NHR and NFDI

Bridging the Gap

IT Center | RWTH AACHEN UNIVERSITY

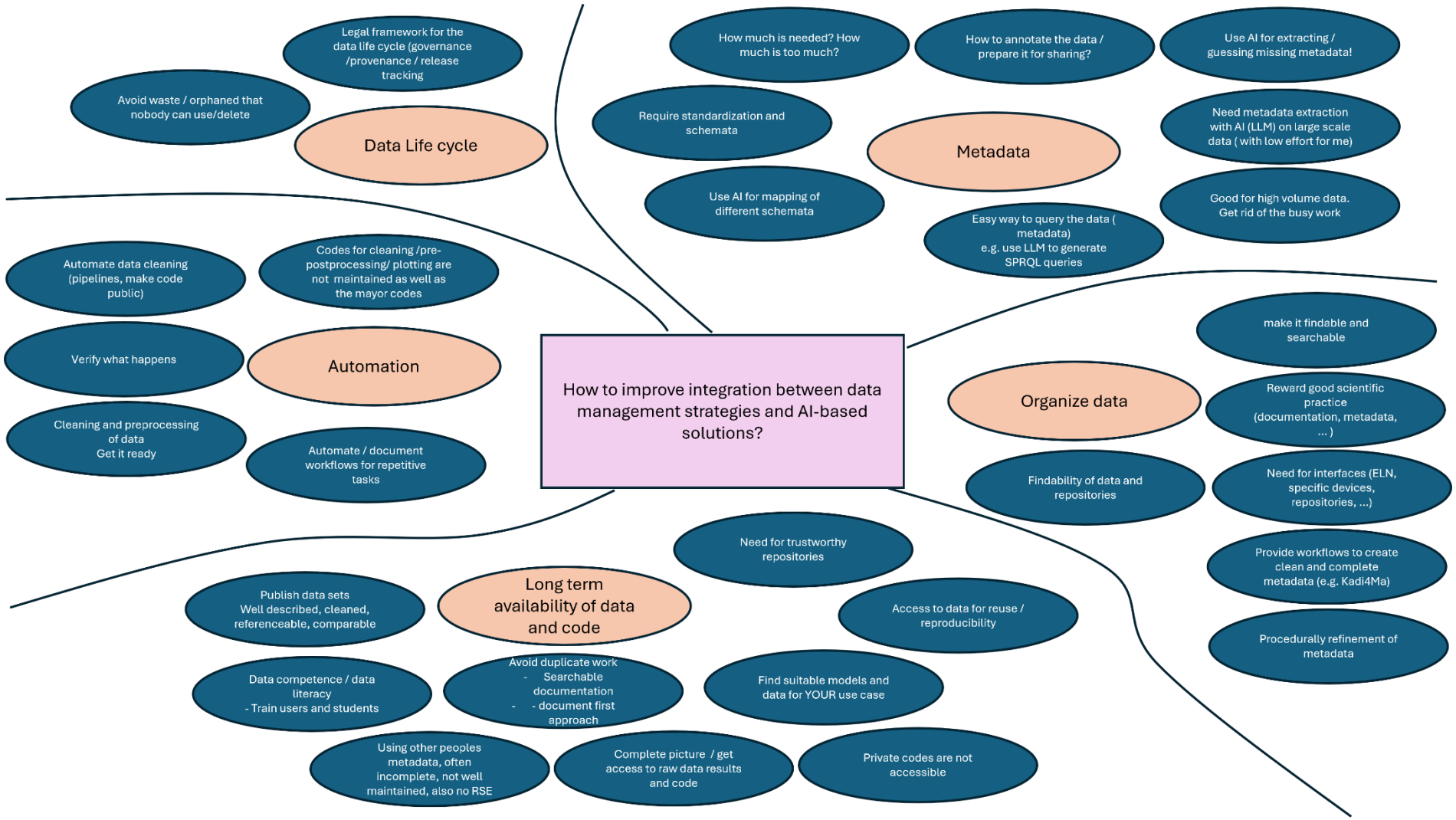
NATIONALES HOCHLEISTUNGS RECHNEN

What is the NHR?

