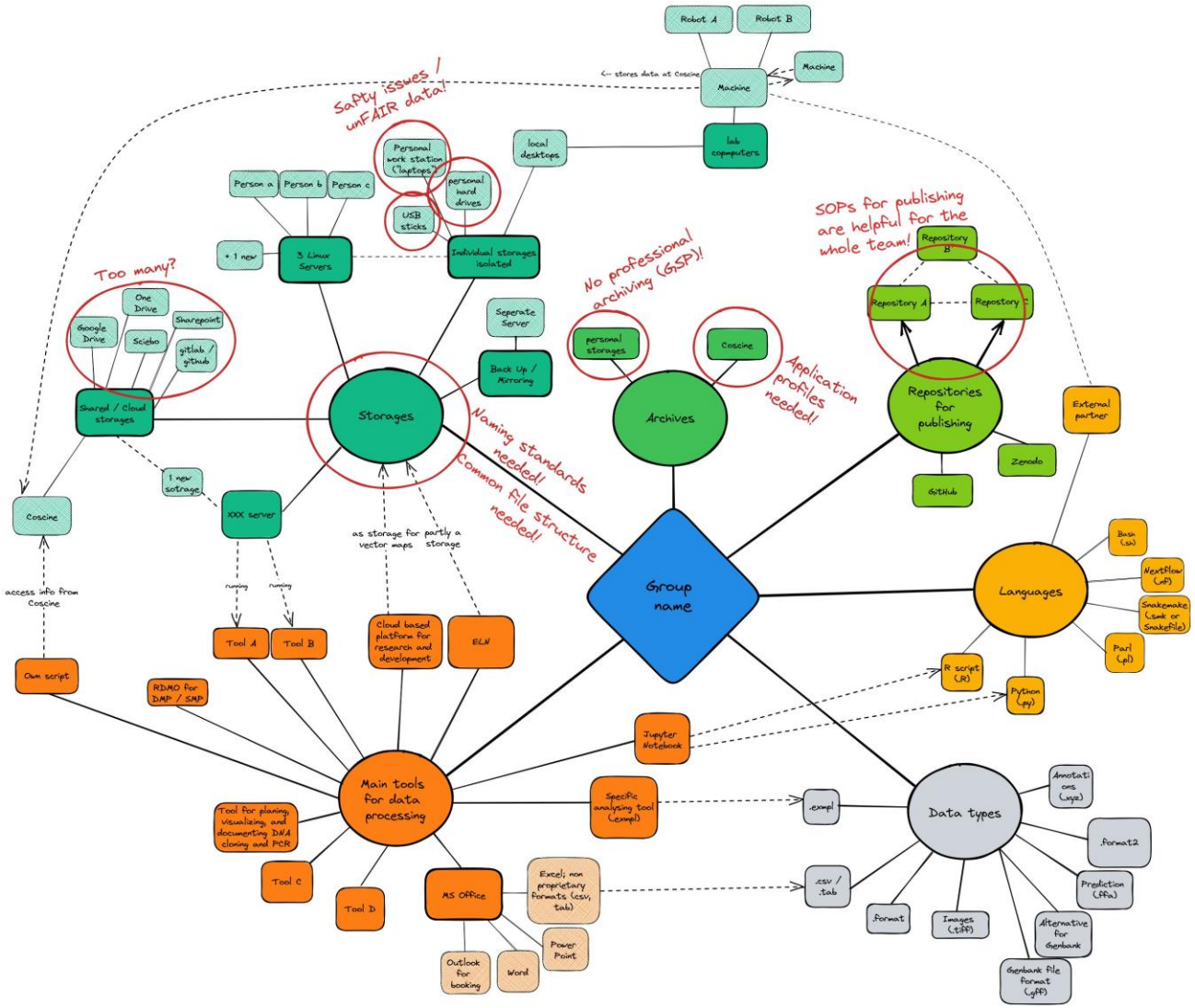
HPC.NRW

- Research projects are a long term commitment

- Data is constantly created
  - First draft / proposal
  - Setup
  - Measurement
  - Analysis
  - Reports
  - Reusage



Know your environment



FAIR – what is fair?

HPC.NRW

- FAIR
  - Findable
    - For humans and for machines
  - Accessible
    - Providing information and infrastructure to access the data
  - Interoperable
    - Interaction with applications, for analysis, storage and processing
  - Reusable
    - Metadata and data need to be well described to allow reproducibility and reusage in other contexts

FAIR – what is fair?

