Digital Materials Environment – An Architecture and Tools based on FAIR Digital Objects for the NFDI-MatWerk

NFFA-Europe Workshop
26.09.2023

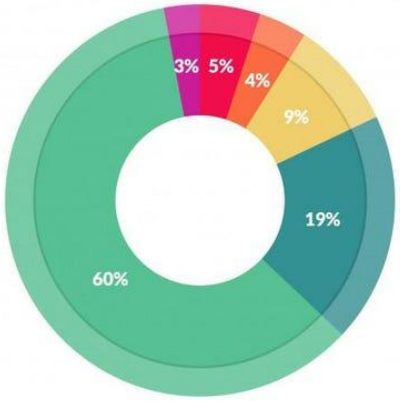
Marius Politze on behalf of NFDI-MatWerk TA-MDI

G. Abdildina, R. Aversa, I. Bierenbaum, E. Bitzek, M. Chmielowski, N. Garabedian, N. Golowin, K. Grünwald, P. Gumbsch, B. Heinrichs, S. Hunke, R. Joseph, A. Kirar, A. Moghaddam, M. Müller, P. Ost, M. Politze, Y. Shakeel, R. Stotzka, A. Streit, E. Vitali

NFDI-MatWerk is funded as part of the National Research Data Infrastructure (NFDI) following a recommendation of the German Joint Science Conference (GWK). The funding is provided by the Federal Government and the Heads of Government of the Länder and managed by the German Research Foundation (DFG) - project number 460247524.

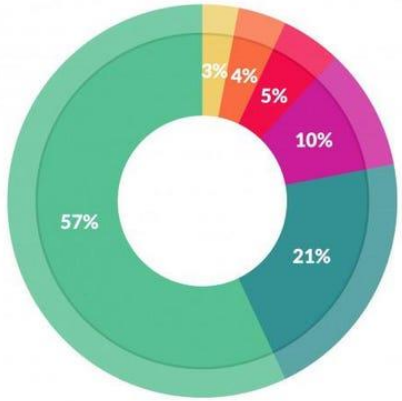
IT Center | RWTH AACHEN UNIVERSITY

Why Should I care?



What data scientists spend the most time doing

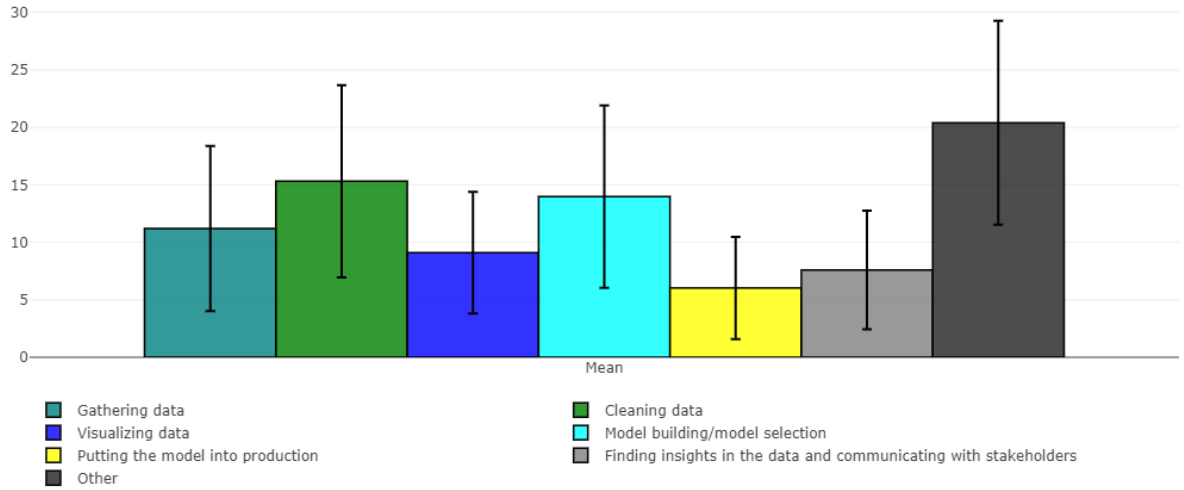
- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%



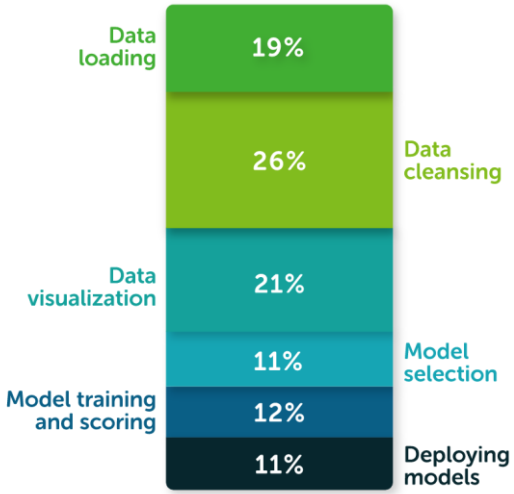
What's the least enjoyable part of data science?

- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%

Press, G (2016): Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says, https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#1b963786f637



Mooney, P (2018): Kaggle Machine Learning & Data Science Survey, https://www.kaggle.com/code/paultimothymooney/2018-kaggle-machine-learning-data-science-survey/notebook



Anaconda Inc. (2020): 2020 State of Data Science, https://www.anaconda.com/state-of-data-science-2020

Agenda

- NFDI
- NFDI-MatWerk
- TA-MDI
- Architecture for DME
- FAIRDO
- Selection of Tools

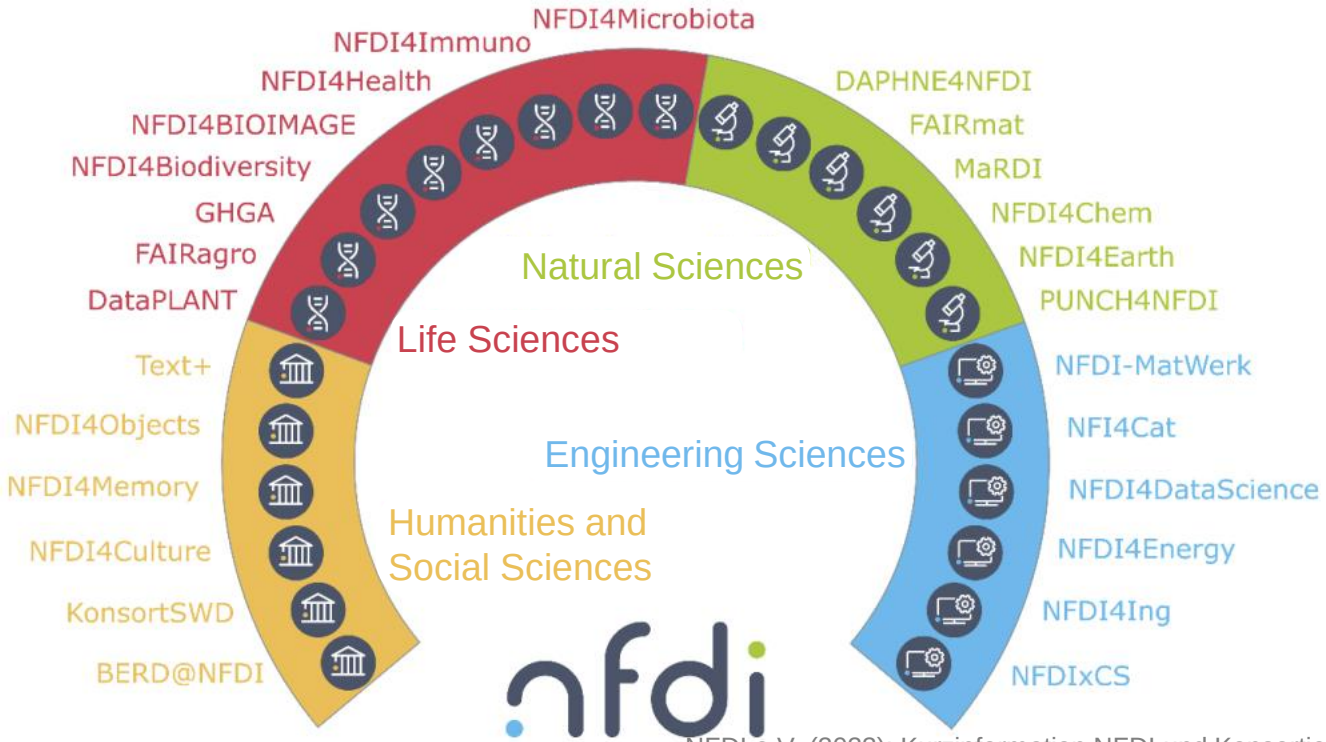
Agenda

- NFDI
- NFDI-MatWerk
- TA-MDI
- Architecture for DME
- FAIRDO
- Selection of Tools

https://www.etsy.com/de/listing/474171815/down-the-rabbit-hole-by-melody-smith

About NFDI

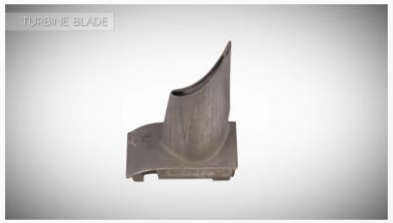
- NFDI – (German) National Research Data Infrastructure
- A network of 26 consortia from different scientific disciplines
- Base4NFDI to provide scalable „basic“ services
- NFDI sections for cross cutting topics
  - Common Infrastructures
  - Training & Education
  - Metadata
  - Industry Engagement
- Goal:
  Establish discipline specific infrastructure services „products“ for their respective scientific communities

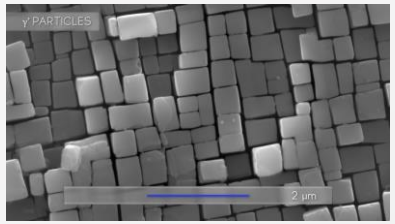


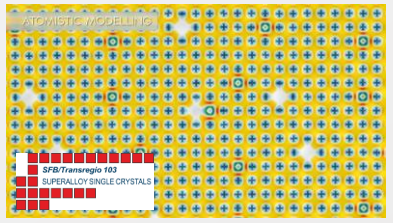
NFDI e.V. (2023): Kurzinformation NFDI und Konsortien
https://www.nfdi.de/wp-content/uploads/2023/03/Einseiter-Konsortien.pdf

NFDI MatWerk | IT Center | RWTH AACHEN UNIVERSITY

How are we connected within NFDI?







Application level

NFDi4ing

Mechanical and Civil Engineering, Production, Data Science

NFDI Mat Werk

Atomistic / molecular level

FAIRmat

NFDI4Phys

nfdi4chem

InnoMatSafety

Condensed matter physics and chemistry

Scientific tools, methodologies

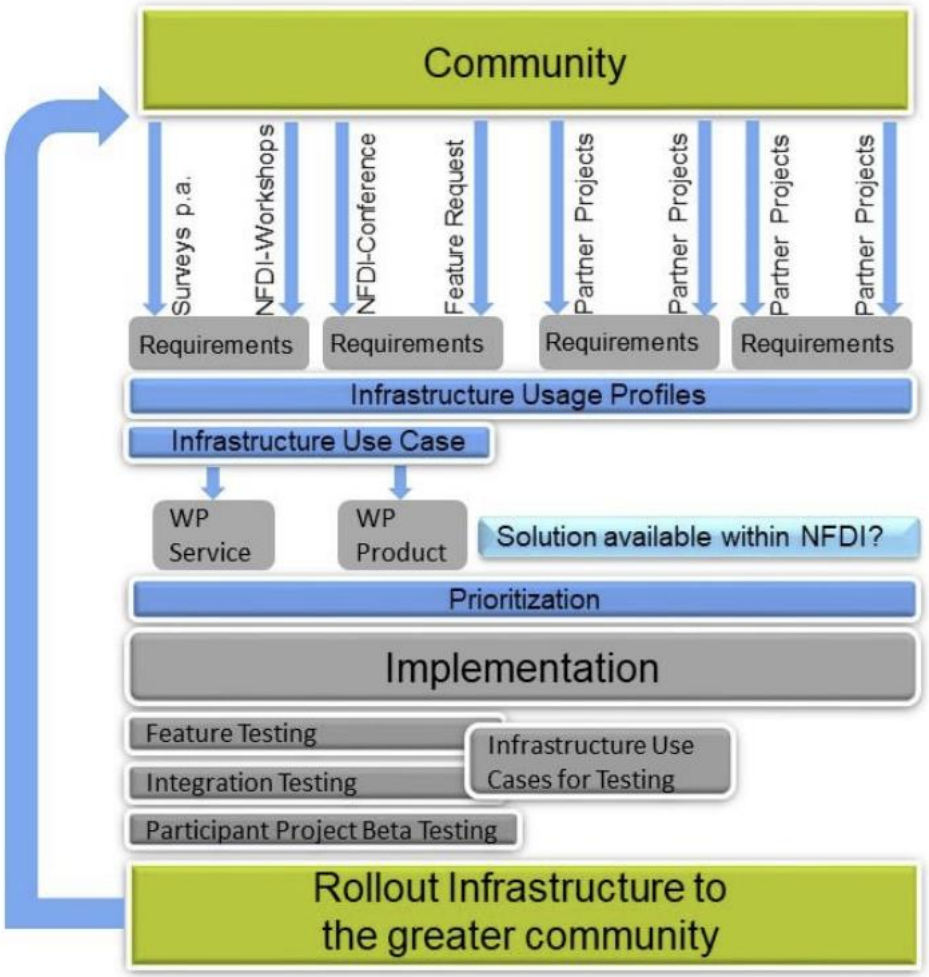
DataPLANT

MaRDI

Daphne

NFDI4HPC

NFDI-MatWerk – Community Driven Work Process



E.g. PP02 Defect phases in structural materials (CRC 1394)