- Tier 2 High performance computing in Germany
- 9 Centers – 13 Universities
- Each Center has different scientific domains it excels in
- Expertise in HPC and AI

IT Center | RWTH AACHEN UNIVERSITY

NHR4CES

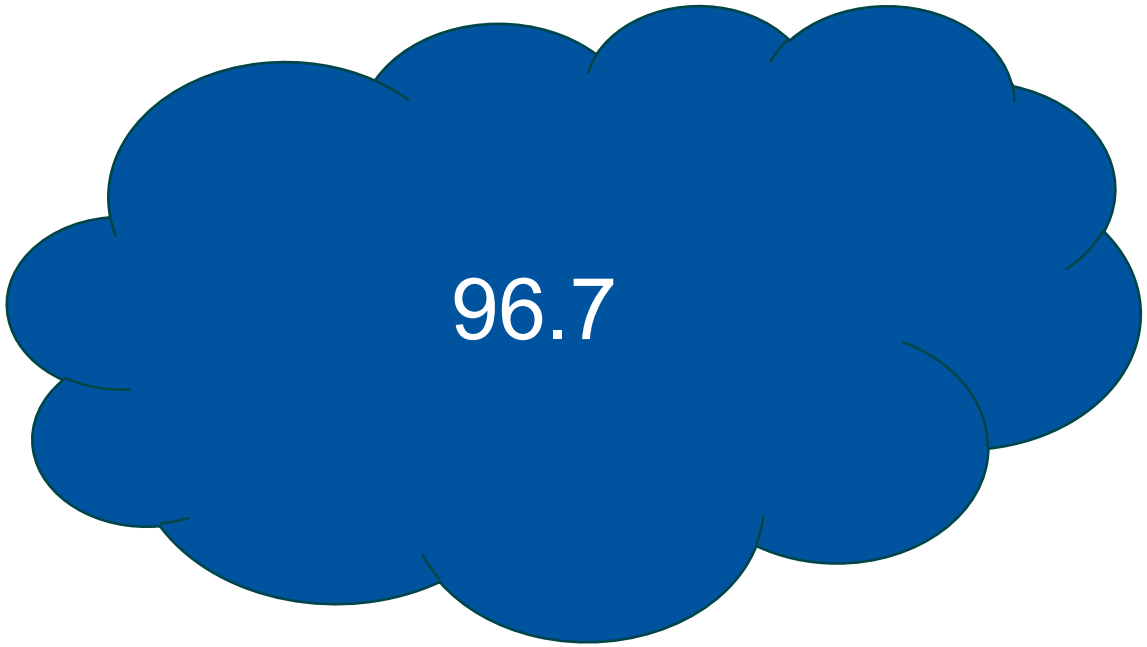


IT Center | RWTH AACHEN UNIVERSITY

What is Metadata?

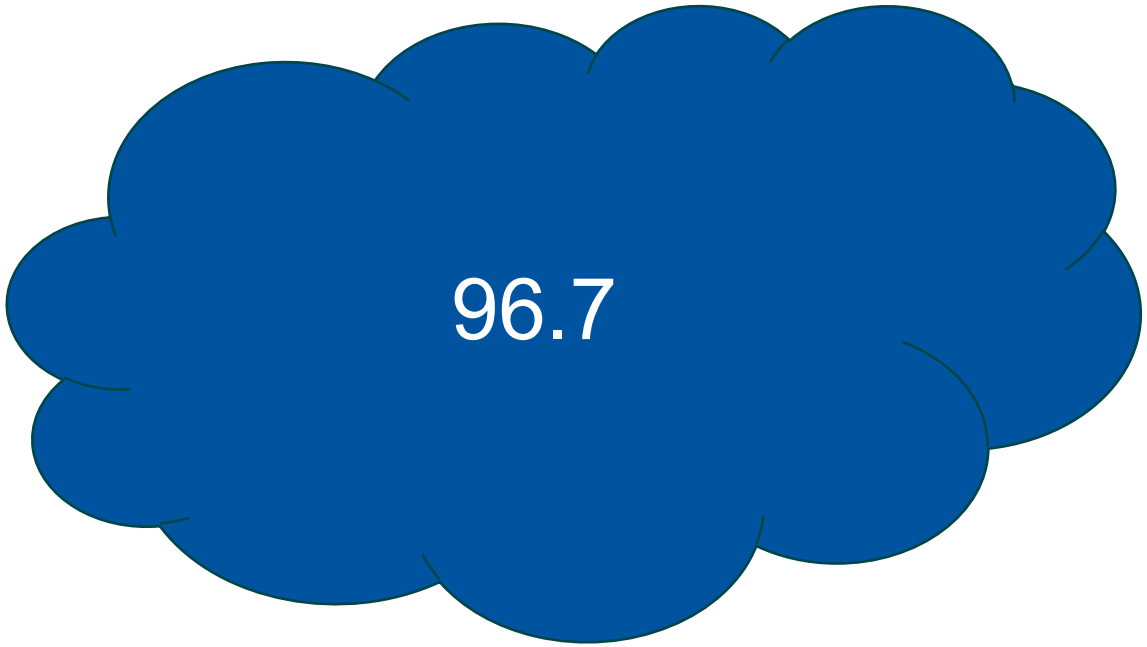
(And why is it so important?)

