Von Bits zu FAIRen Daten

Aus der Praxis an der RWTH Aachen und im Land NRW

Marius Politze
0000-0003-3175-0659

RDS and RDS.NRW are funded by Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen (MKW:124-4.06.05.08-139057, DFG: INST222/1261-1). DataStorage.nrw is funded by Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen (MKW: 214-76.01.09-7-7937 DFG: INST 222/1530-1). Coscine.nrw is funded by Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen as a central Service.
The conceptual work was supported with resources granted by NFDI4Ing, funded by Deutsche Forschungsgemeinschaft (DFG) under project number 442146713, NFDI-MatWerk, funded by Deutsche Forschungsgemeinschaft (DFG) under project number 460247524 and FAIR Data Spaces, funded by the German Federal Ministry of Education and Research (BMBF) under funding reference FAIRDS11.

IT Center @ RWTH Aachen University

Mission

IT-Service Provider for RWTH Aachen University

- From network infrastructure to HPC systems
- E-Learning and SLCM
- Responsible to support Research Data Management at RWTH

Mission@NRW

IT-Service Provider for NRW Universities

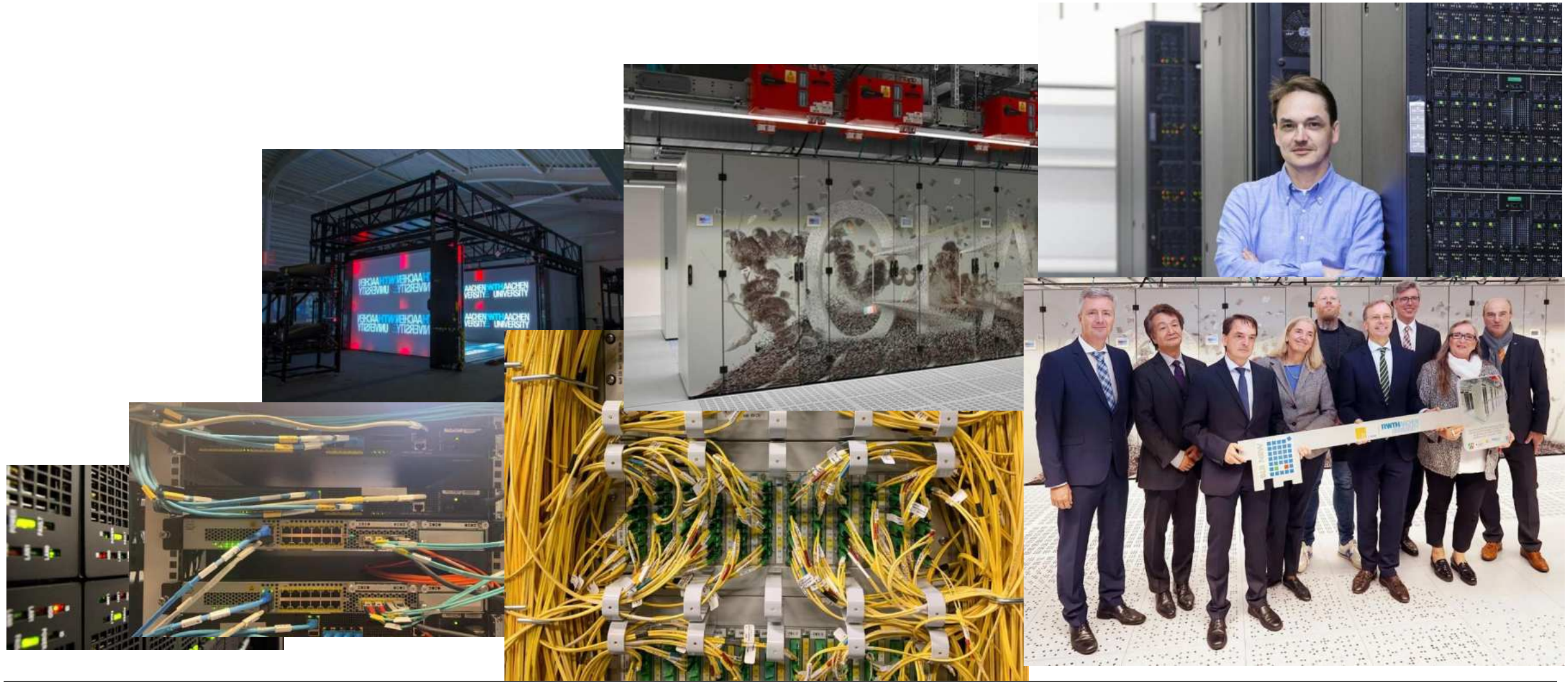
- Strong role to support Research Data Management at NRW
- Backup / Data protection
- Strategy development

National Mission

- **HPC for Computational Engineering Sciences (NHR4CES)**
- **Important node of the NFDI network**

