

Zeyd Boukhers
@ VisGap'25 | Luxembourg

FAIR Datamanagement in the context of scientific computing

The FIT principle

**enabling.
digital.
spaces.**

Fraunhofer FIT designs solutions for digital self-determination, productive value creation and a fair and sustainable society.

We understand humans.

Human beings are at the center of our actions. This is how we ensure that digital technologies are used responsibly for a better world.

We master technology.

We have extensive expertise in the field of digital key technologies. We not only work with technical excellence, but also build applicable technical solutions.

We show profile.

We work independently and with high standards. Together with our partners, we passionately drive digital transformation in business, environment, and society.

We build bridges.

We connect science with practice as well as perspectives from various disciplines. Diversity and interdisciplinarity fuel our creativity and innovative strength.

We own methods.

We have years of experience in applying and developing scientific methods for practical use. We pay attention to details while keeping the bigger picture in mind.

FIT in numbers



Director

Prof. Dr. Stefan Decker



> 24 Mio €
third-party funding
(industrial contracts, national
research programs, EU)



40
years
of experience



6
branch offices
(Sankt Augustin, Aachen, Augsburg,
Bayreuth, Hamm-Lippstadt, Hürth)



> 350
scientists



> 20
professors



> 15
affiliated chairs

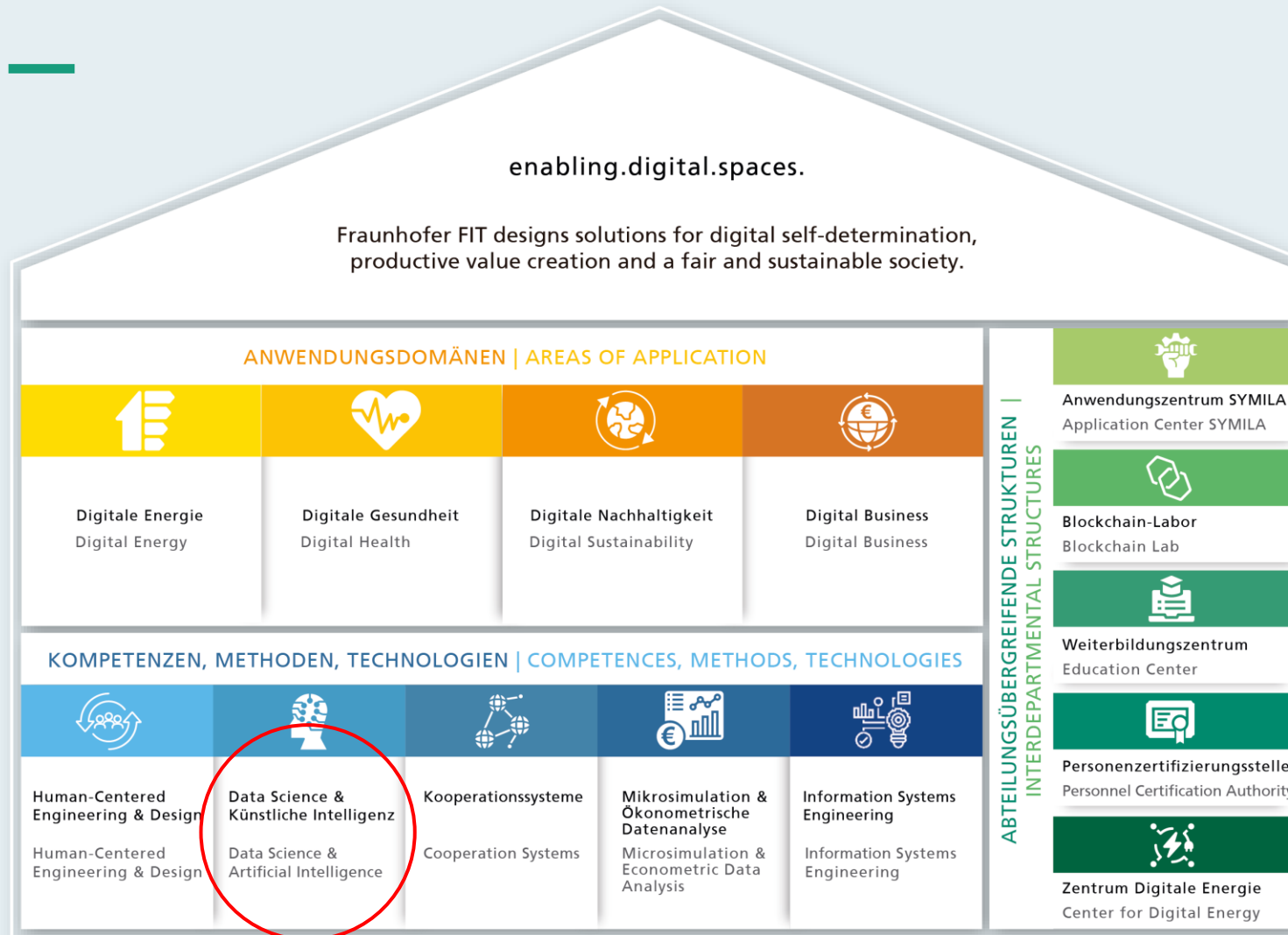


Ø 40
dissertations
per year



Ø 300
publications
per year

The FIT principle



We understand humans.

Human beings are at the center of our actions. This is how we ensure that digital technologies are used responsibly for a better world.

We master technology.

We have extensive expertise in the field of digital key technologies. We not only work with technical excellence, but also build applicable technical solutions.

We show profile.

We work independently and with high standards. Together with our partners, we passionately drive digital transformation in business, environment, and society.