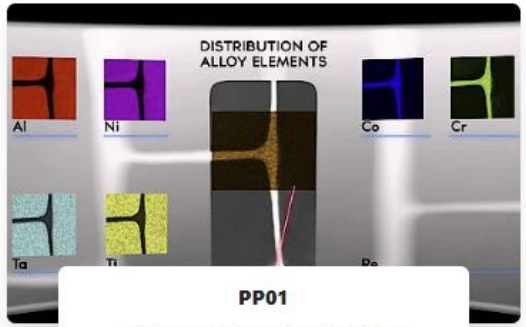
Requirements: seamless integration of mechanical simulations and heterogeneous databases for data analysis and prediction via machine learning tools.

Usage profile: Experts implement individual workflows to optimize of 3D microstructure data from various sources and correlate them with crystal plasticity. Experts seamlessly access the heterogeneous distributed data and metadata through standardized interfaces, visualize the analyzed data and develop automated processes/workflows which are used from others within the project.

Infrastructure Usecases:
- Semantic description of workflows
- (semantic) standardized interfaces
- Access to computing resources
- Access to storage: metadata, raw data, making data available (rights management), archiving

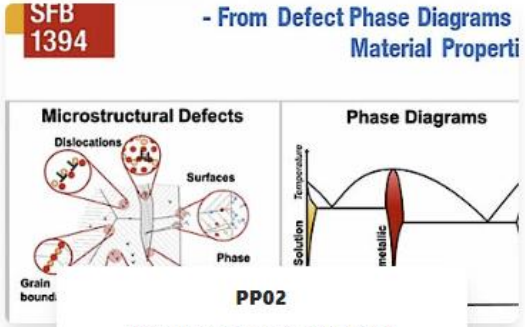
Eberl, Christoph, Niebel, Markus, Bitzek, Erik, Dahmen, Tim, Fritzen, Felix, Gumbsch, Peter, Hickel, Tilmann, Klein, Stefan, Mücklich, Frank, Müller, Matthias S., Neugebauer, Jörg, Pauly, Christoph, Politze, Marius, Sack, Harald, Sandfeld, Stefan, Schwaiger, Ruth, Slusallek, Philipp, Stotzka, Rainer, Streit, Achim, & Zimmermann, Martina. (2021). Consortium Proposal NFDI-MatWerk. Zenodo. https://doi.org/10.5281/zenodo.5082837

NFDI-MatWerk - Participant Projects (PP)



PP01

Defect phases in structural materials



PP02

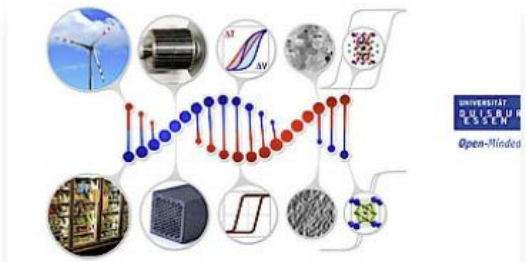
Defect phases in structural materials

PP03

Method development for mechanical joinability in versatile process chains

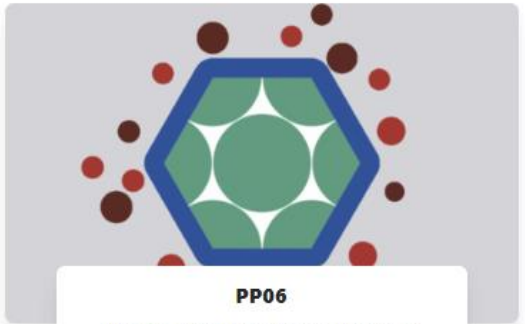
PP04

CRC 1261 - Magnetoelectric Sensors: From Composite Materials to Biomagnetic Diag-nostics



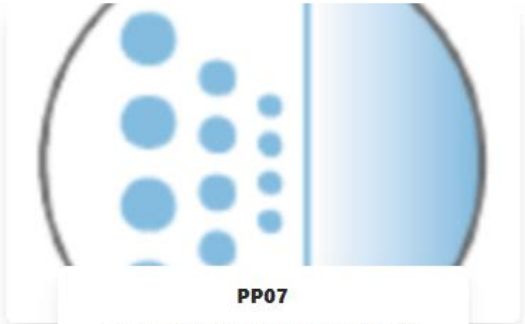
PP05

HoMMage – Hysteresis design of magnetic materials for efficient energy conversion



PP06

Working group modeling of oxygen-free production



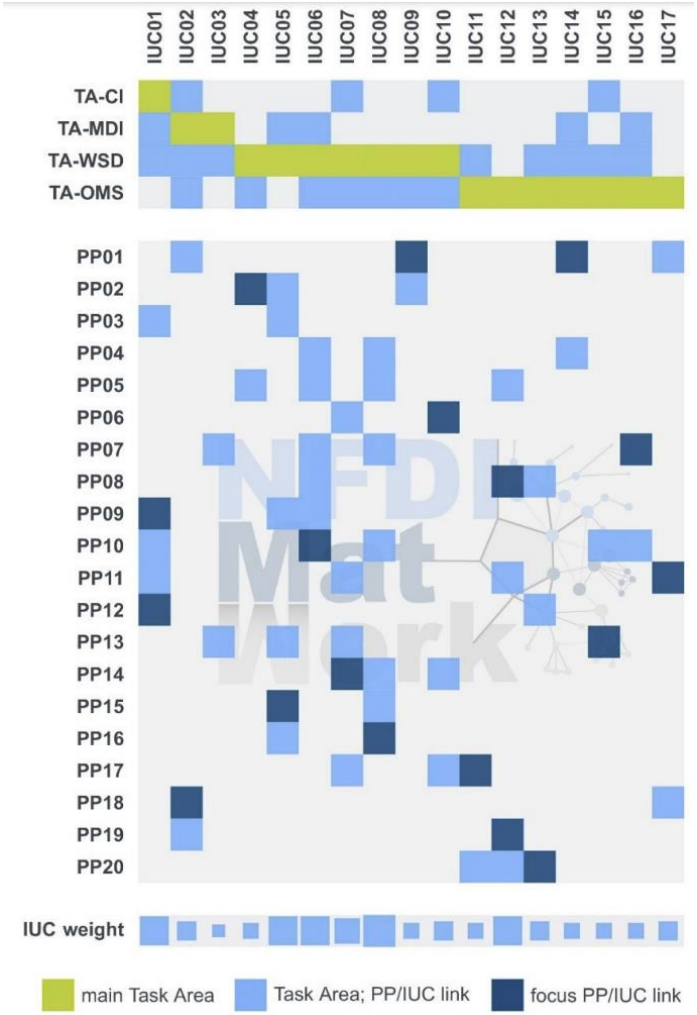
PP07

Processing uncertain microstructural data

PP08

Cluster of Excellence Living, Adaptive and Energy-autonomous Materials Systems

NFDI-MatWerk - Infrastructure Use Cases (IUC)



IUC03 Storage concepts for large hierarchical datasets
Joint Committee

Committee members: 0

Topics

Tomographic data from different techniques (e.g. Atom Probe Tomography, FIB/EBSD serial sectioning) and related simulations need to be stored in the DME in accordance with the underlying ontology for materials tomography and with reference coordinate systems that allow for the fusion of the datasets. A particular challenge here is the large amount of data that needs to be integrated and accessed, testing the performance and scalability of the DME.

Main requirements

- Digital Materials Environment (DME) Backend (MDI)
- Materials Data Infrastructure Support (MDI)
- Connection of workflows and ontologies (WSD)

Main Task Area: TA-MDI
Other related Task Areas: TA-WSD, TA-Strategy

Possible connections within NFDI: NFDI4Ing, FAIRmat

Material/Data: (mainly) metallic materials/tomographic data and related simulations with focus on large datasets

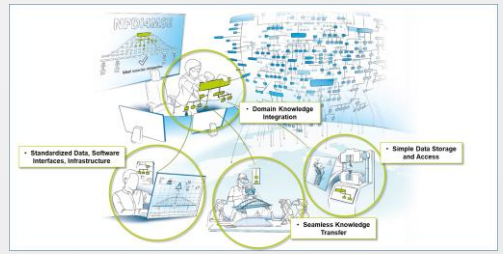
Main Success Scenario: Large amounts data from various, scale-bridging tomography methods of a given sample as well as simulations can be fused, specific volume elements can be accessed at different scales.

Added value for the MatWerk community: An exemplary implementation of a storage concept for large fused and linked datasets which is accessible through the DME and a web-based interface.

Eberl, Christoph, Niebel, Markus, Bitzek, Erik, Dahmen, Tim, Fritzen, Felix, Gumbsch, Peter, Hickel, Tilmann, Klein, Stefan, Mücklich, Frank, Müller, Matthias S., Neugebauer, Jörg, Pauly, Christoph, Politze, Marius, Sack, Harald, Sandfeld, Stefan, Schwaiger, Ruth, Slusallek, Philipp, Stotzka, Rainer, Streit, Achim, & Zimmermann, Martina. (2021). Consortium Proposal NFDI-MatWerk. Zenodo. https://doi.org/10.5281/zenodo.5082837