What is Metadata?



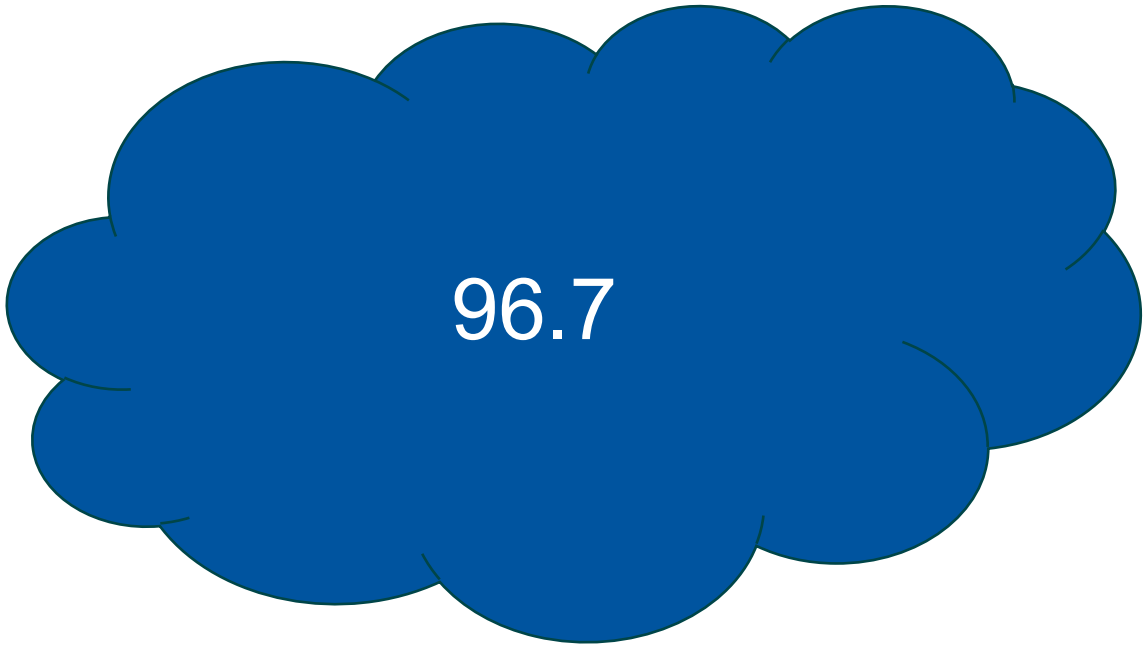
What is Metadata?

Unit:
Degree Celsius



What is Metadata?