HPC.NRW

- **Findability**
  - (meta)data are assigned a globally unique and persistent identifier
  - Data are described with rich metadata
  - Metadata clearly and explicitly include the identifier of the data it describes
  - (meta)data are registered or indexed in a searchable resource

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

FAIR – what is fair?

HPC.NRW

- **Accessibility**
  - (meta)data are retrievable by their identifier using a standardized communications protocol
    - The protocol is open, free, and universally implementable
    - The protocol allows for an authentication and authorization procedure, where necessary
  - Metadata are accessible, even when the data are no longer available

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

FAIR – what is fair?

HPC.NRW

- **Interoperability**
  - (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
  - (meta)data use vocabularies that follow FAIR principles
  - (meta)data include qualified references to other (meta)data

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

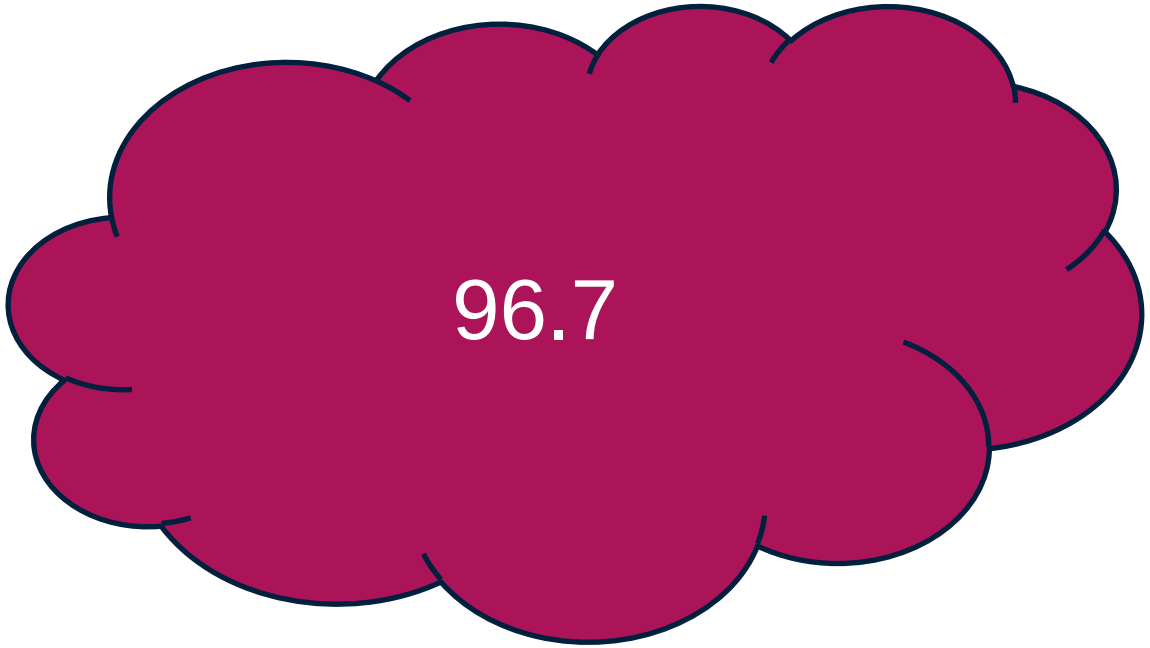
FAIR – what is fair?

HPC.NRW

- **Reusability**
  - (meta)data are richly described with a plurality of accurate and relevant attributes
    - (meta)data are realized with a clear and accessible data usage license
    - (meta)data are associated with detailed provenance
    - (meta)data meet domain-relevant community standards