IT Center | RWTH AACHEN UNIVERSITY

Wenn Ich an Rechenzentrum Denke, Denke Ich an…

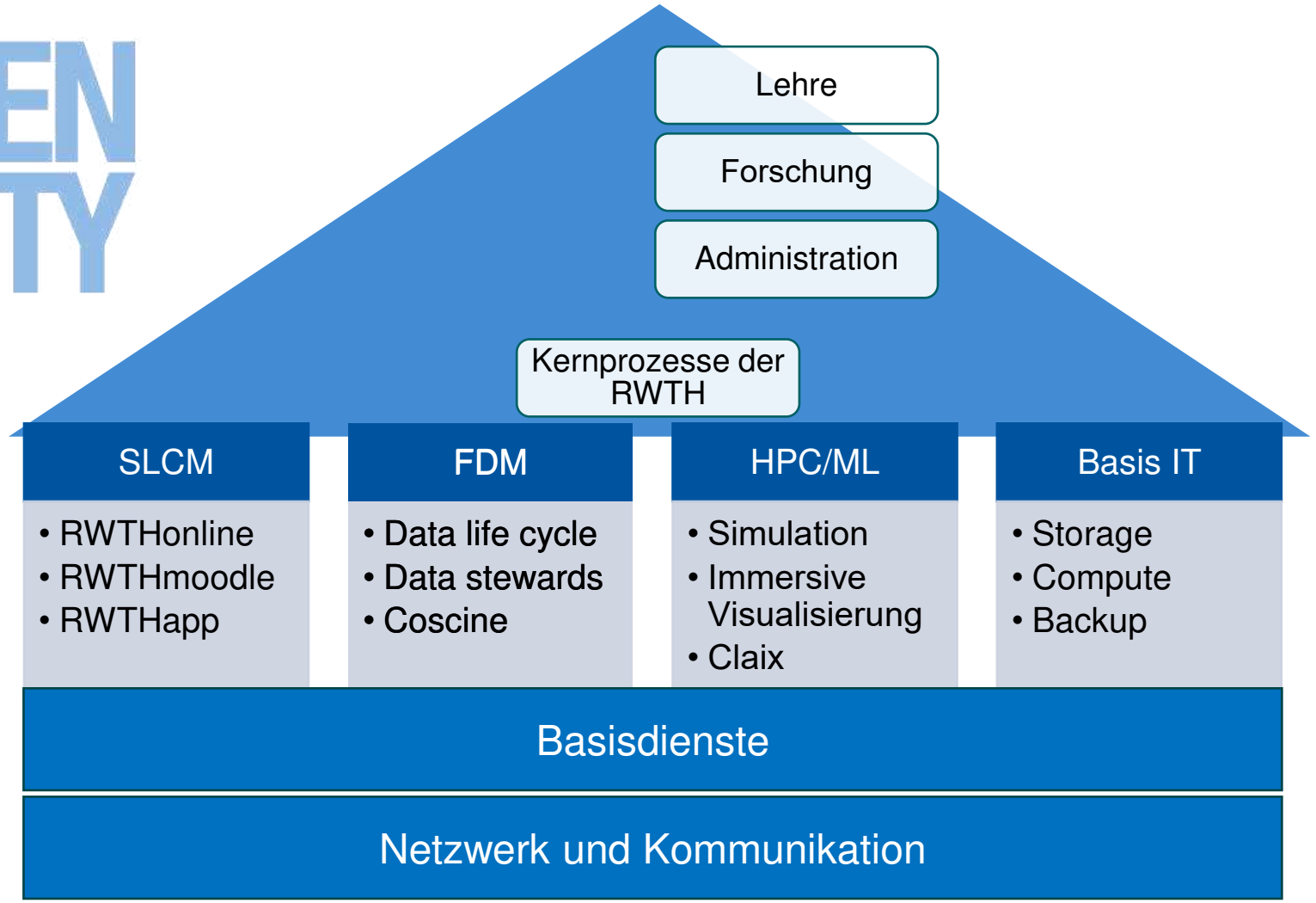


IT Center | RWTH AACHEN UNIVERSITY

Zukunftsorientierte Gestaltung der Kernprozesse der RWTH

RWTH AACHEN UNIVERSITY

IT Center



IT Center | RWTH AACHEN UNIVERSITY

Digitalisierung des Kernprozesses Forschung

- Unterstützung der guten wissenschaftlichen Praxis
  - FAIR-Prinzipien
  - Open Science
  - Open Source
- Beratung der Forschenden und die Betreuung von Forschungsprozessen
  - Data Stewardship
  - Forschungsdaten-Lebenszyklus
  - Bewirtschaftung wissenschaftlicher Ressourcen
- … auch für externe Forschungseinrichtungen
  - Auf nationaler und internationaler Ebene
  - Mitarbeit in Konsortien, z.B. NFDI oder NHR und wissenschaftlichen Projekten

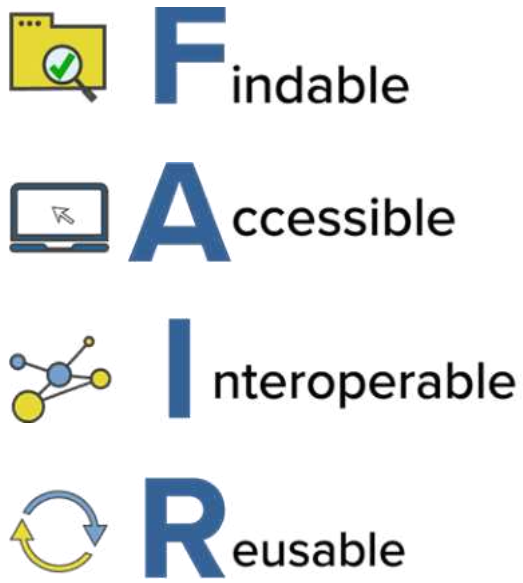


Image of FAIR acronymn with FAIR = Findable, Accessible, Interoperable, Reusable
National Library of Medicine
https://www.nlm.nih.gov/oet/ed/cde/tutorial/02-300.html

Digitalisierung des Kernprozesses Forschung

- Unterstützung der guten wissenschaftlichen Praxis
  - FAIR-Prinzipien
  - Open Science
  - Open Source
- Beratung der Forschenden und die Betreuung von Forschungsprozessen
  - Data Stewardship
  - Forschungsdaten-Lebenszyklus
  - Bewirtschaftung wissenschaftlicher Ressourcen
- … auch für externe Forschungseinrichtungen
  - Auf nationaler und internationaler Ebene
  - Mitarbeit in Konsortien, z.B. NFDI oder NHR und wissenschaftlichen Projekten

Findable

- Globally Unique & Persistent Identifier
- Registriert in einem Suchindex

Accessible

- Austausch über ein offenes Protokoll
- Zugriffsbeschränkungen wo nötig
- Vorhalten eines „Grabsteins“ auch nach Löschung

Interoperable

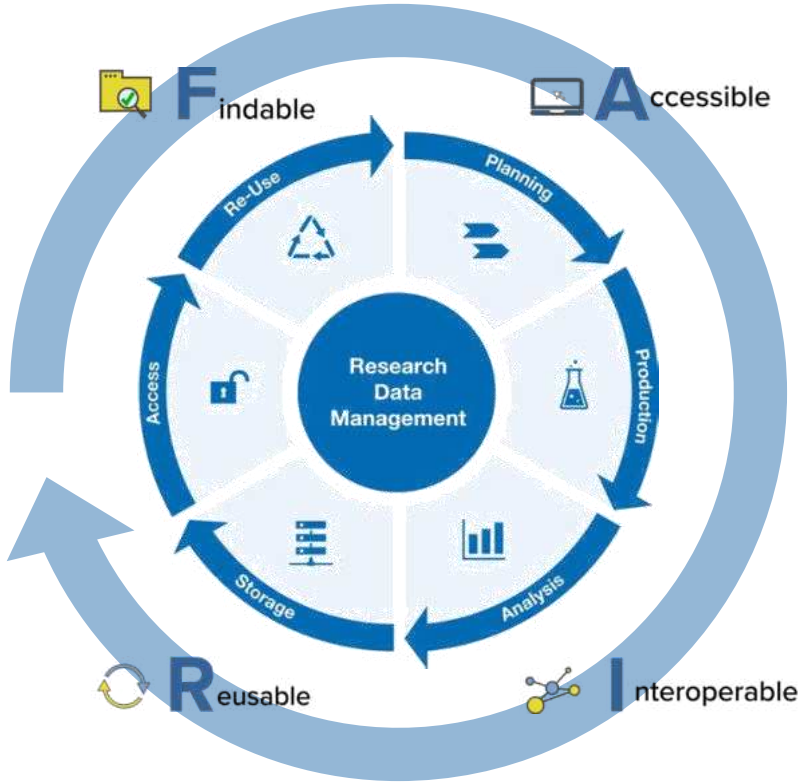
- Dokumentierte und offene Datenformate
- Definierte Datenstrukturen
- Verknüpfung mit anderen Datensätzen