Unit:
Degree Celsius

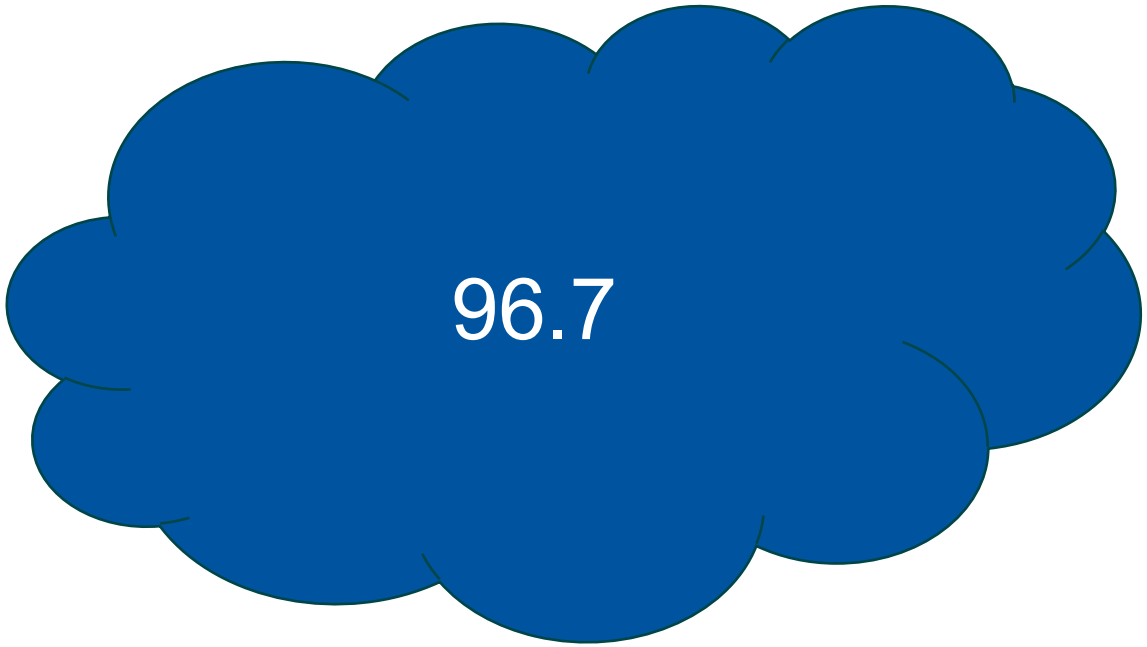


Date:
26 April 1986

Time:
01:23

What is Metadata?

Unit:
Degree Celsius



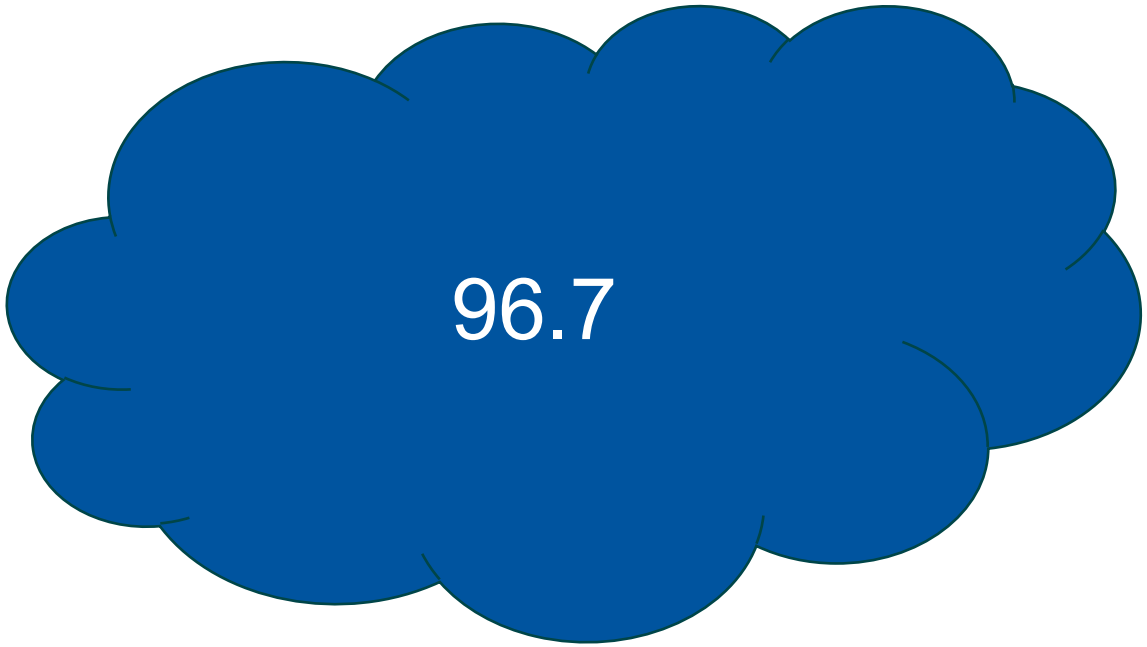
Date:
26 April 1986

Time:
01:23

Sensor:
CPU Temperature

What is Metadata?