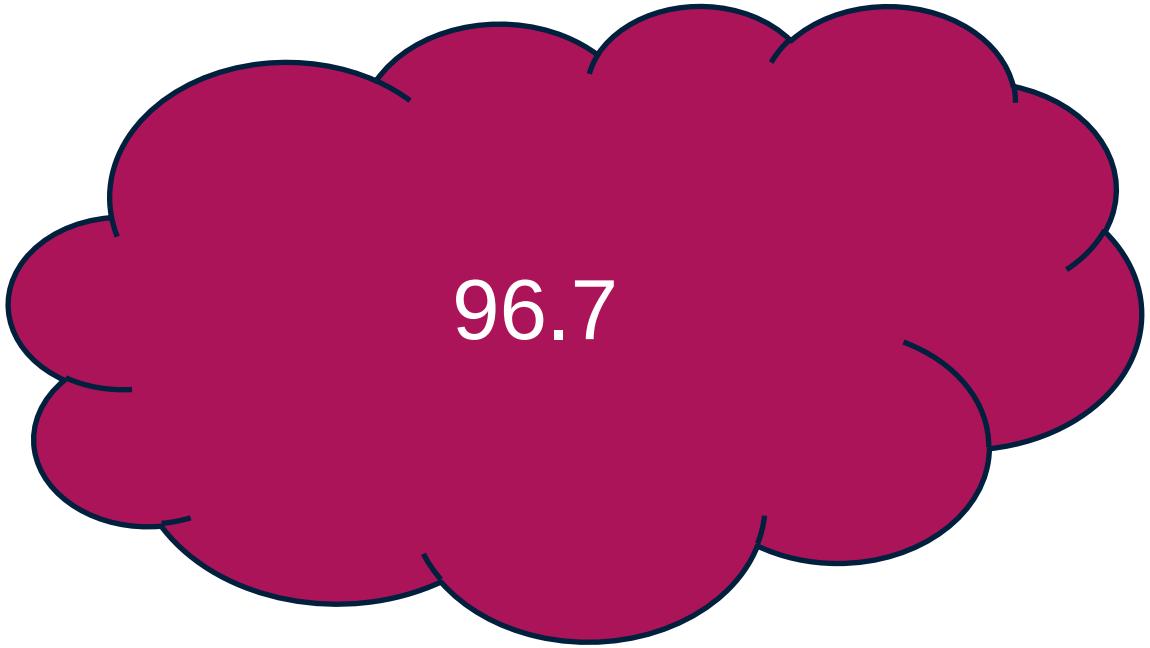
Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

What is metadata?



What is metadata?

Unit:
Degree Celsius

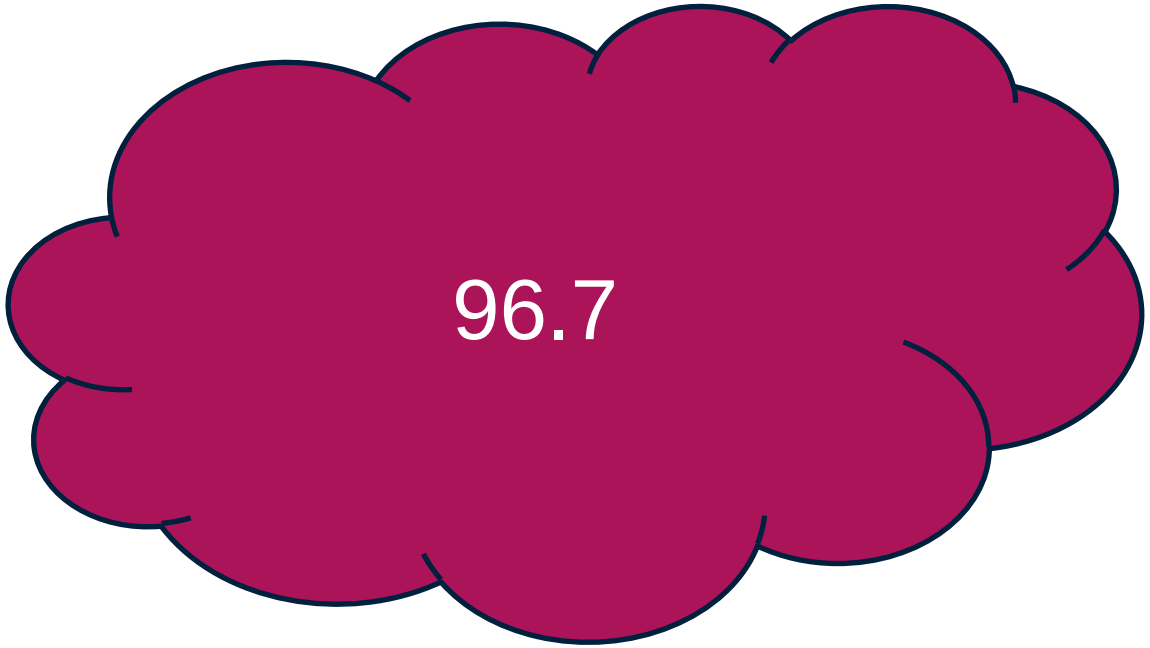


Date:
26 April 1986

Time:
01:23

What is metadata?