Reusable

- Dokumentation entsprechend Anforderungen der Domäne
- Datenherkunft ist Dokumentiert
- Lizenz und Nutzungsbedingungen

Digitalisierung des Kernprozesses Forschung

- Unterstützung der guten wissenschaftlichen Praxis
  - FAIR-Prinzipien
  - Open Science
  - Open Source
- Beratung der Forschenden und die Betreuung von Forschungsprozessen
  - Data Stewardship
  - Forschungsdaten-Lebenszyklus
  - Bewirtschaftung wissenschaftlicher Ressourcen
- … auch für externe Forschungseinrichtungen
  - Auf nationaler und internationaler Ebene
  - Mitarbeit in Konsortien, z.B. NFDI oder NHR und wissenschaftlichen Projekten



Digitalisierung der Forschung im Wandel der Zeit

“pre FDM” “Nadelstichtaktik” “Integration” “Skalierung”

2012 2014 … 2017 2018 2019 2020 2021 2022 2023 2024

ProjektRepository
- Disziplinübergreifende Kollaboration
- Basiert auf SharePoint
- RDF Metadaten

MetadataManager
- Metadaten Generator
- Disziplinspezifische Schemata
- RDF-basiert

simpleArchive
- Webbasierte Archivierung
- IBM Spectrum Protect Backend
- PIDs für Archivknoten

TECHNISCHE UNIVERSITÄT DARMSTADT

RDS.NRW
- Objektspeicher für Forschungsdaten
- Georedundante Speicherung
- 10 Jahre Vorhaltezeit

DataStorage.nrw

Research Data Storage
- Objektspeicher für Forschungsdaten
- Standortredundante Speicherung
- 10 Jahre Vorhaltezeit

COSCINE.NRW

Coscine

eLabFTW

SFB985 Probenmanagement
- Basiert auf SharePoint
- PID für chemische Proben
- Zentrale Datensammlung in Kooperation

AIMS

NFDi4ing

NFDI MatWerk

IAM4.nfdi

IT Center | RWTH AACHEN UNIVERSITY

Coscine – Unsere Motivation

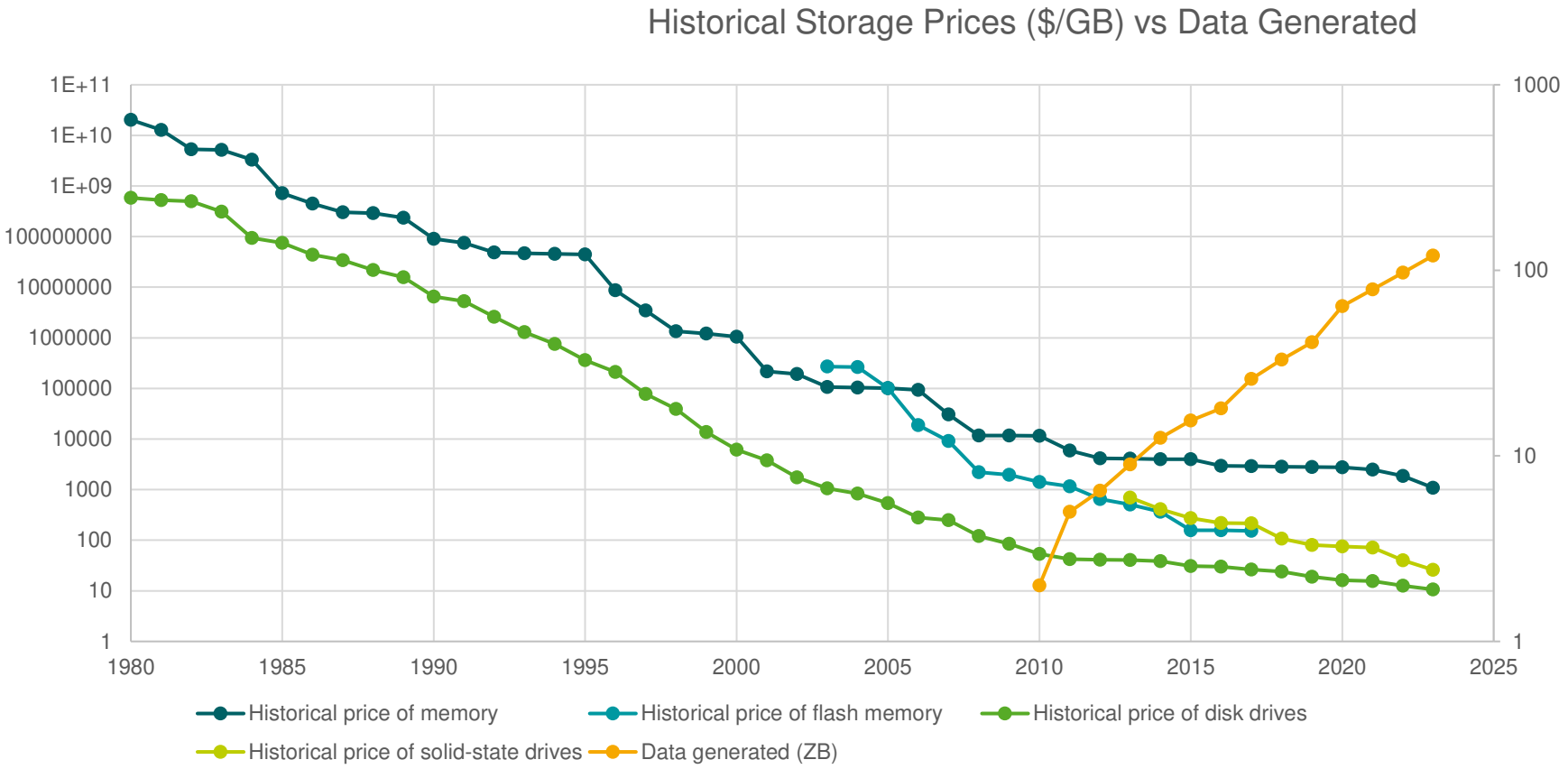
Problem

- Der FAIRe Umgang mit Forschungsdaten ist keine alltägliche Routine für Forschende
- Wenn überhaupt, wird nur ein kleiner Teil der Forschungsdaten nach den FAIR-Prinzipien behandelt - wenn sie veröffentlicht werden
- Eine beträchtliche Menge an wertvollen Informationen (Metadaten) geht verloren
- Keine Abbildung eines Daten-Life-Cycle (Planung, Zugang, Publikation, Vorhaltezeiten, …)