Unit:
Degree Celsius



Date:
26 April 1986

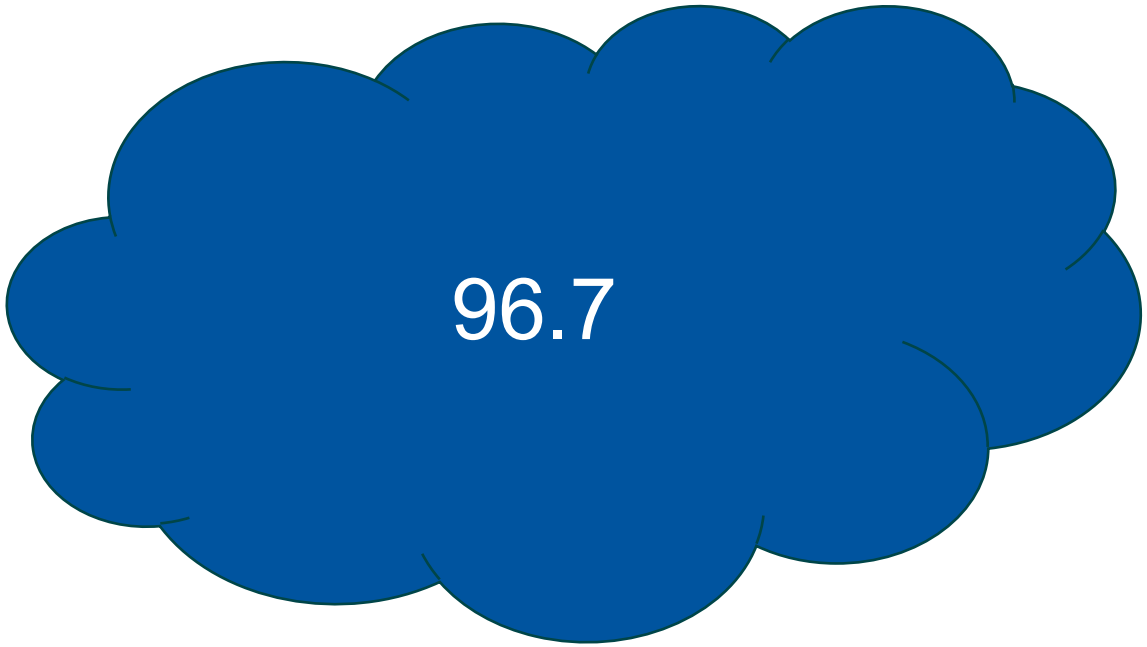
Time:
01:23

Location:
Chernobyl
Reactor Block 4

Sensor:
CPU Temperature

What is Metadata?

Unit:
Degree Celsius



Date:
26 April 1986

Time:
01:23

Location:
Chernobyl
Reactor Block 4

Sensor:
CPU Temperature

Metadata gives data meaning and makes it valuable

You have a fever…

You have a fever…

What is the NFDI?

- A nationwide, interdisciplinary initiative funded by the German Federal Ministry of Education and Research (BMBF) to build a sustainable, federated research-data infrastructure for all scientific disciplines in Germany.

- Important parts:
  - Consortia: 26 domain specific alliances of researchers
  - Sections: 6 Cross-sectional topics across consortia
  - Base Services: NFDI-wide services for common, interoperable solutions

The consortium for the engineering consortium within NFDI

- NFDI4ING is one of the first and biggest consortia within NFDI, established in 2017 with now over 50 active members and participants. Bringing together the **diverse engineering communities** to develop **solutions for research data management**, NFDI4ING leverages self-contained data entities embedded in knowledge graphs, and implements FAIR digital objects as well as established vocabularies to ensure that research data is **FAIR and AI-ready**. These developments are made **available to all engineering scientists** and ensure German- or European-wide **scalability**.