Unit:
Degree Celsius



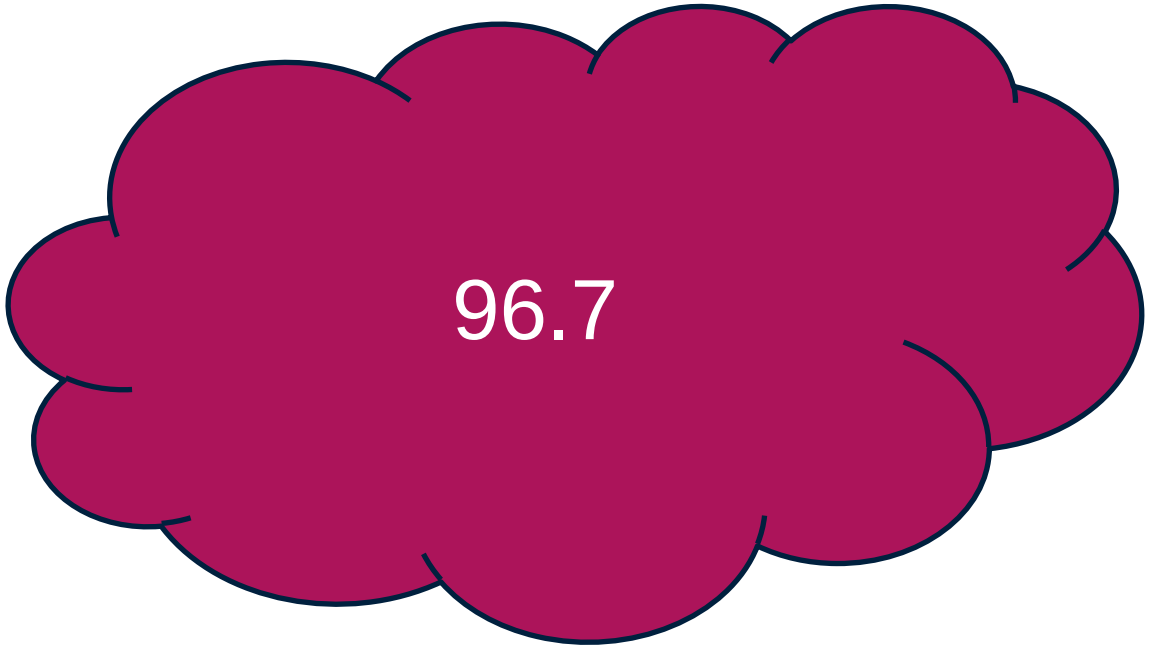
Date:
26 April 1986

Time:
01:23

Sensor:
CPU Temperature

What is metadata?

Unit:
Degree Celsius



Date:
26 April 1986

Time:
01:23

Location:
Chernobyl
Reactor Block 4

Sensor:
CPU Temperature

What is metadata?