Unsere Lösung

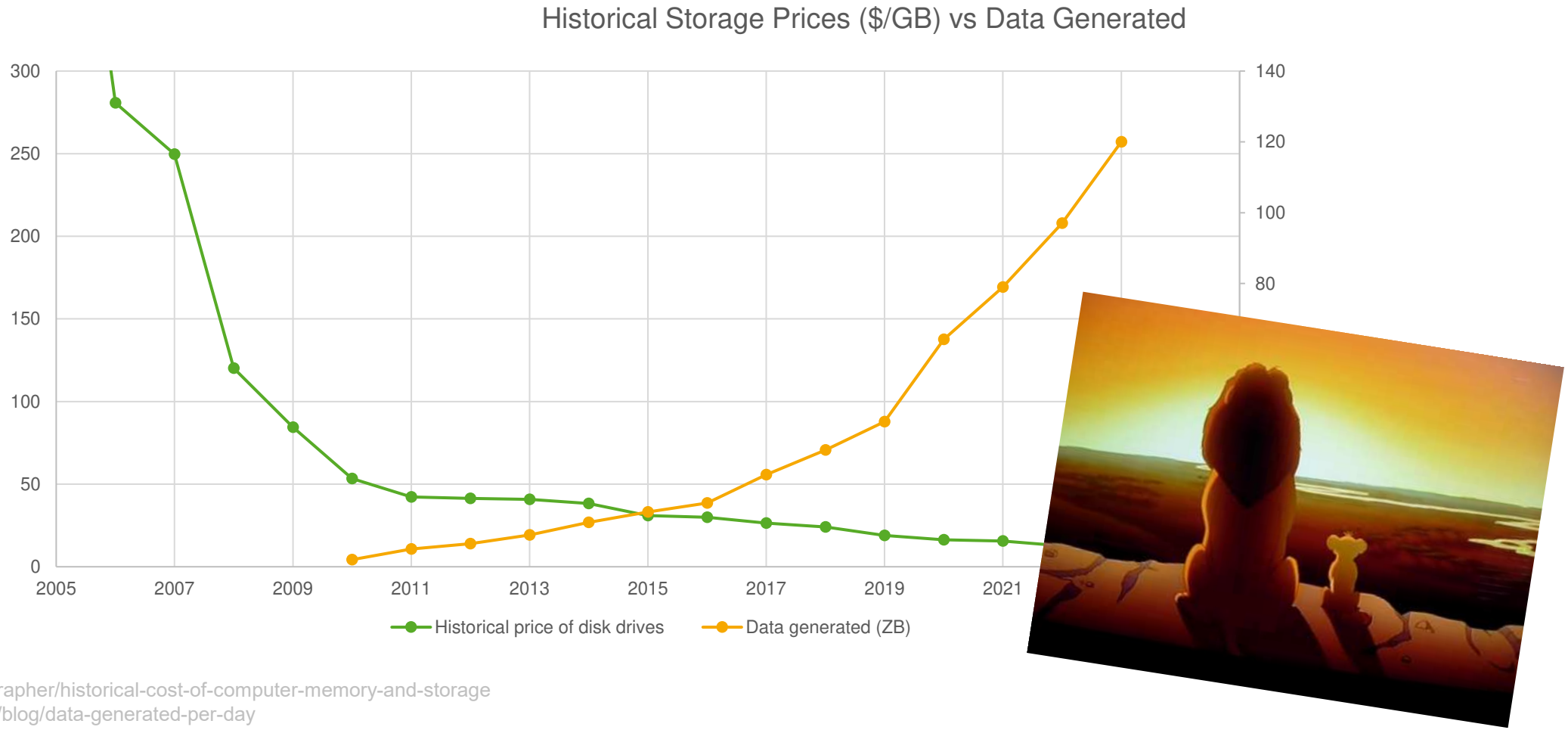
- Eine Umgebung zur Datenspeicherung und -verknüpfung, die implizit die FAIR-Prinzipien für beliebige Datenquellen implementiert

Warum machen wir uns die Mühe?



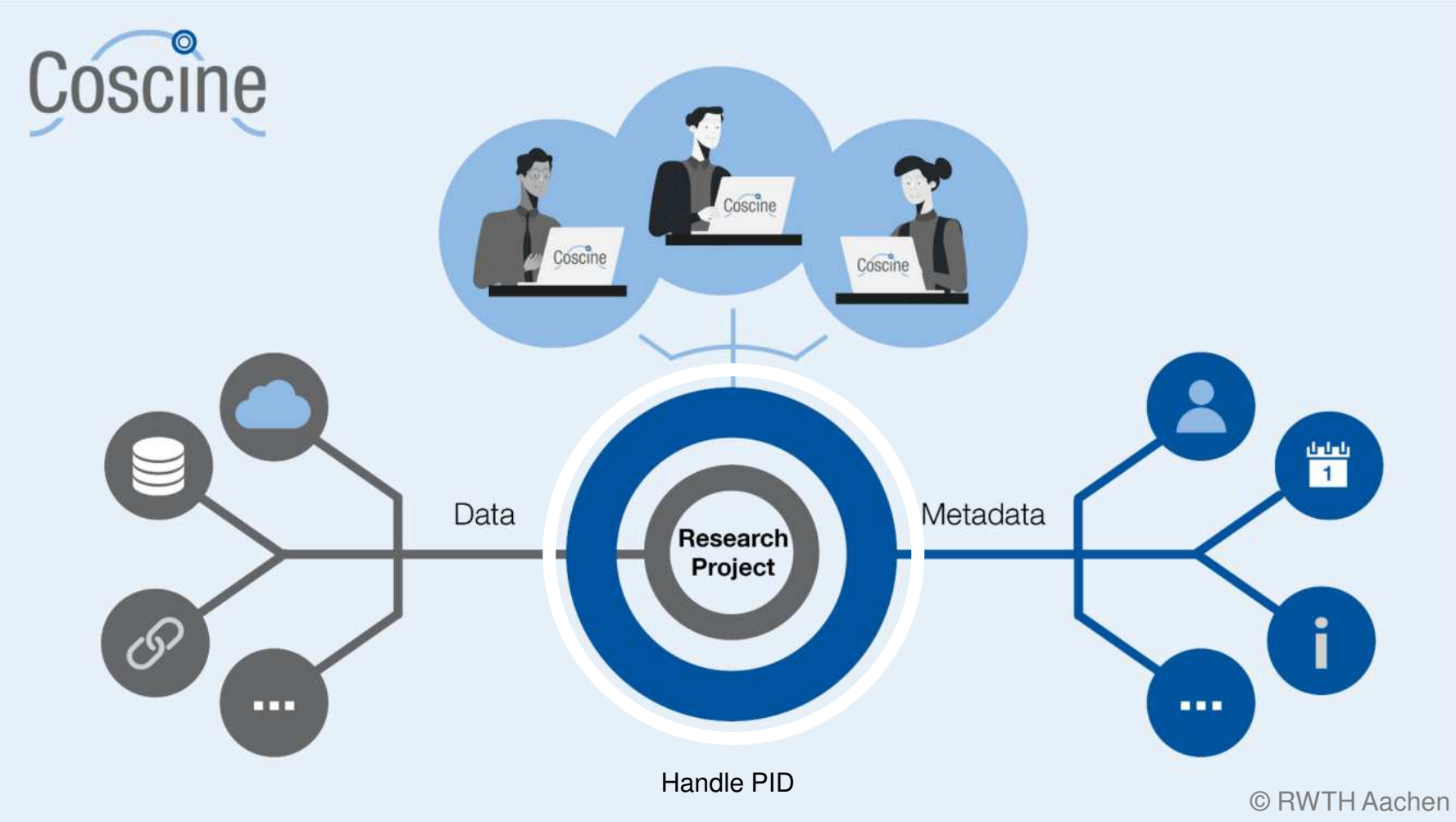
https://ourworldindata.org/grapher/historical-cost-of-computer-memory-and-storage
https://explodingtopics.com/blog/data-generated-per-day