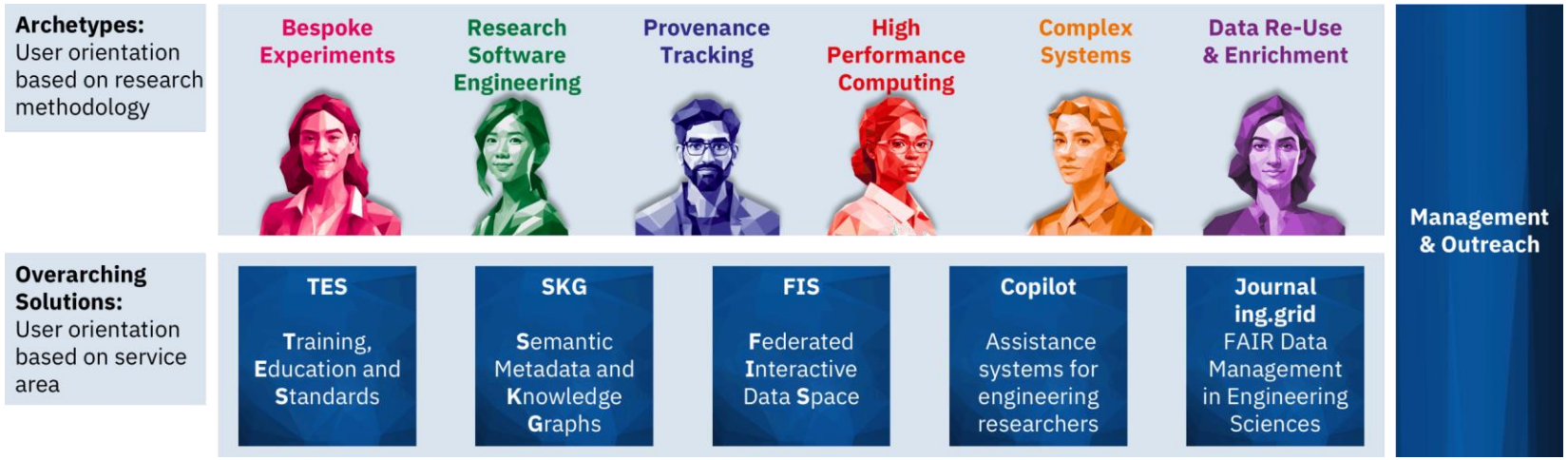


Image: https://www.nfdi.de/nfdi4ing-2/

NFDI4DataScience

- Within the NFDI there exists a consortium dedicated to Data Science and Artificial Intelligence, **NFDI4DS**. Its goal is the development, establishment, and sustainment of a **national research data infrastructure for the Data Science and AI community** in Germany. It works towards increasing the transparency, reproducibility, and fairness of Data Science and AI projects, by **making all digital artifacts available, interlinking them, and offering innovative tools and services**. In the initial phase, the focus is on four areas: **language technology, biomedical sciences, information sciences, and social sciences**.

NFDI4ING and others

- Many, if not most or even all, consortia now work with AI in some form

- In NFDI4ING this is focused on **AI assistance systems** for now
  - A chatbot for questions on RDM and NFDI4ING
  - JupyterHub enhancement through code quality checks and AI-assisted development support

- AI interfaces in some services for these systems

- Metadata from repositories provide **input for Data Science**

RDM supported repositories

- FAIR data requires said data to be properly **annotated with semantically rich metadata**. From this follow **requirements for repositories and storage** in general that need more than merely cloud space for said data. If metadata is instead supported and enforced by the repository and readily accessible, this information can **more easily be used in the processing** of said data. This should be especially **useful for Machine Learning** approaches that may lead to new insights from consistently annotated metadata.

Coscine & RADAR

- NFDI4ING provides two main repositories for storage with metadata

- Coscine is the solution for **hot data** with editing of metadata throughout the project
  - Metadata profiles can be created by users
  - An approval process aims to ensure that data is annotated and not just dumped
  - Data and metadata accessible via REST API

- RADAR4ING is for **cold data**, i.e. archival and publishing of a dataset
  - Metadata cannot be freely edited, changes require a new dataset instead
  - Similar approval process that focuses on making sure it belongs to the engineering sciences
  - Accessible via REST API

File Transfer Service

- Research data, with the potential volumes of data involved, may have requirements for transfers that are not fulfilled by common file transfer solutions. Besides the **scale required** to transfer large volumes of data in a timely manner, the larger the transfer, the more likely it is that an **error occurs, and part of the transfer must be repeated**. Both this scalability and reliability are offered by the File Transfer Service developed at CERN for the WLCG. For our needs at RWTH and NFDI4ING we are setting up **our own FTS instance** to get data to where it is needed for processing or storage.

HPC implications

- Users **submit transfers** to the service and must **trust it with access** to a storage
- Storage must support a **protocol supported by FTS** (i.e. S3, WebDAV)
- Getting data to or from an **HPC system requires file system access via these protocols**
- Solution still an open question