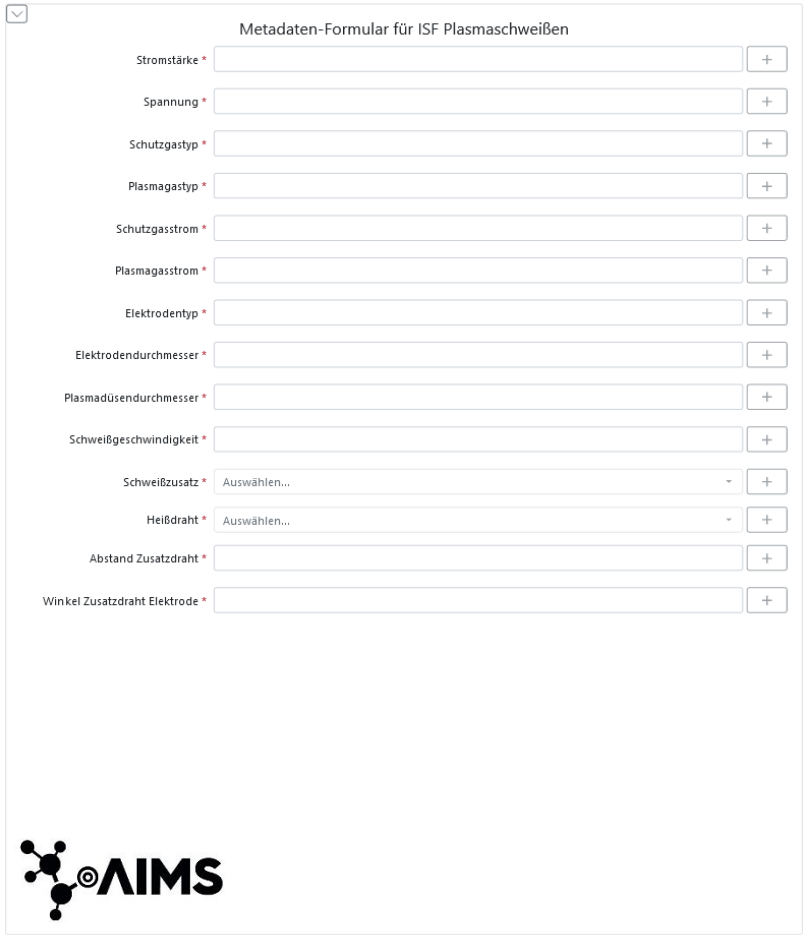
HPC.NRW

Metadata gives meaning to data!

What is metadata?

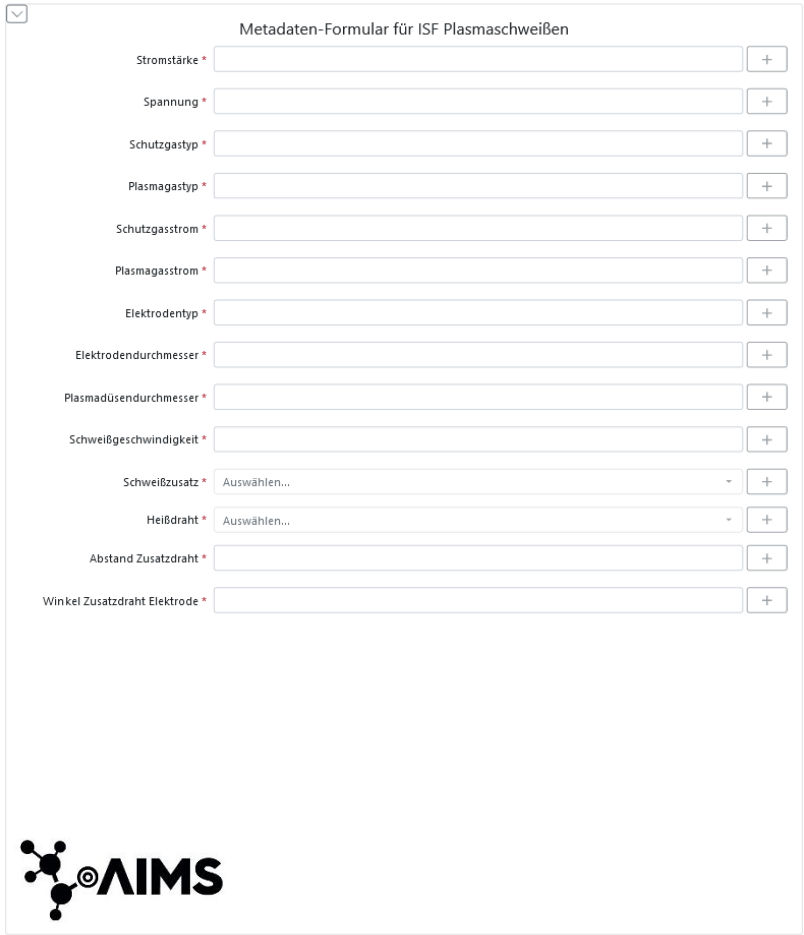
Application profiles / Metadata schema

- Every field of research is different
- Application profiles must be customizable
- Researchers and institutes must be able to create their own profiles
  - If the application profile does not fit the workflow, it will not be used correctly
  - If the application profile is not relevant, it will not be adapted
- Using the application profiles must be easy
  - Entering metadata for ~50 different fields by hand is tedious
  - Support of automation
- Reduce the amount of work for the researchers