Warum machen wir uns die Mühe?



https://ourworldindata.org/grapher/historical-cost-of-computer-memory-and-storage
https://explodingtopics.com/blog/data-generated-per-day

Coscine – Struktur



Coscine – Unsere Motivation

Hauptgründe für die Nutzung (Perspektive Nutzende)

1. Zugang zu Speicherplatz (z.B. RDS.NRW)
2. Erfüllung von Förderrichtlinien
3. Archivierung von Forschungsdaten für 10 Jahre

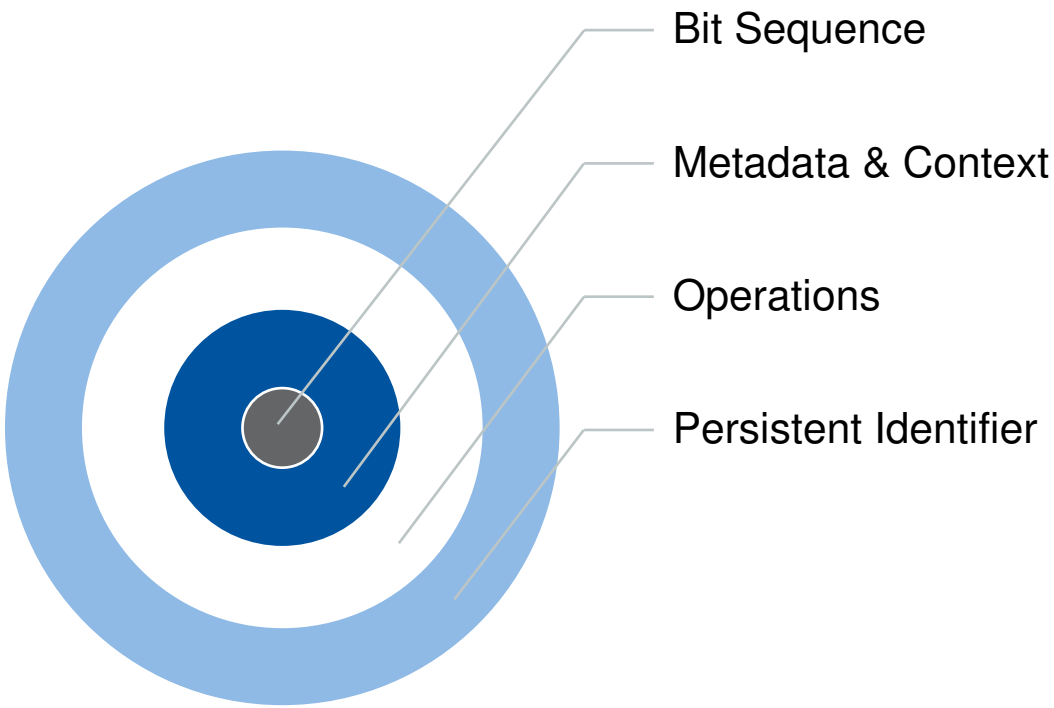
Hauptgründe für die Entwicklung (Hochschule / NFDI Perspektive)

- Erfüllung von Anforderungen für FDM (insb. Metadatenmanagement)
  - z.B. im NRW-Hochschulgesetz verankerte Anforderungen an die Gewährleistung guter wissenschaftlicher Praxis (NRW HG § 3 Abs. 1)
- Implementierung eines Daten-Life-Cycle

The FAIR Digital Object Concept

A FDO is **a unit of data**, represented as a sequence of bits that binds all critical information about an entity **in one place** and creates a new kind of **actionable**, meaningful and technology-independent object.

https://fairdo.org/

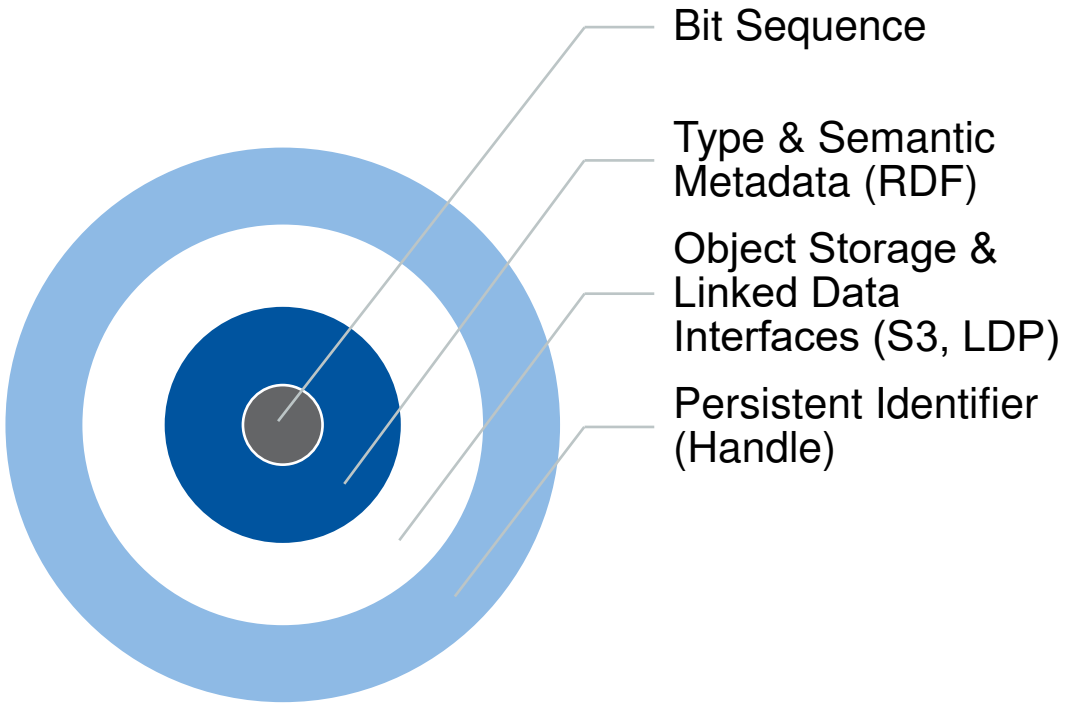


"The object itself can be digital data, code, metadata or the digital representation of a physical object, etc."