Consortium PI – Domain Expertise Meets Technical Enablers

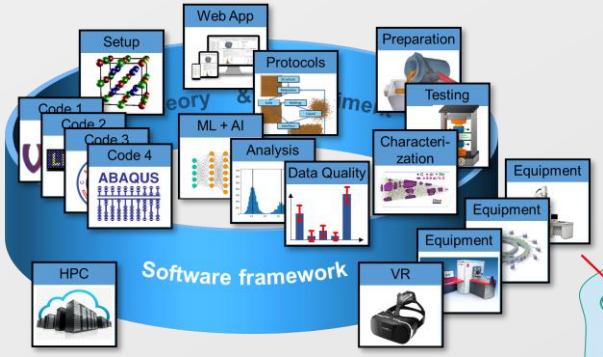
TA Strategy Development

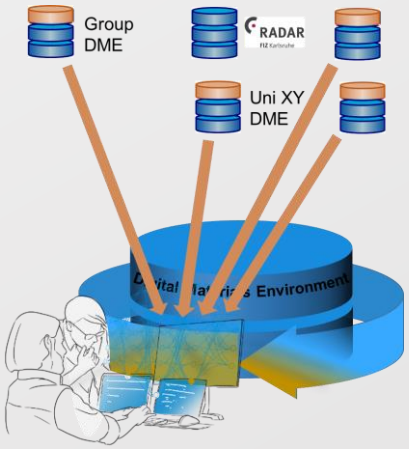


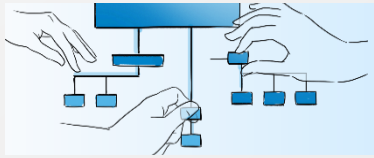
TA Community Interaction

TA Software Development

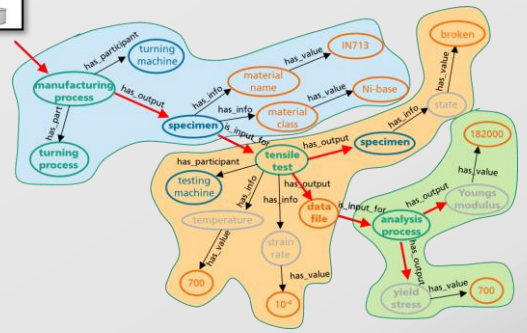
TA Materials Data Infrastructure





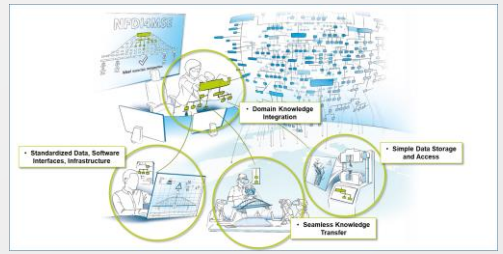


TA Ontologies for Materials Sciences



Consortium PI – Domain Expertise Meets Technical Enablers

TA Strategy Development

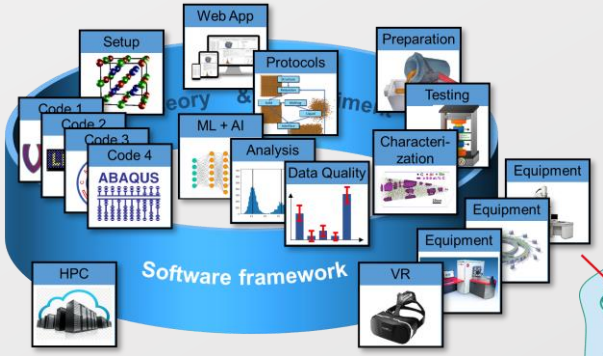


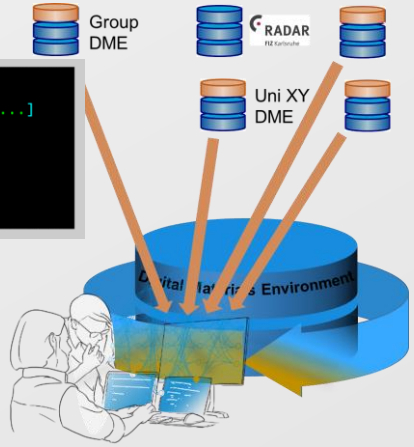
TA Community Interaction

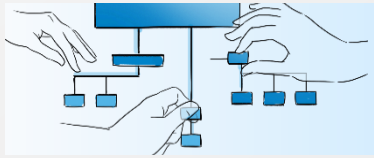
TA Software Development

```
NFDI4MSE_user> property = property_data.rate.min
property = creepRateMinimum == [0.00016, 0.00019, ...]
NFDI4MSE_user>
```

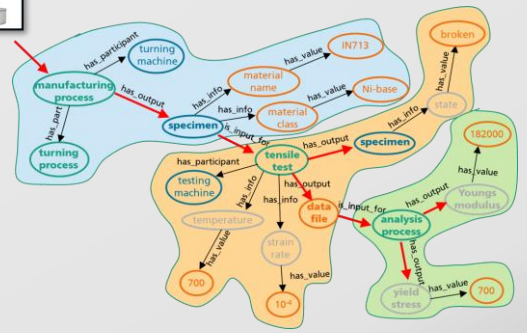
TA Materials Data Infrastructure







TA Ontologies for Materials Sciences



About Task Area Materials Data Infrastructure (TA-MDI)

- Collaboration between
  - Material Scientists at KIT, FAU and MPIE
  - IT-Service-Providers at KIT and RWTH

- Goals:
  - Design and implement scalable services to support IUCs
  - Ensure compatibility and interoperability with other NFDI consortia

- Measures:
  - Digital Materials Environment (DME) Backend
  - Web Frontends
  - Metadata Services
  - Data Governance Services
  - Support

Friedrich-Alexander-Universität Erlangen-Nürnberg

RWTH AACHEN UNIVERSITY

Karlsruher Institut für Technologie

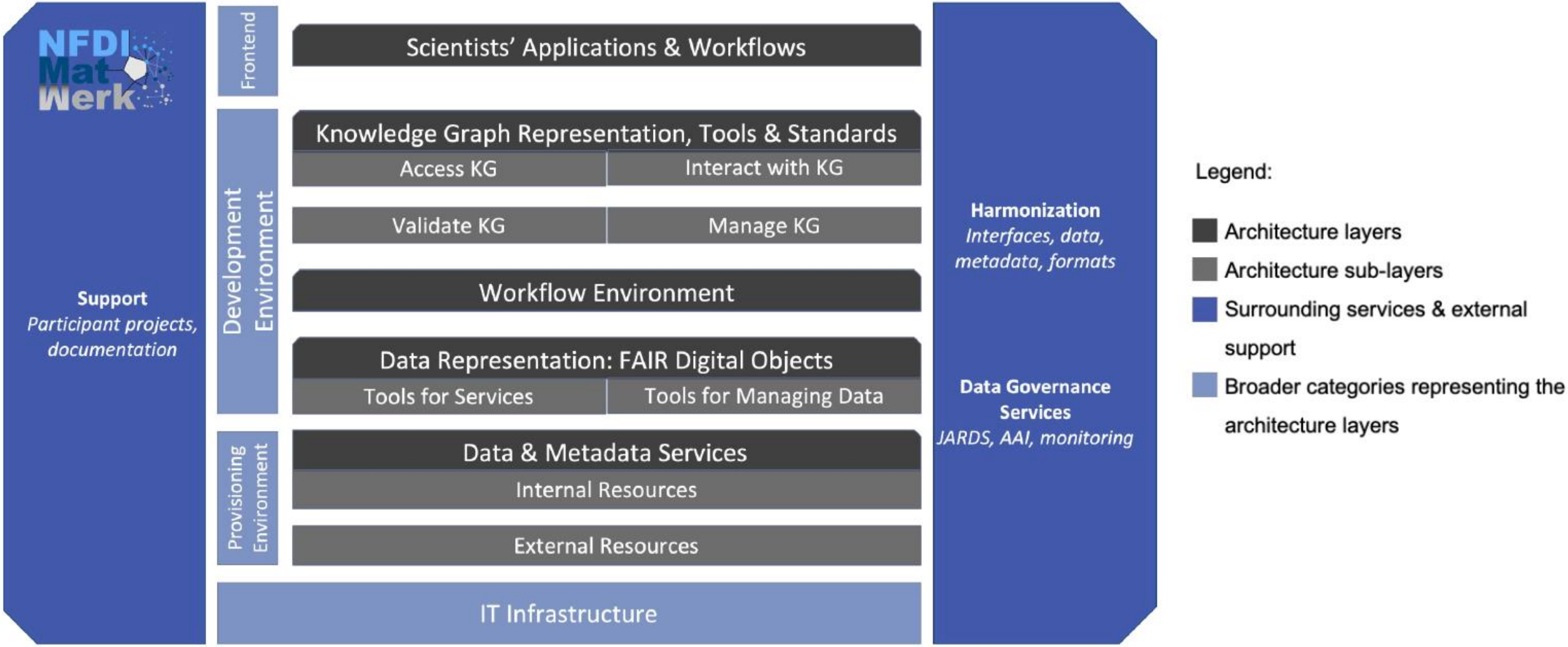
MAX-PLANCK-INSTITUT FÜR EISENFORSCHUNG GmbH

NFDI MatWerk

IT Center

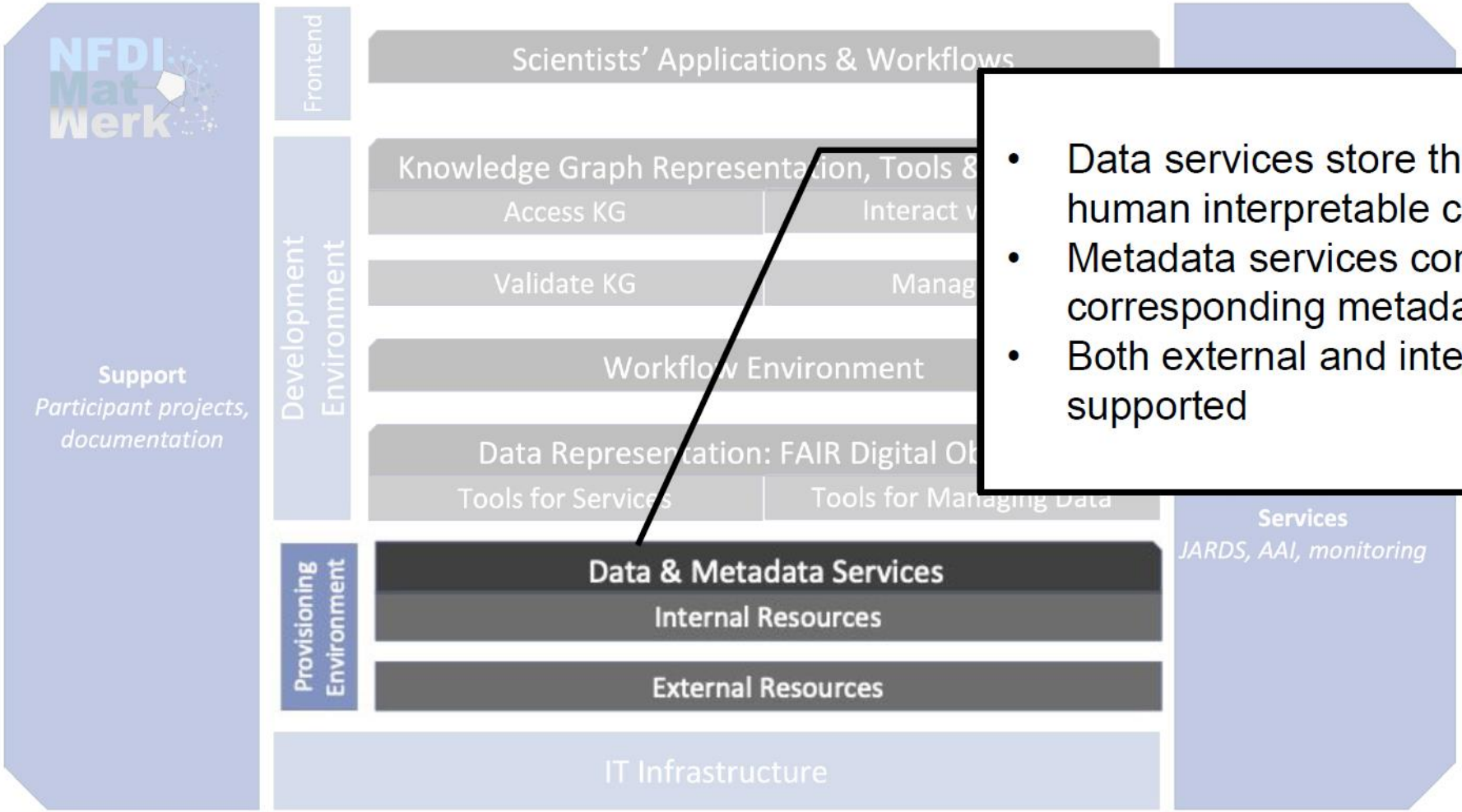
RWTH AACHEN UNIVERSITY

TA-MDI - Digital Materials Environment Architecture



M. Politze, Y. Shakeel, S. Hunke, P. Ost, R. Aversa, B. Heinrichs, l. Lang (2023): Long Term Interoperability of Distributed Research Data Infrastructures, https://doi.org/10.52825/CoRDI.v1i.348