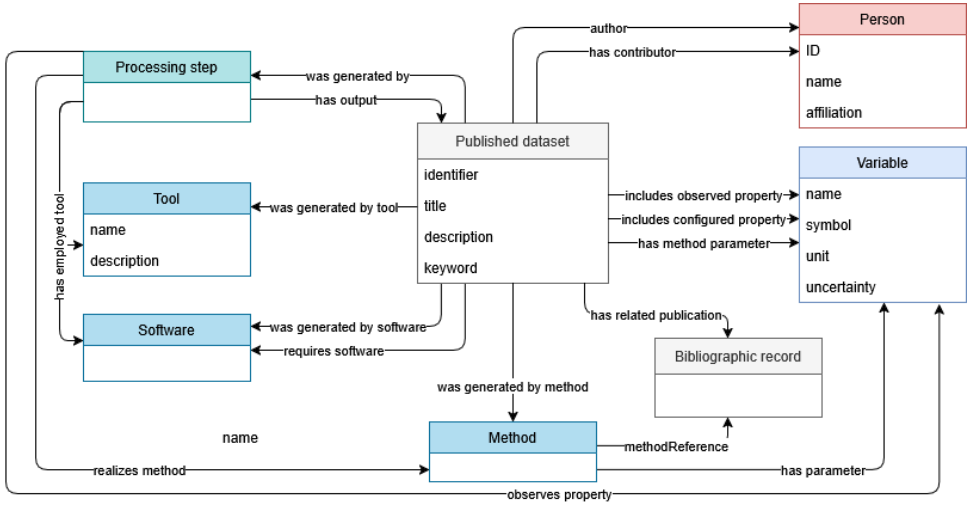
It‘s just semantics, right?

Image: https://xkcd.com/1860/

It‘s just semantics, right?

Common Information Model

- While being able to annotate data with metadata and even enforce its use technically are useful features within the repositories, **how is metadata formed and understood**? In the first place, **what is good metadata** to have and **what do I describe** with the metadata? **What does my description mean**? If a researcher wants to combine the data from multiple datasets, is there consistency or is a custom mapping required? To address all of these problems, the **semantics of the metadata must be agreed upon**. For this purpose, a Common Information Model for engineering is being developed and provided.



It‘s just semantics, right?

Common Information Model

- Based on **models and ontologies within and outside NFDI4ING**
- Use **standards** for documenting and providing data
- **Adaptable** to specific requirements and **integrated** into various services
- Basis of the **NFDI4ING knowledge graph**

Findability

The insight needle in the data haystack

- This treasure trove of data accumulated in the repositories has limited use if it cannot be discovered by interested parties. Findability is therefore an important aspect of reusable research data. To both **search the NFDI4ING's own repositories**, but also those **existing outside of the consortium**, several solutions are provided.

NFDI4ING AI assistant

ORKG

Data Search Index

re3data.org

Research Software Finder

FAIR Digital Object (FDO)

- FAIR digital objects are the **technical implementation of the FAIR principles**. A digital object can be data, metadata, documents, software, semantic assertions, and so on. FDO is **machine-readable and machine actionable** unit of data **identified by a PID** and **described by a record**, which includes the most critical information about the **object‘s context** and **possible operations**. The context should be a minimal set of metadata for discoverability and resuability. Operations define access to the data and metadata. Thus long-term interoperability is eased.

FAIR on the technical level

Implementing FAIRness in NFDI4ING

- FDOs are not a standardized concept yet
- NFDI4ING is only starting to implement the ideas
- Coscine resources given PIDs
- Further plans to extend this to all data and metadata within resources

An Overview

- Base4NFDI integrates and establishes basic services as common, **interoperable** solutions. Already existing services are adapted or extended to be usable for researchers from other disciplines. This way, parallel developments are avoided - since many scientific fields have similar requirements for many research data management services.

base4nfdi Basic Services for NFDI

Data Management Plans

- Centralised service for **data management plans** (DMPs) and **software management plans** (SMPs) across NFDI
- Hosting of the **open-source DMP tool RDMO**, coordination of template creation, support in standardisation and interoperability of templates, guidance and support for consortia staff
- Establish and foster the NFDI-wide use and application of **machine-actionable** DMPs (maDMPs) and maSMPs and analogues to maximise the benefits of their widespread adoption

Research Data Management Training

- Modular collection of **foundational RDM training materials** (train-the-trainer approach)
- Develop **concepts and training formats, didactic concepts**, and **certification**
- Strengthen the **network of RDM experts**

Persistent Identifier Services

- Consolidate and evolve the **PID service landscape** within NFDI at all **levels**: **technical**, **organisational**, **methodological**, and in **communication**.