Tobias Weigel et. al., RDA Recommendation on PID Kernel Information FINAL

FAIR in the FAIR Digital Object

- The PID makes the record *findable*.
- Operations based on general protocols make data and metadata *accessible*.
- A broadly understandable metadata record and context in the record makes it *interoperable*.
- The record indicates *re-usability* of the data.



Coscine – (k)ein Repositorium?

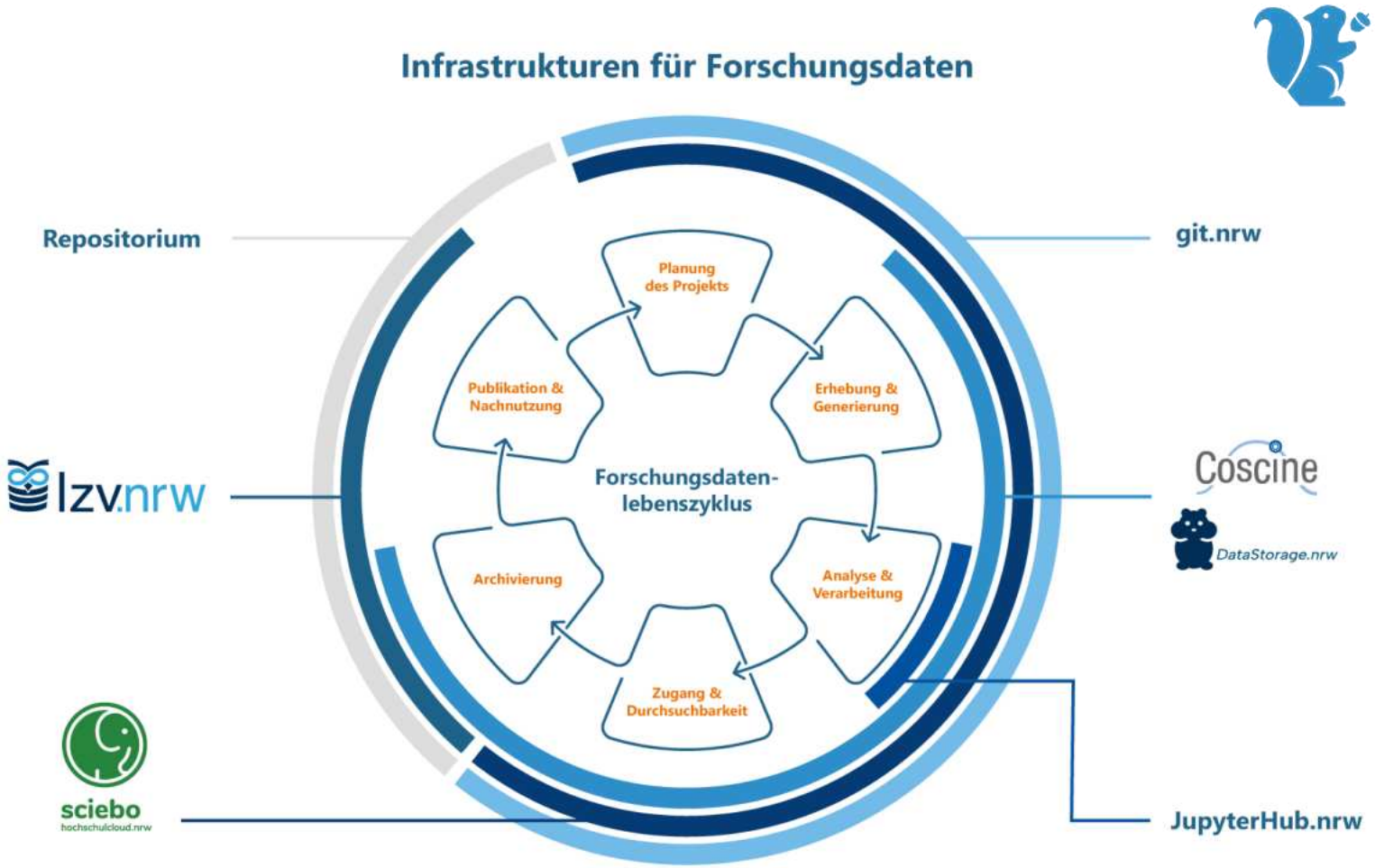
Coscine ist (k)ein Repositorium ?…

- für die **aktive Projektphase** gedacht um Forschungsdaten mit Metadaten zu speichern
- Metadaten sind obligatorisch aber es erfolgt **keine Kuration** durch die Betreiberin
- Datensätze erhalten PIDs aber **keine Veröffentlichung** via DOI, sondern Zugriffsmanagement über AAI
- Archivierung max. 10 Jahre nach Projektende aber **keine Langzeitarchivierung**

Definition

„Repositorien sind [verwaltete] Speicherorte für digitale Objekte, die diese für einen öffentlichen oder beschränkten Nutzerinnen- oder Nutzerkreis zur Verfügung stellen.“ (forschungsdaten.info)

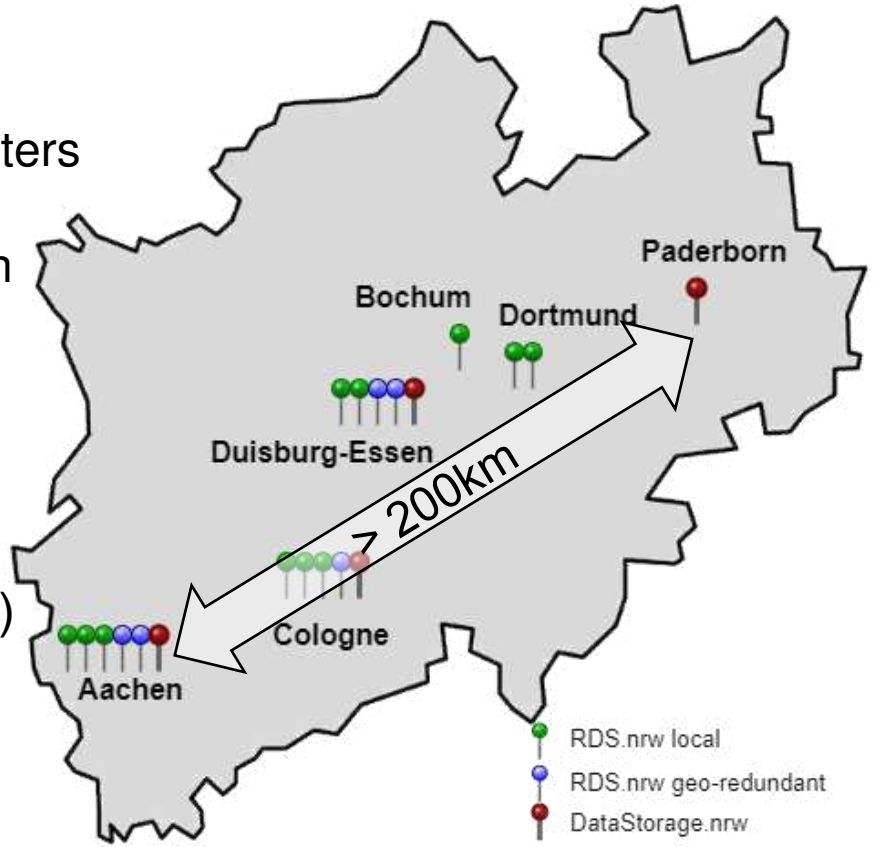
FDM-Landesdienste entlang des Forschungsdatenlebenszyklus