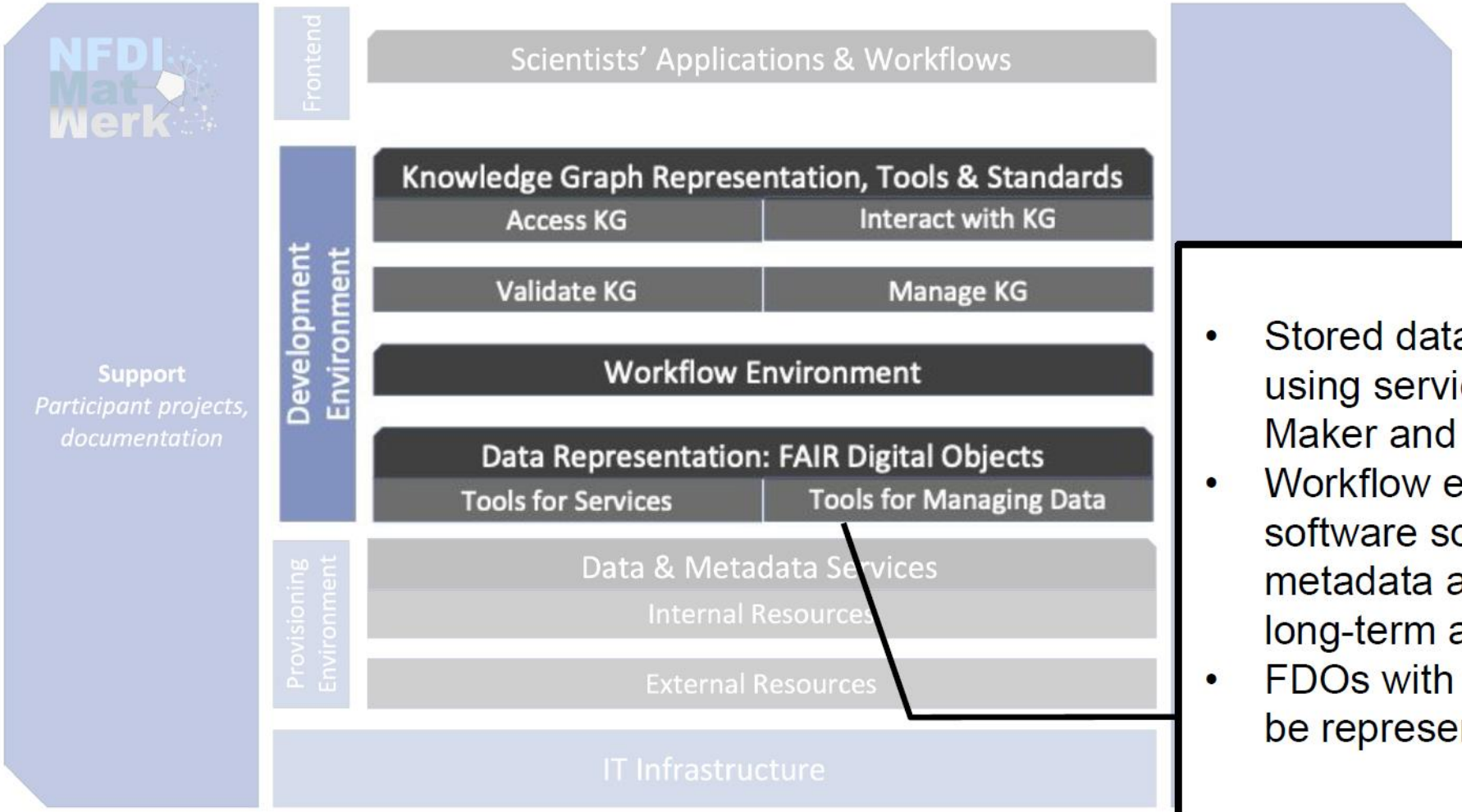
NFDI-MatWerk Shared Service Architecture



- Data services store the bit sequences as human interpretable content
- Metadata services complement by storing the corresponding metadata of the data stored
- Both external and internal resources are supported

M. Politze, Y. Shakeel, S. Hunke, P. Ost, R. Aversa, B. Heinrichs, l. Lang (2023): Long Term Interoperability of Distributed Research Data Infrastructures, https://doi.org/10.52825/CoRDI.v1i.348

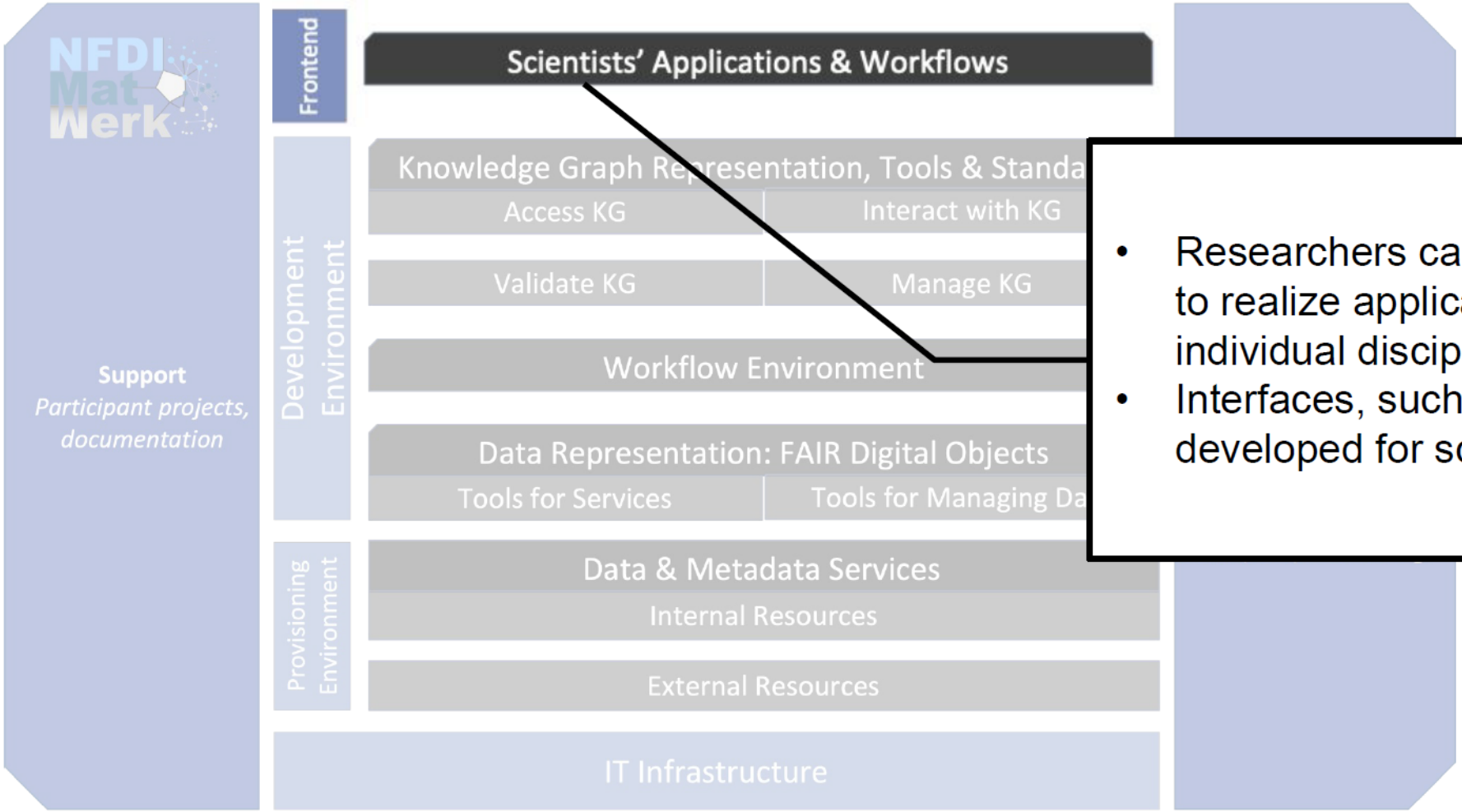
NFDI-MatWerk Shared Service Architecture



- Stored data is represented as FDO, using services such as, Typed PID Maker and FDO-scope
- Workflow environment supports software solutions to share data, metadata and workflows, ensuring their long-term availability
- FDOs with their interconnections can be represented and explored as KG

M. Politze, Y. Shakeel, S. Hunke, P. Ost, R. Aversa, B. Heinrichs, l. Lang (2023): Long Term Interoperability of Distributed Research Data Infrastructures, https://doi.org/10.52825/CoRDI.v1i.348

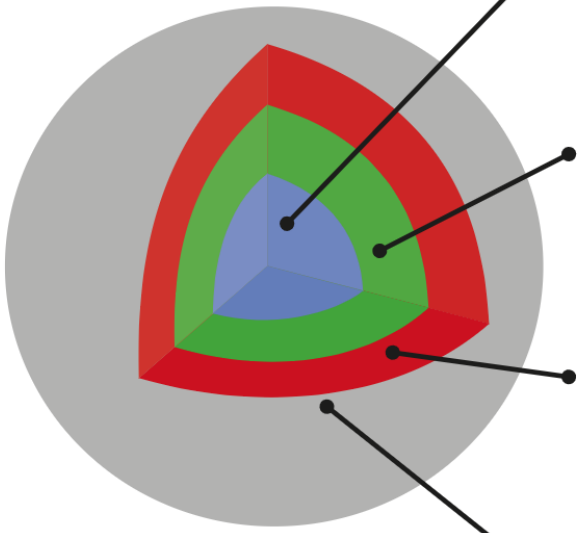
NFDI-MatWerk Shared Service Architecture



- Researchers can build on the workflows to realize applications relevant for their individual discipline
- Interfaces, such as ELNs can be developed for scientists to interact with

M. Politze, Y. Shakeel, S. Hunke, P. Ost, R. Aversa, B. Heinrichs, l. Lang (2023): Long Term Interoperability of Distributed Research Data Infrastructures, https://doi.org/10.52825/CoRDI.v1i.348

Implementing FAIR Principles: FAIR Digital Objects



DIGITAL OBJECT

Data, code and other research outputs

At its most basic level, data or code is a bitstream or binary sequence. For this to have meaning and to be FAIR, it needs to be represented in standard formats and be accompanied by Persistent Identifiers (PIDs), metadata and documentation. These layers of meaning enrich the object and enable reuse.

IDENTIFIERS

Persistent and unique (PIDs)

Digital Objects should be assigned a unique and persistent identifier such as a DOI or URN. This enables stable links to the object and supports citation and reuse to be tracked. Identifiers should also be applied to other related concepts such as the data authors (ORCIDs), projects (RAIDs), funders and associated research resources (RRIDs).

STANDARDS & CODE

Open, documented formats

Digital Objects should be represented in common and ideally open file formats. This enables others to reuse them as the format is in widespread use and software is available to read the files. Open and well-documented formats are easier to preserve. Data also need to be accompanied by the code use to process and analyse the data.

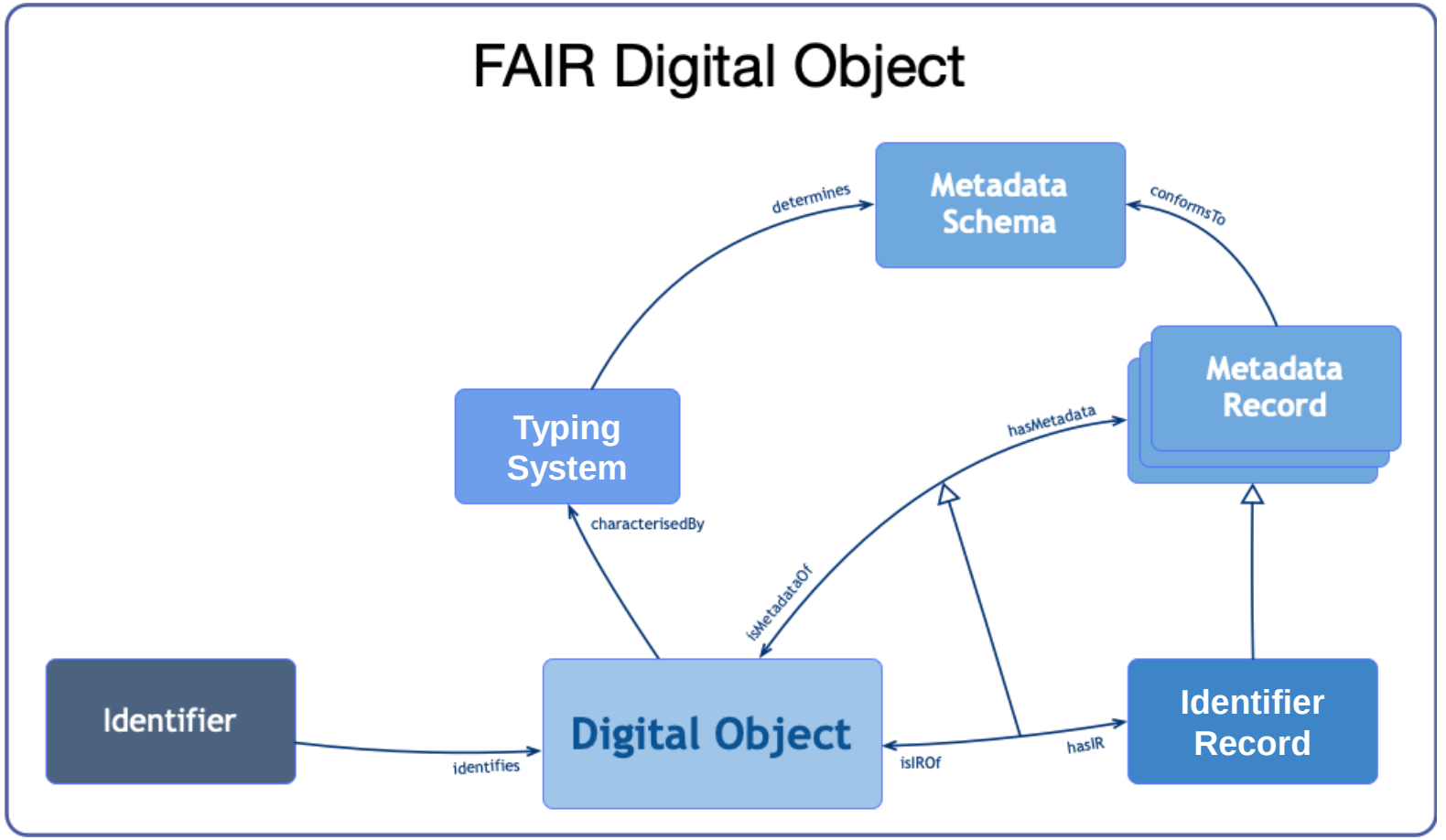
METADATA

Contextual documentation

In order for Digital Objects to be assessable and reusable, they should be accompanied by sufficient metadata and documentation. Basic metadata will enable data discovery, but much richer information and provenance is required to understand how, why, when and by whom the objects were created. To enable the broadest reuse, they should be accompanied by a plurality of relevant attributes and a clear and accessible usage license.