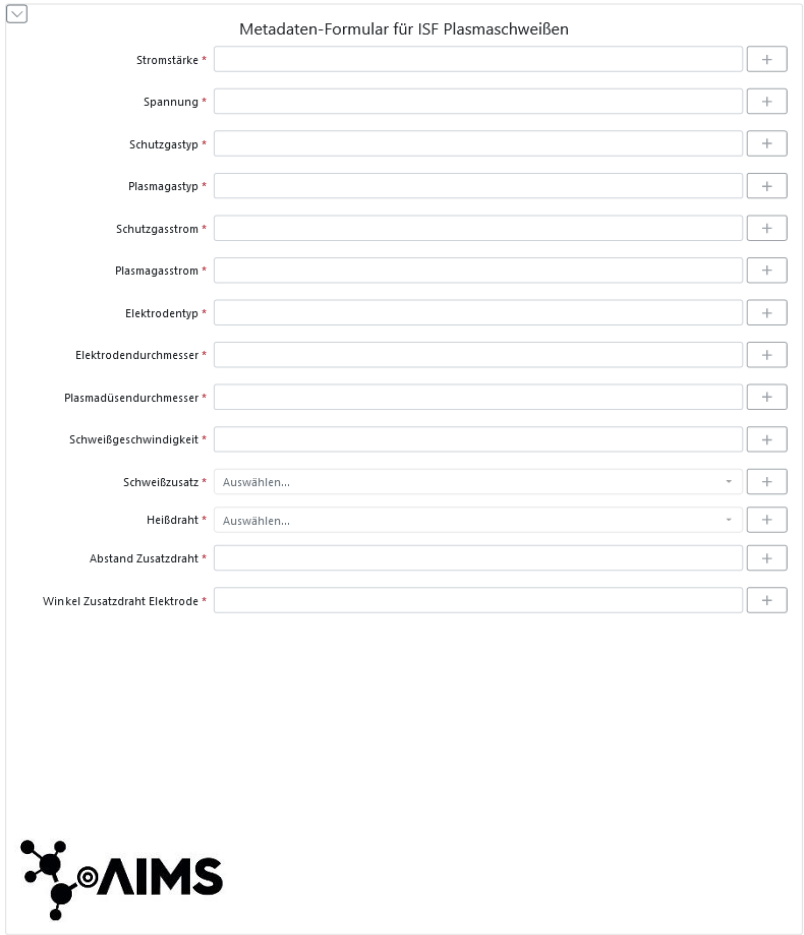
Application profiles / Metadata schema

- Create a definition of the required metadata
  - Usually called Application Profile or Metadata schema
- General description of the metadata fields
- Additional information on
  - Expected data types
  - Length of fields
  - Ranges of input fields
- Usage of vocabularies
  - Closed list of possible values (e.g., DFG categories)
- Multitude of input fields
  - For example, each paper must have at least one author but can have multiple authors



Application profiles / Metadata schema

- Metadata can be validated with the application profiles
  - Complete information
  - Quality standard
- Well defined metadata
  - Interoperability
  - Allows reuse and comparison
- Extensive set of metadata for research data and processes
  - Analyze
  - Share
  - Reuse



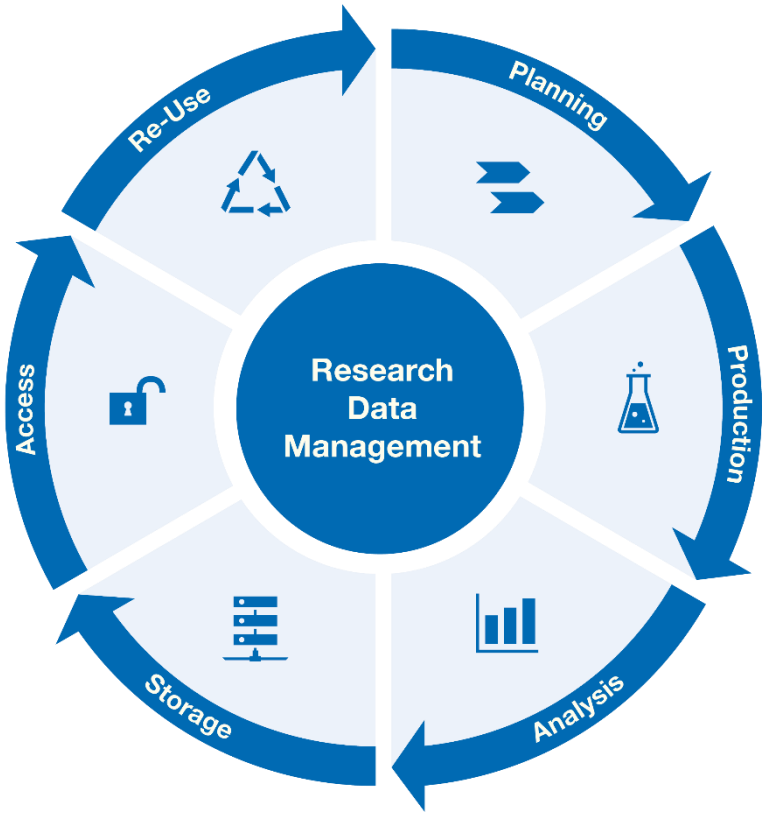
Showcase: DMP / RDMO

HPC.NRW

- Short definition:
  - “A data management plan (DMP) structures the handling of research data of a scientific project. It describes how the data is handled during the term and after the end of the project.“