FDM-Landesdiensten entlang des Forschungsdatenlebenszyklus, fdm.nrw, CC BY 4.0, https://fdm-nrw.coscine.de/#/LaDi

Research Data Storage Infrastructure for NRW

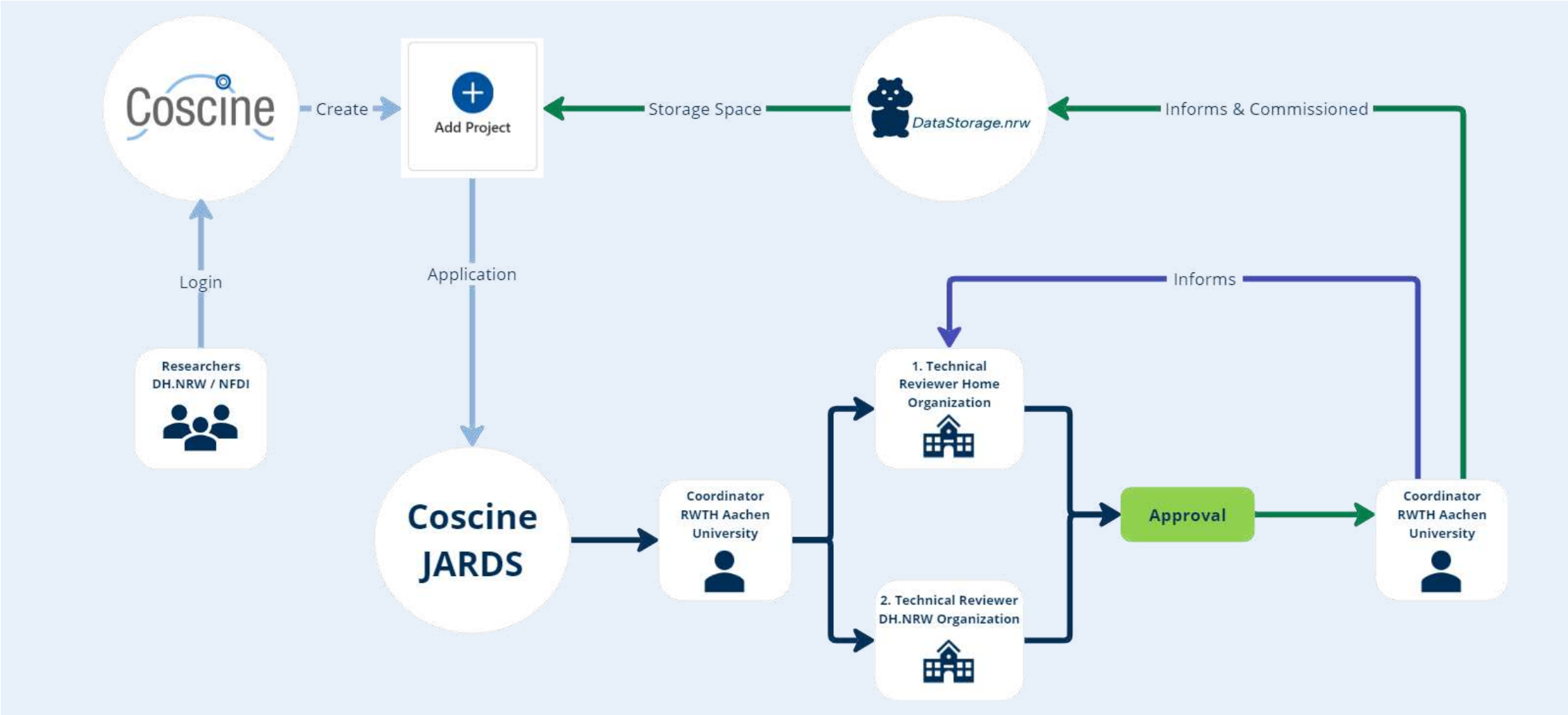
- Zentraler Service für Forschungsprimärdaten (“hot” / “warm”)
- HA geo-redundant gesichert Gegen Ausfall eines kompletten Datacenters
- Bewirtschaftung nach wissenschaftsgeleiteten Peer-Review-Verfahren
- Zusage eingelieferte Daten für mindestens 10 Jahre vorzuhalten
- 1. Generation (2020 – 2025): „RDS.nrw“ (~10PB usable capacity)
- 2. Generation (ab 2024): „DataStorage.nrw“ (>>10PB usable capacity)
- Skalierbarkeit auf >>100 PB
- Kontinuierliche (nutzungsbasierte) Erweiterung