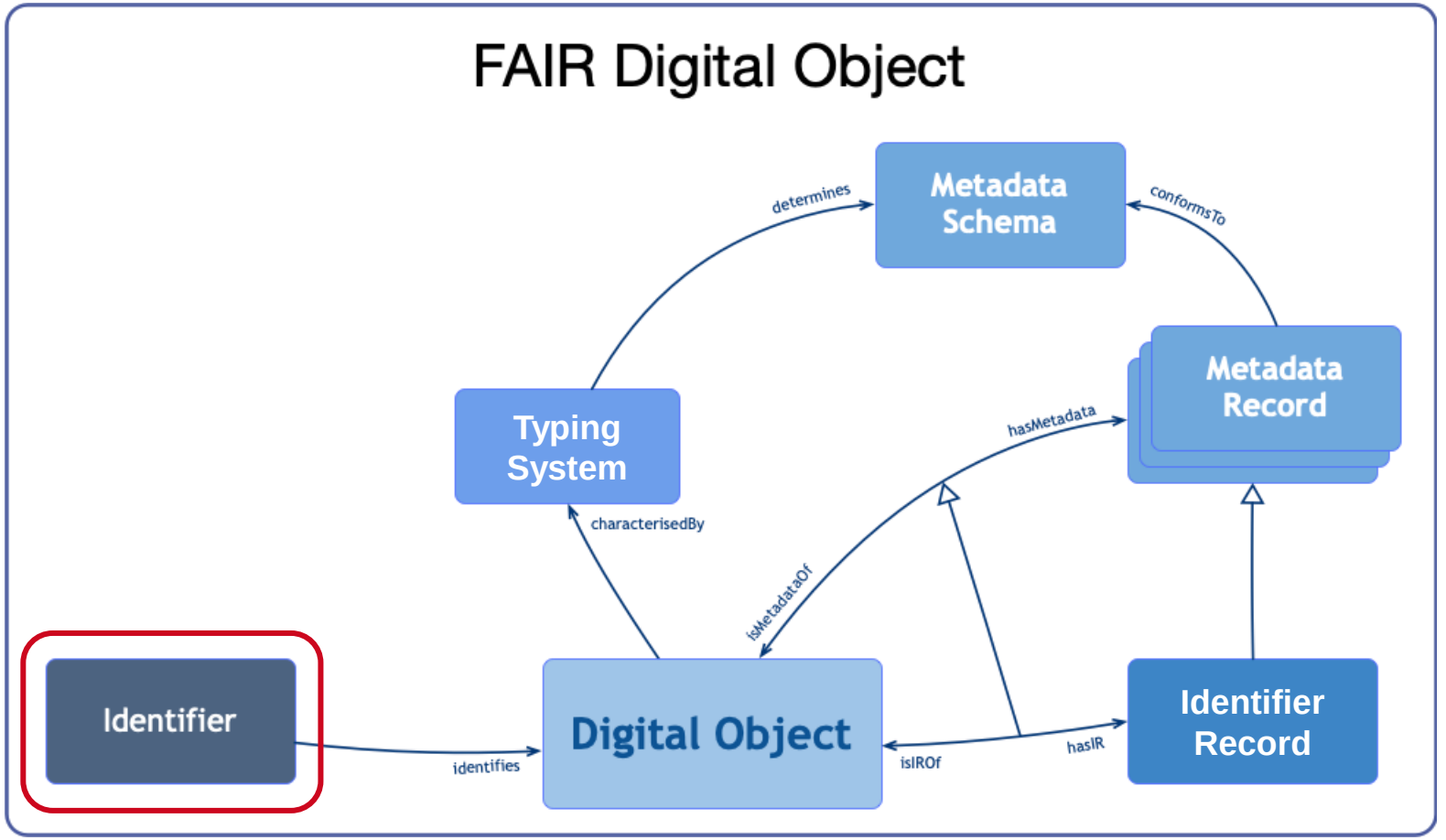
European Commission, Directorate-General for Research and Innovation, Turning FAIR into reality : final report and action plan from the European Commission expert group on FAIR data, Publications Office, 2018, doi:10.2777/1524

FAIR Digital Object – A Structured Set of Links



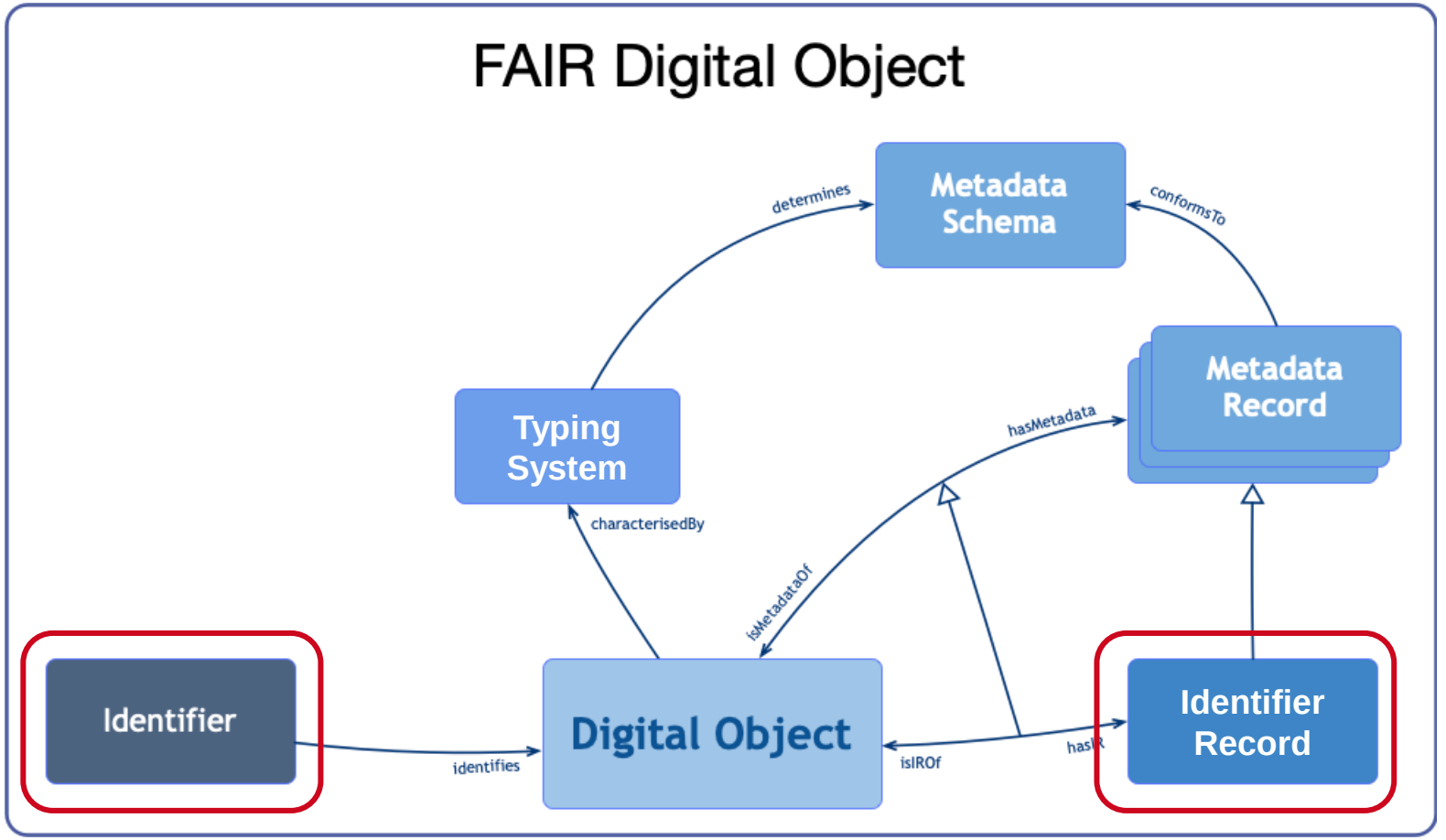
Bonino da Silva Santos, L. O. (2021): FAIR Digital Object Framework Documentation. https://fairdigitalobjectframework.org/

FAIR Digital Object – A Structured Set of Links



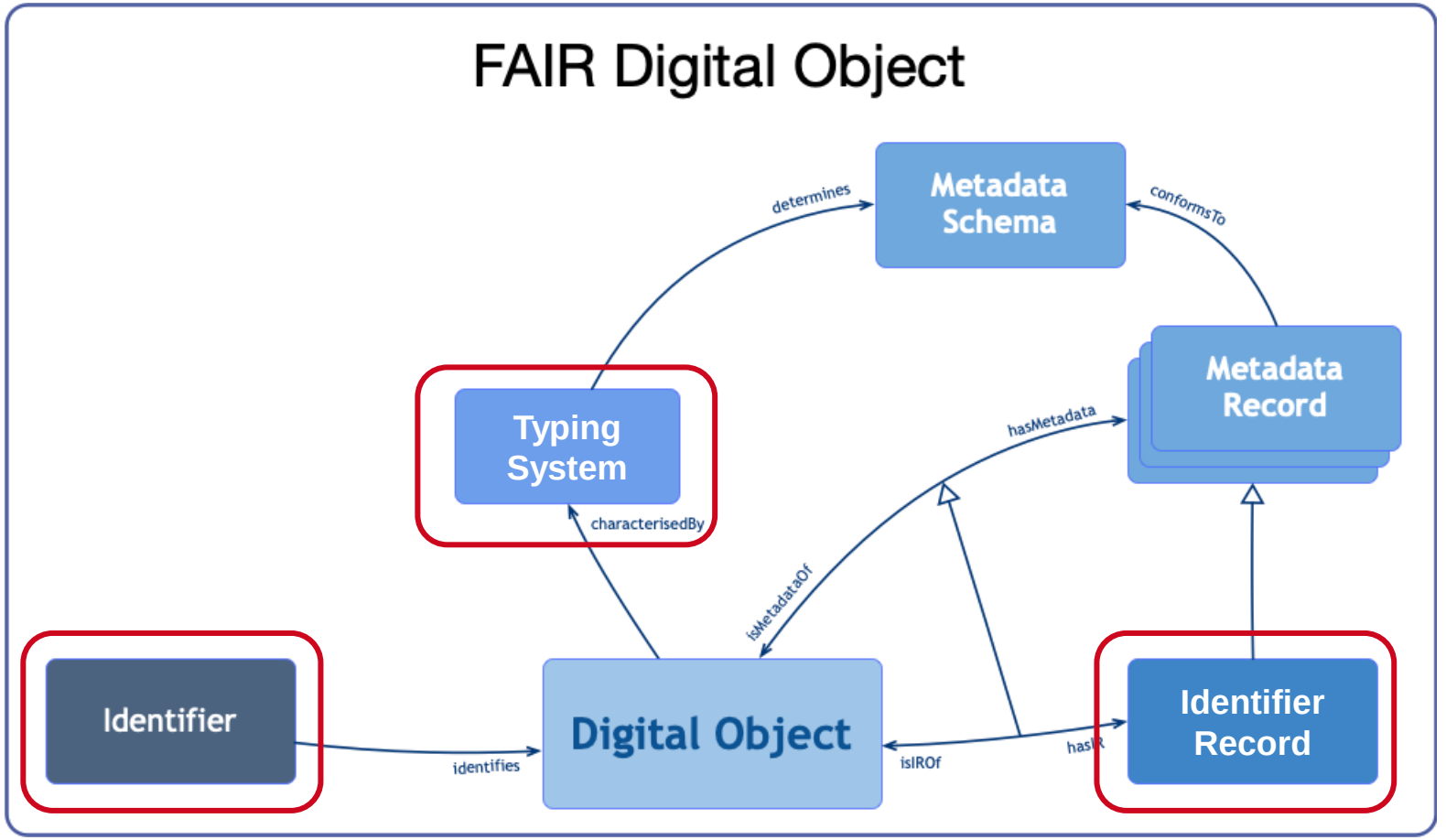
Bonino da Silva Santos, L. O. (2021): FAIR Digital Object Framework Documentation. https://fairdigitalobjectframework.org/

FAIR Digital Object – A Structured Set of Links



Bonino da Silva Santos, L. O. (2021): FAIR Digital Object Framework Documentation. https://fairdigitalobjectframework.org/

FAIR Digital Object – A Structured Set of Links



Bonino da Silva Santos, L. O. (2021): FAIR Digital Object Framework Documentation. https://fairdigitalobjectframework.org/

Ok, nice but…

… is there an app for that?