    Source: https://forschungsdaten.info/themen/informieren-und-planen/datenmanagementplan/ 2023-09-17.

- A document which…
  - …describes the handling of data from the collection to archiving, as well measures to ensure availability and reusability
  - …is prepared before the project begins and is updated regularly („living document“)
  - …describes the whole data life cycle of research data

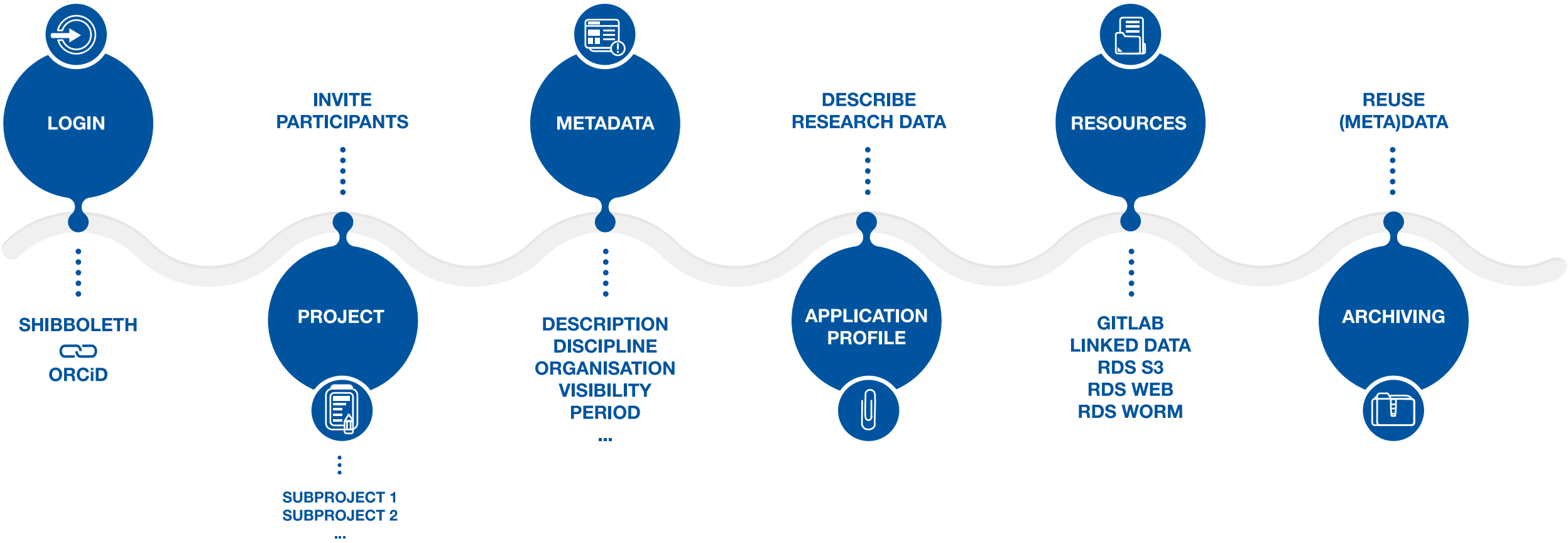


Showcase: Coscine

- Problems for Researchers:
  - Limited knowledge about RDM and FAIR principles
- Additional work before publishing
  - Metadata not complete
  - Repeat experiments
  - Ensure Reuse and Interoperability
- Solution:
  - Integration of RDM in daily workflow
  - Use a platform which helps to follow FAIR principles

Showcase: Coscine

HPC.NRW



Coscine

Stay tuned…

HPC.NRW

- Regular talks on RDM topics

- General tools
  - Coscine
  - RDMO / DMP
  - Electronic Lab books

- Domain specific tools
  - Nomad
  - …

Thank you for your attention!