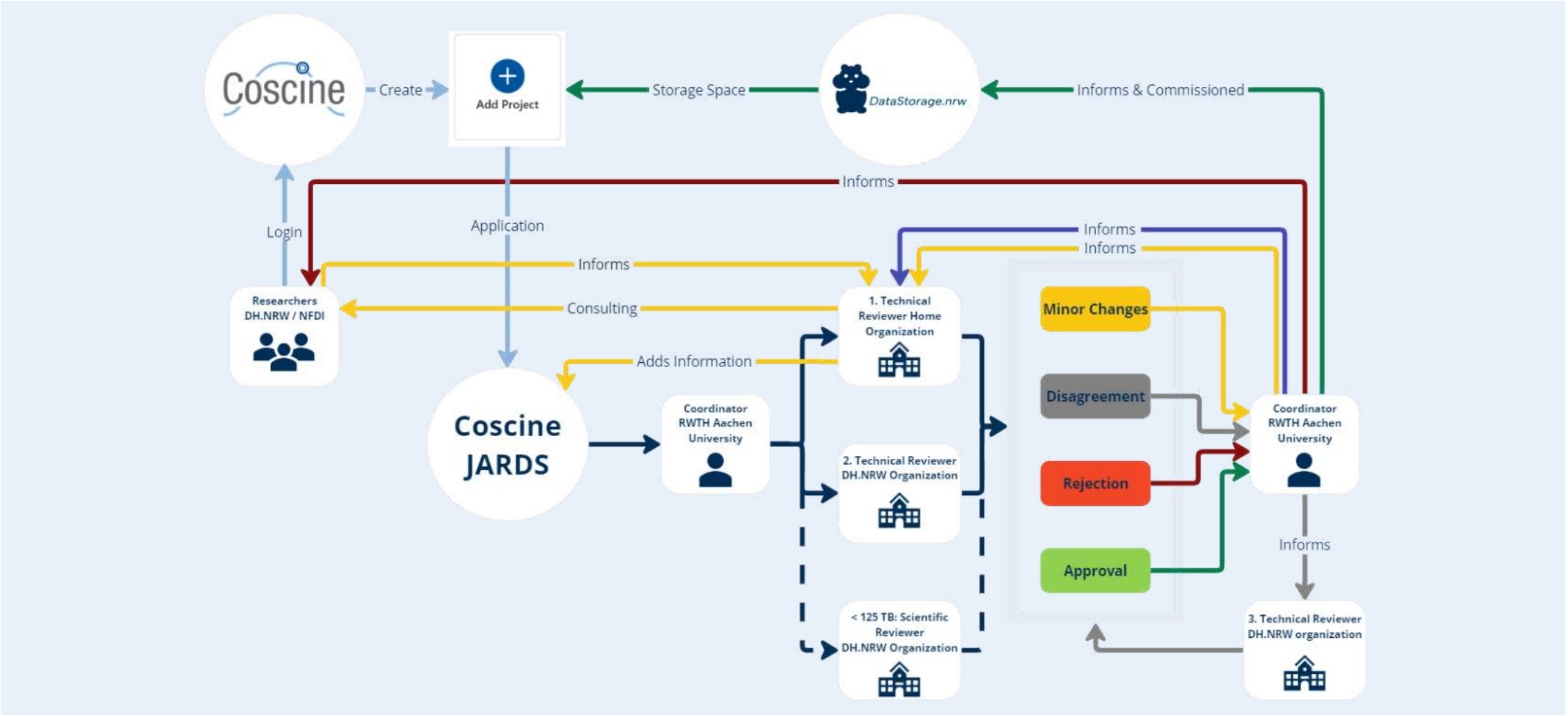
Coscine – Struktur

IT Cente

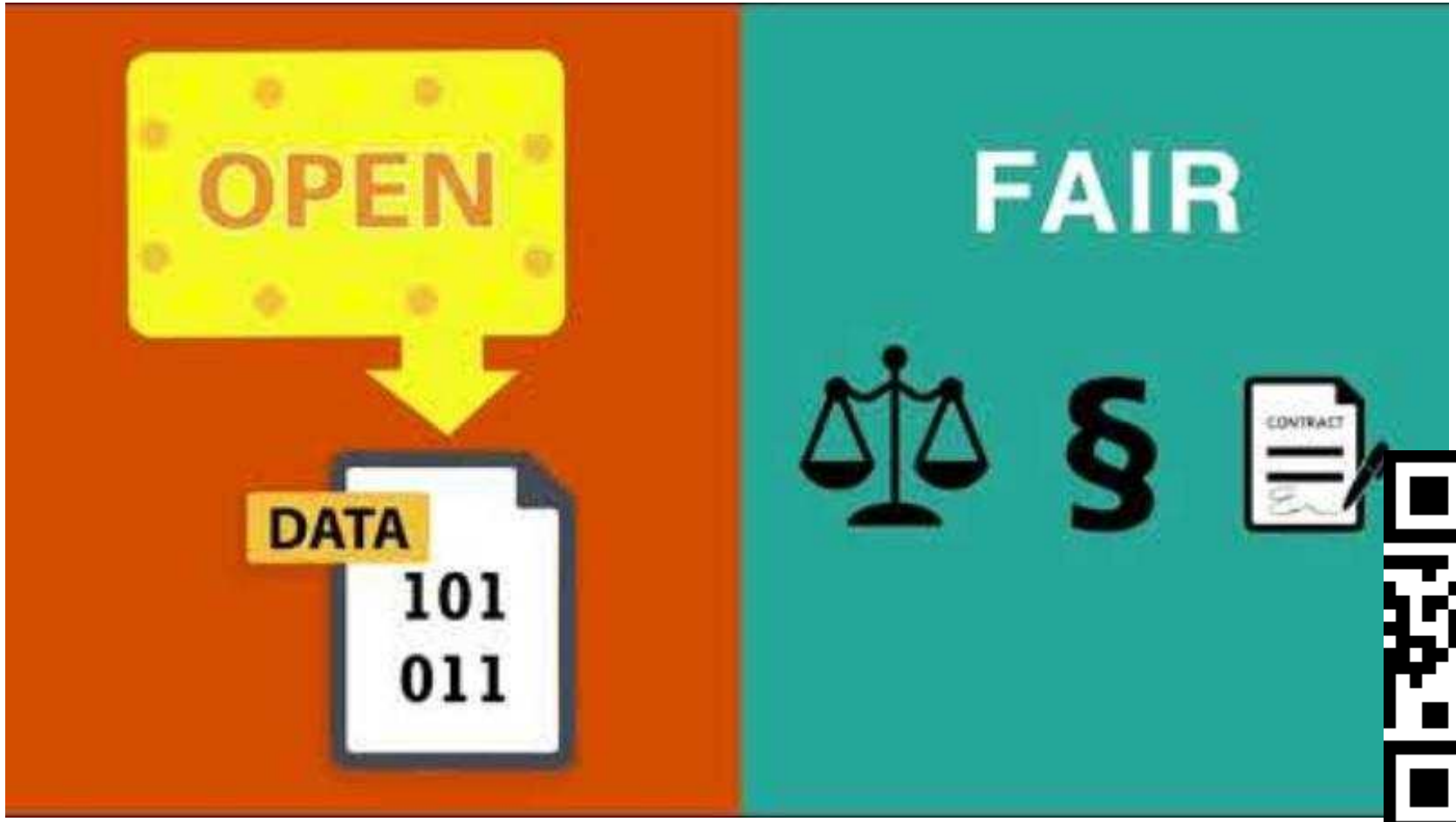
Provisionierung DataStorage.nrw: einfache Anträge



Provisionierung DataStorage.nrw: komplexe Anträge



FAIR Principles Explained…



https://www.youtube.com/watch?v=5OeCrQE3HhE

Vielen Dank
für Ihre Aufmerksamkeit

Dr. Marius Politze
0000-0003-3175-0659

politze@itc.rwth-aachen.de

RDS and RDS.NRW are funded by Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen (MKW:124-4.06.05.08-139057, DFG: INST222/1261-1). DataStorage.nrw is funded by Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen (MKW: 214-76.01.09-7-7937 DFG: INST 222/1530-1). Coscine.nrw is funded by Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen as a central Service.
The conceptual work was supported with resources granted by NFDI4Ing, funded by Deutsche Forschungsgemeinschaft (DFG) under project number 442146713, NFDI-MatWerk, funded by Deutsche Forschungsgemeinschaft (DFG) under project number 460247524 and FAIR Data Spaces, funded by the German Federal Ministry of Education and Research (BMBF) under funding reference FAIRDS11.