Ok, nice but…

… is there an app for that?

... well …

Ok, nice but…

… is there an app for that?

... well …

… no …

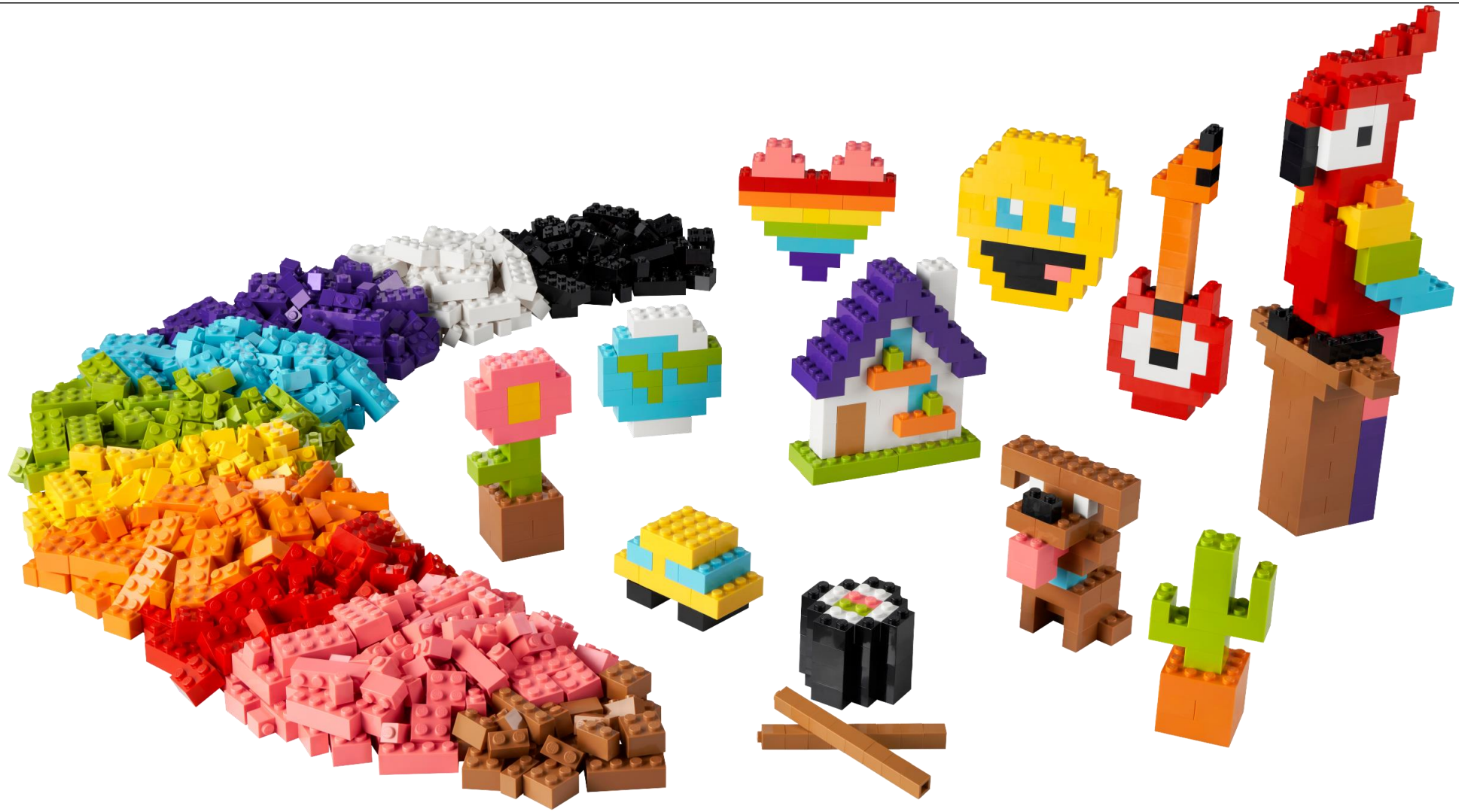
Ok, nice but…

… is there an app for that?

... well …

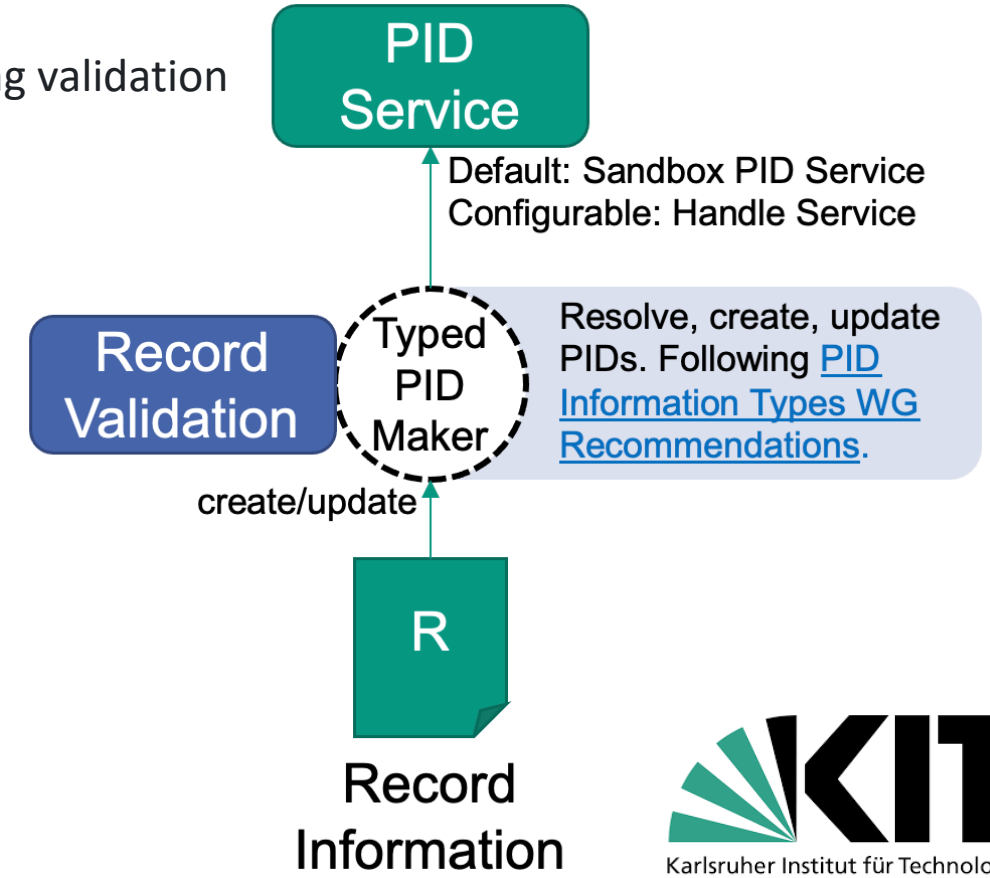
… no …

… and yes …



Create FAIR DOs with Typed PID Maker

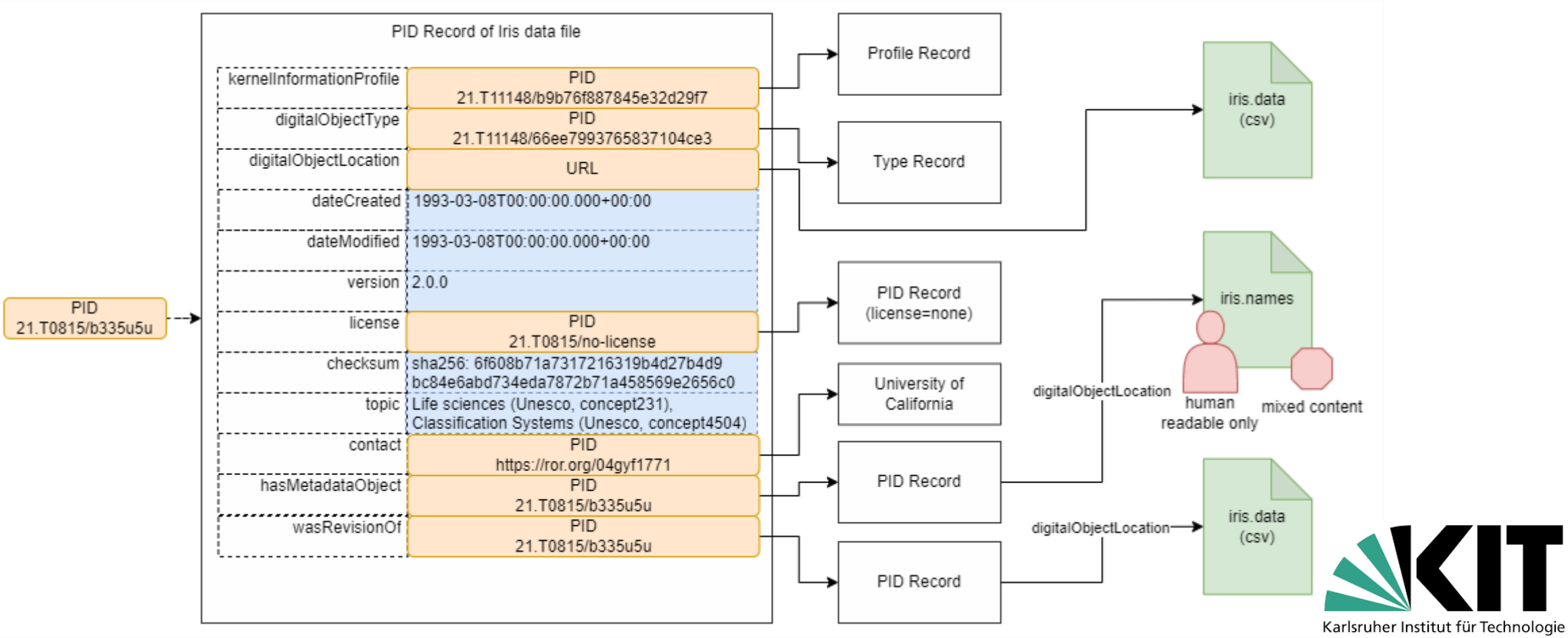
- Enables the creation, maintenance, and validation of PIDs
- Ensures the PID contains typed, machine-actionable information using validation
- Requires a reference to a registered Kernel Information Profile (see recommendations of RDA).
- Uses/Requires Handle PIDs (Handle prefix not included).
- Supports sandboxed PIDs, which require no external service.
- Can be integrated in other applications as REST endpoint.