Terminology services

- Aims to facilitate consensus-building and interoperability of services across disciplines to achieve a **shared knowledge representation**
- Provide a harmonized and centralized **single point of access** to other (NFDI) terminology services
- Is a cross-domain service for the **provision**, **curation**, **development**, **harmonization** and **mapping of terminologies**

Knowledge Graph Infrastructure

- Establish a **central, reusable knowledge graph infrastructure** (KGI) to improve interoperability within the research domain
- Provide essential components like a **knowledge graph registry** and services for accessing knowledge graphs across projects
- Empower research communities to create **decentralized knowledge graphs** using standardized methods and technologies
- Support optimised **knowledge graph creation** and foster **ontology harmonization** to support FAIR data principles and international initiatives like the European Open Science Cloud (EOSC)

Creating a software marketplace for NFDI

- Establish a **central marketplace for research software**, including **metadata enrichment** and support of the FAIR principles
- Establish a **dynamic search and recommendation system**

base4nfdi Basic Services for NFDI

Centralized Service for Jupiter Notebooks in NFDI

- Establish a **central Jupyter Hub** that is accessible via a single entry point
- Provide a foundational set of **computing and data resources**
- Foster the creation of **transparent** and **reproducible** scientific results
- In NFDI4ING ease use through **curated engineering images**

base4nfdi Basic Services for NFDI

Identity and Access Management

- A single way to “sign on” to all NFDI services and beyond
- Develop a **scalable**, **federated** IAM system that connects existing and emerging **authentication and authorization** solutions
- Ensure interoperability and integration with infrastructures like the European Open Science Cloud (EOSC) and other international research platforms

ACCT4nfdi
Accounting Infrastructures for the German National Research Data Infrastructure

Accounting and reporting infrastructures

- Build a **federated, standards-based** accounting system for the whole NFDI ecosystem.
- Transparent tracking of storage, compute, AAI, data-license services.
- Enables reliable cost estimation, planning, and long-term sustainability.

base4nfdi Basic Services for NFDI

IT Center | RWTH AACHEN UNIVERSITY

Funding

- Initialization (duration 1 Year)
  - KGI4NFDI
  - Nfdi.software
  - RDMT4NFDI
  - Accounting4NFDI
- Integration (duration 2 Year)
  - PID4NFDI
  - TS4NFDI
  - DMP4NFDI
  - jupiter4NFDI
- Ramp-Up (duration 3 Year)
  - IAM4NFDI

base4nfdi Basic Services for NFDI

Bridging the Gap

NHR to NFDI

- High-performance compute clusters with scalable storage tiers
- Dedicated support & consulting on HPC optimization, data-intensive workflows, and AI
- Experts from different scientific domains that work in HPC
- Focus on automation of workflows

Image: https://www.arc.ed.tum.de/sd/forschung/built-demonstrators/bridge-the-gap/

NFDI to NHR

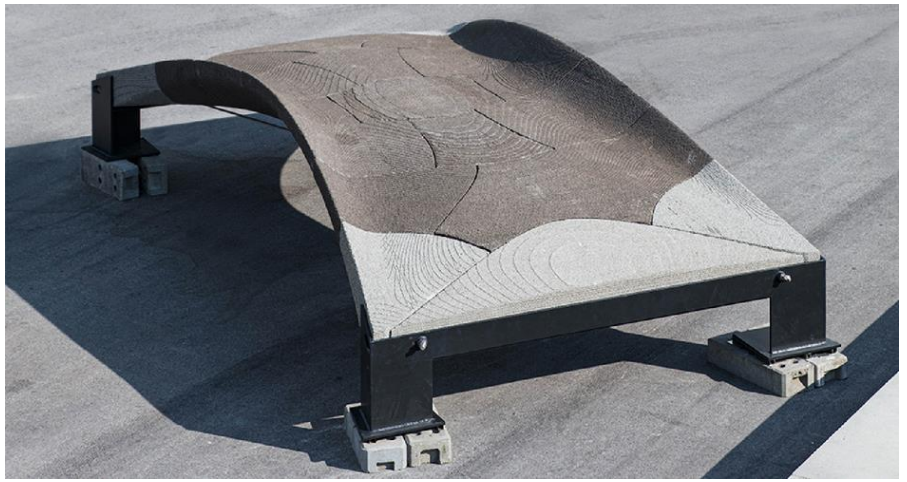
- Many researchers with a focus on annotating their research data with metadata
- Expertise on RDM from different scientific domains
- Standardization of annotation eases Data Science and AI use
- Processing of data needs corresponding computing resources

Image: https://www.arc.ed.tum.de/sd/forschung/built-demonstrators/bridge-the-gap/

Conclusion

A good mix of similarities and synergies

- NHR brings the power, NFDI brings the tasks
- Standardization of metadata schemata and automated extraction
- Strong interest in utilizing AI on a large scale on rich metadata
- Offer joint trainings
- Support researchers during the entire research project



Thank you very much for your attention!