Karlsruher Institut für Technologie

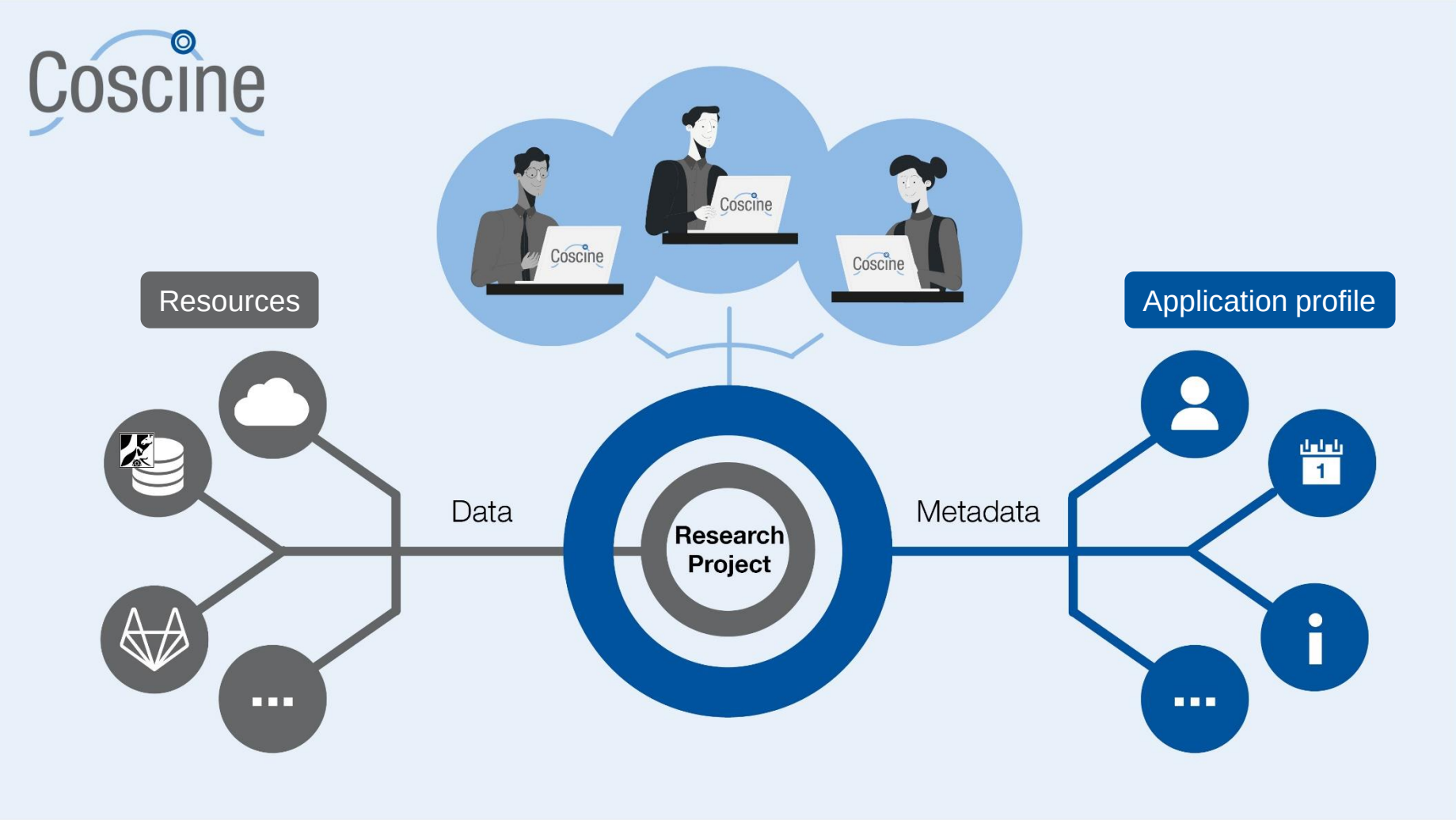
Based on: https://kit-data-manager.github.io/fairdo-cookbook

A Typed PID Record - Typed PID Maker



Based on: https://kit-data-manager.github.io/fairdo-cookbook

Coscine - Short Introduction

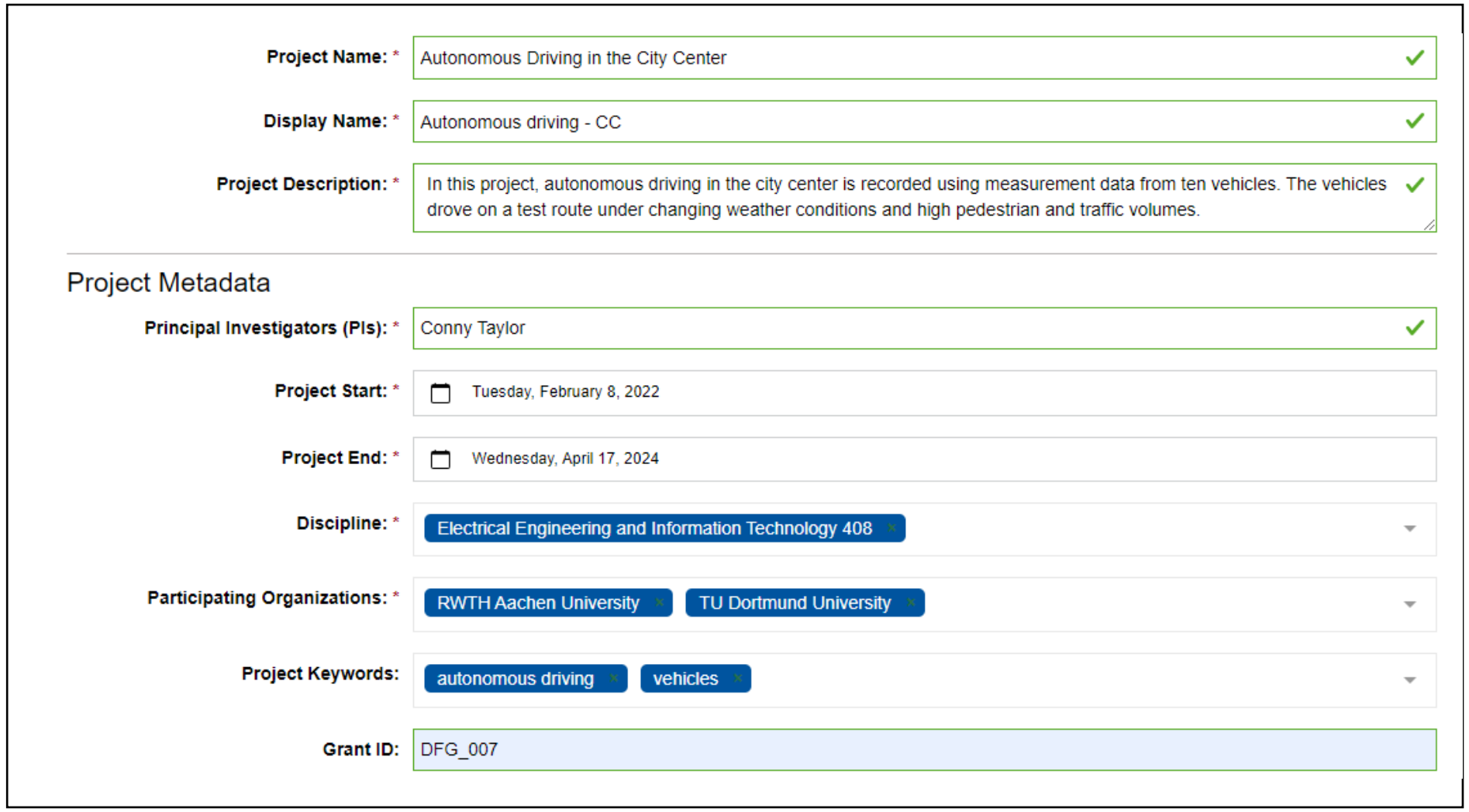


Coscine…

- …maps a project structure
- …integrates different storage systems
- …maintains authorizations at the project level
- …helps to describe all data with structured metadata

Coscine – Metadata Management

1. Project Level

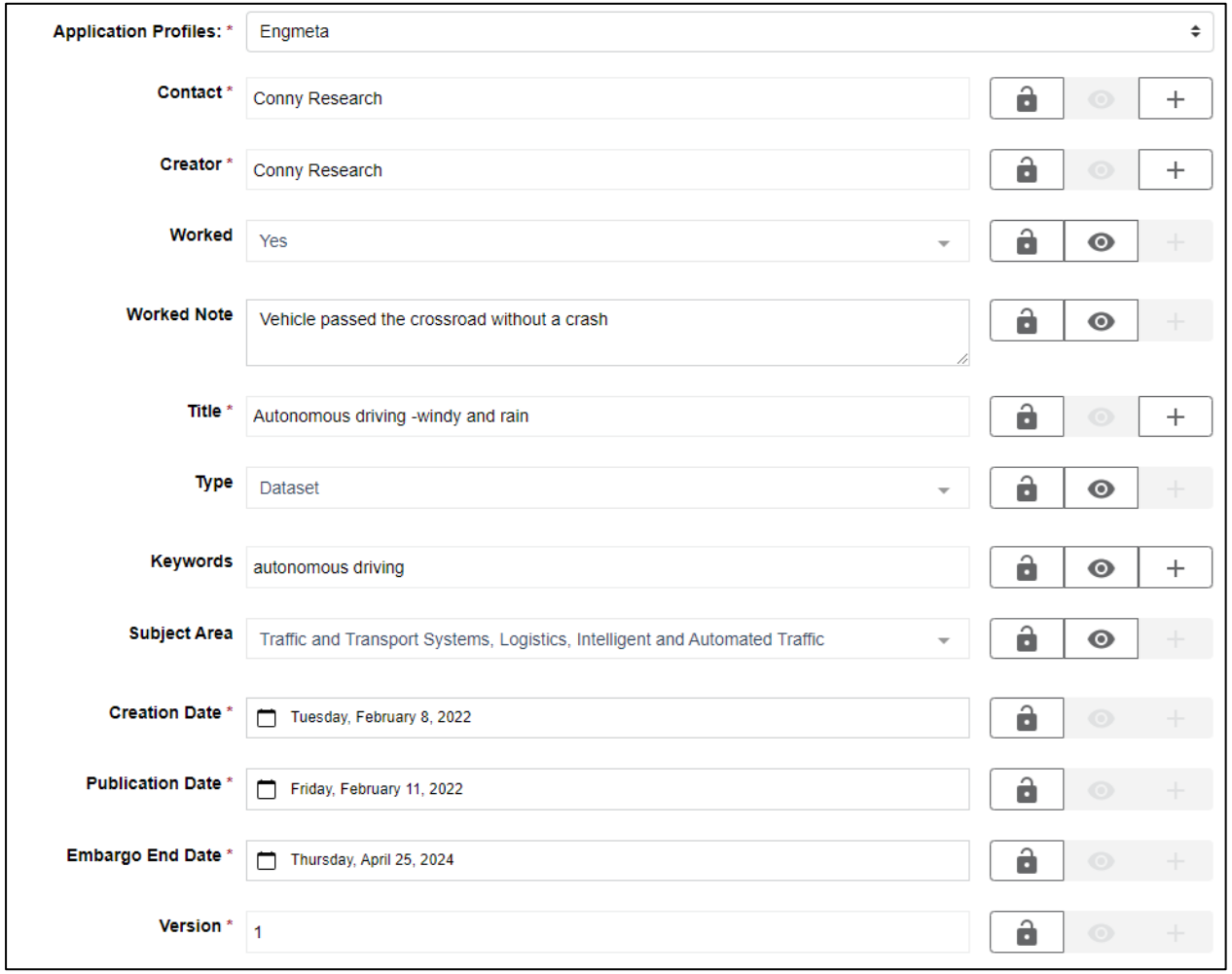


Coscine – Metadata Management

2. Resource Level

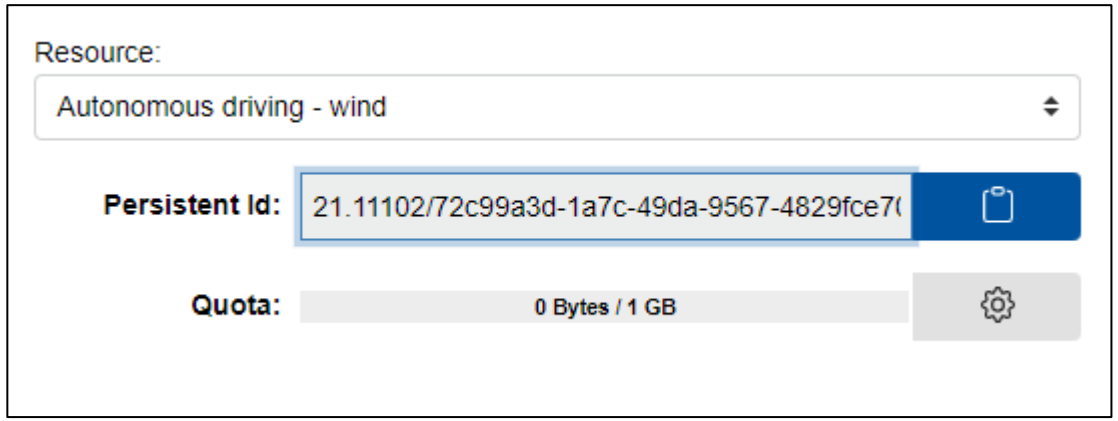
Options:

- **Pre-implemented** application profiles (e.g. EngMeta, see figure)
- Create and request an **individual** application profile



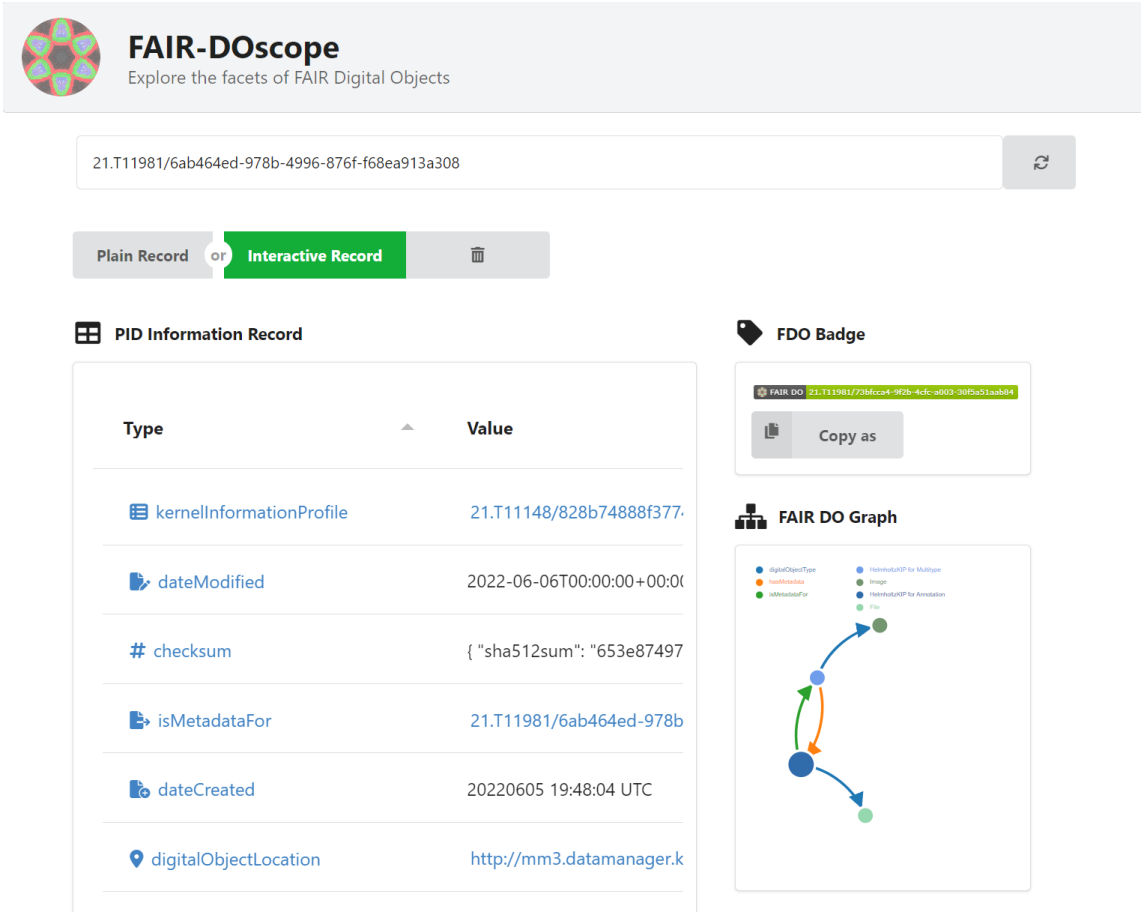
Coscine – Persistent Identifiers (PIDs)

- Coscine uses PIDs to uniquely reference resources
- Each resource in a project is automatically assigned a PID
- The URL contains the handle-prefix followed by a PID
  - Example: *http://hdl.handle.net/21.11102/7599d318-99f3-4385-ace9-7aeb9cf3bXXX*
- PIDs can be used to link resources and make them accessible to others



Explore FAIR DOs with FAIR DO Scope

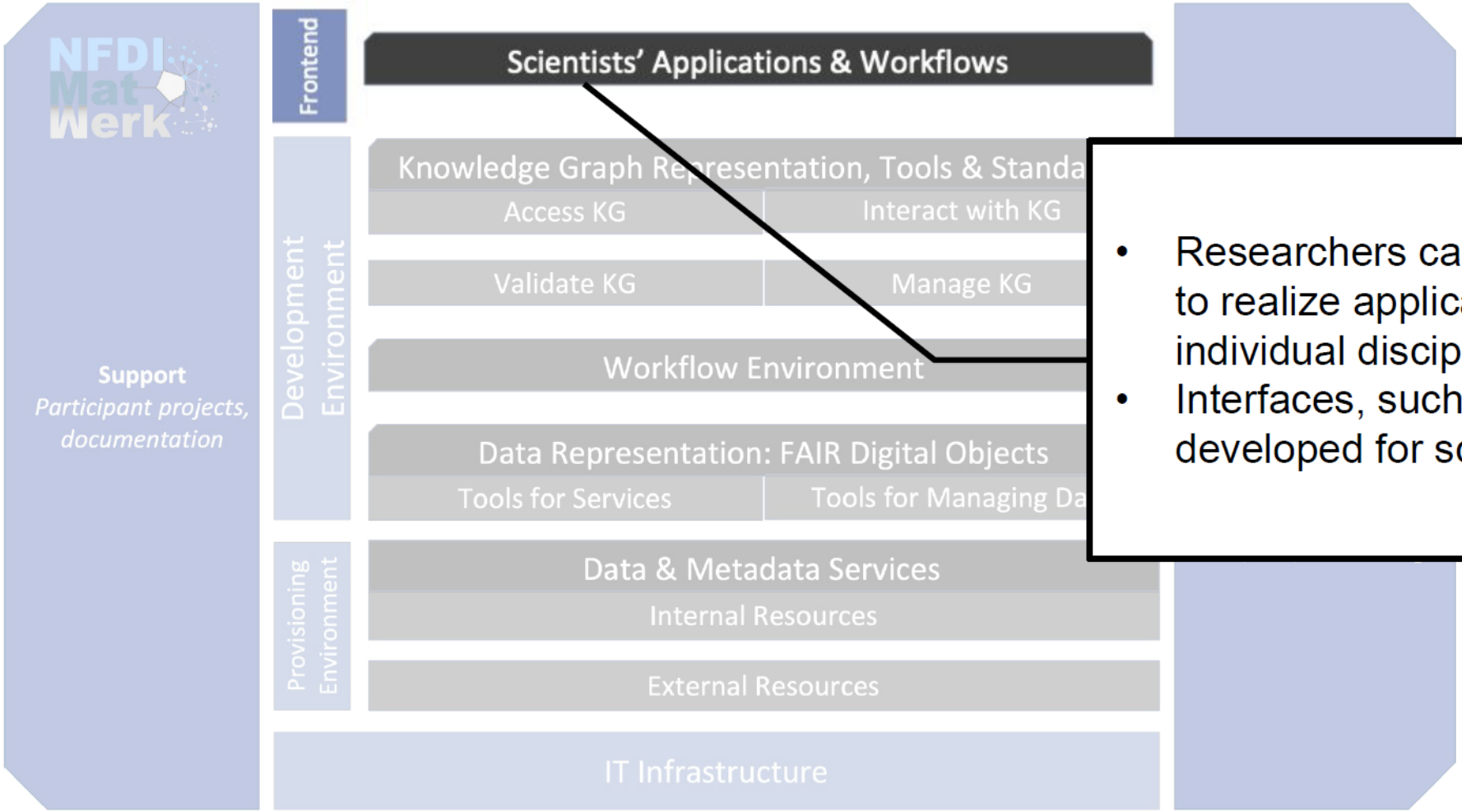
- Easy-to-use, generic FAIR Digital Object viewer and browser
- Presenting the PID record in a graphical and user-friendly way
  - Tabular view and
  - Graphical representation of related FAIR DOs.
- Can be used directly in the browser:
  - https://kit-data-manager.github.io/fairdoscope/



Karlsruher Institut für Technologie

NFDI MatWerk
IT Center
RWTH AACHEN UNIVERSITY

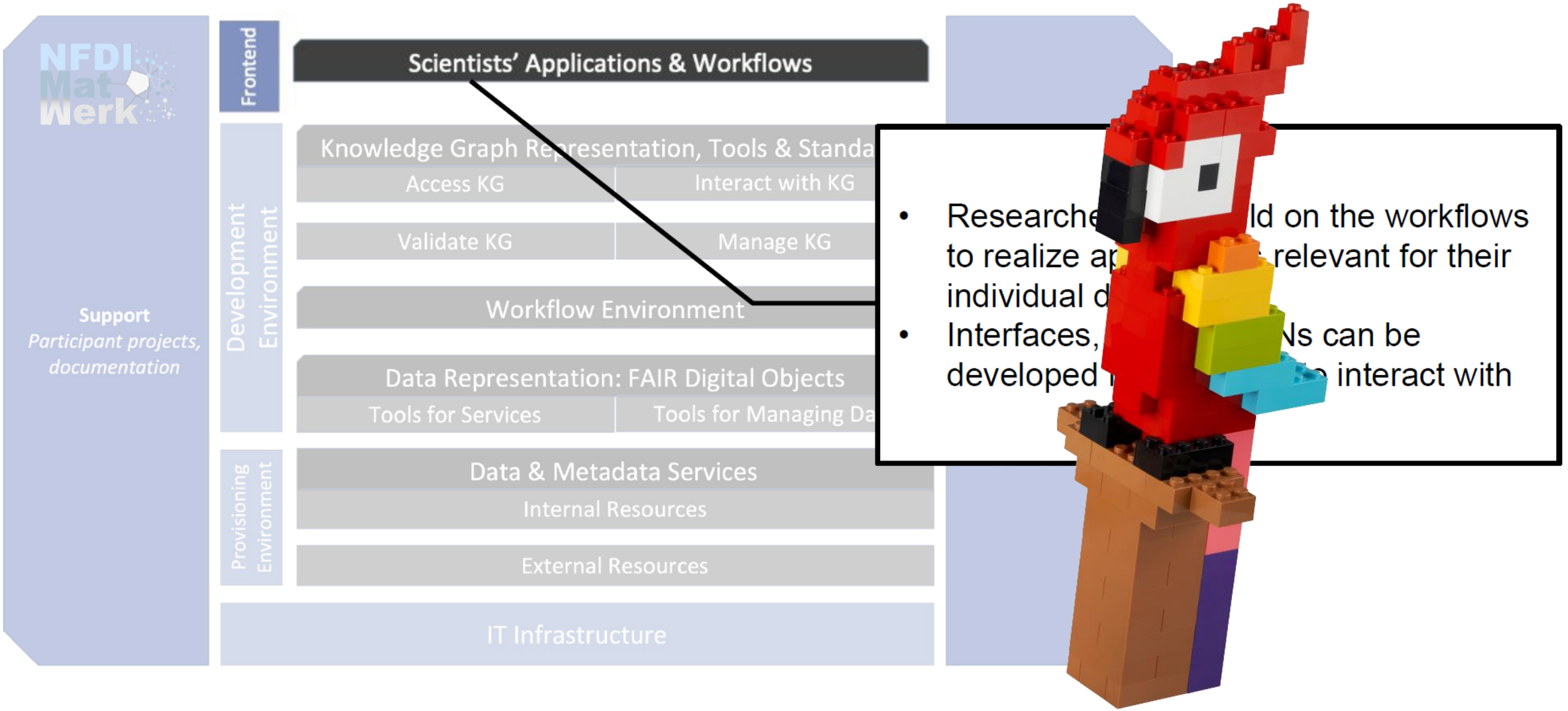
NFDI-MatWerk Shared Service Architecture



- Researchers can build on the workflows to realize applications relevant for their individual discipline
- Interfaces, such as ELNs can be developed for scientists to interact with

M. Politze, Y. Shakeel, S. Hunke, P. Ost, R. Aversa, B. Heinrichs, l. Lang (2023): Long Term Interoperability of Distributed Research Data Infrastructures, https://doi.org/10.52825/CoRDI.v1i.348

NFDI-MatWerk Shared Service Architecture



M. Politze, Y. Shakeel, S. Hunke, P. Ost, R. Aversa, B. Heinrichs, l. Lang (2023): Long Term Interoperability of Distributed Research Data Infrastructures, https://doi.org/10.52825/CoRDI.v1i.348

Vielen Dank
für Ihre Aufmerksamkeit

Dr. Marius Politze
0000-0003-3175-0659

politze@itc.rwth-aachen.de