We build bridges.

We connect science with practice as well as perspectives from various disciplines. Diversity and interdisciplinarity fuel our creativity and innovative strength.

We own methods.

We have years of experience in applying and developing scientific methods for practical use. We pay attention to details while keeping the bigger picture in mind.

Data Science and Artificial Intelligence

Dr. Christoph Lange-Bever

Our offer

- **Processes / systems to process, integrate, organize, and analyze data and knowledge**
- **Data infrastructures and ecosystems**
 - Gaia-X, International Data Spaces, National Research Data Infrastructure, European Open Science Cloud
- **Bespoke systems** using **open source** software
- **Professional training** of data scientists

Your benefits

- **Data discovery** and **simplification** of your data landscape
- Bespoke **compatible services**
- **Compliance with privacy laws** and latest **technical standards**
- Better **understanding of your business and production processes**
- **Decision-making** based on **intelligent analyses**
- Medium-term **cost savings** through optimization



Groups

- Data Management (Dr. Christina Gillmann and Prof. Dr. Christoph Quix)
- FAIR Data & Distributed Analytics (Dr. Zeyd Boukhers)
- Data Protection & Sovereignty (Dr. Avikarsha Mandal)
- Process Mining (Daniel Schuster)
- Intelligent Data Analytics (Prof. Dr. Christian Beecks)
- Large Language Models and Knowledge Graphs (Dr. Diego Collarana)
- Learning Center (Dr. Andreas Pippow)

FAIR Data & Distributed Analytics

Dr. Zeyd Boukhers

Our offer

- **Boost FAIR data ecosystems with innovative tools.**
- **Support distributed AI and analytics with expert data science.**
- **Promote data transparency, reproducibility, and FAIRness.**
- **Enhance data management and reuse.**
- **Offer advanced information retrieval and extraction solutions.**
- **Simplify integration of varied data sources for in-depth analysis.**



Dr. Zeyd Boukhers
Machine Learning
and Information
Retrieval
Group Lead

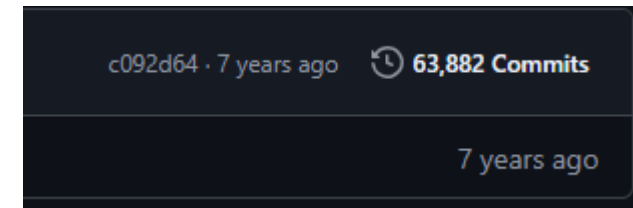
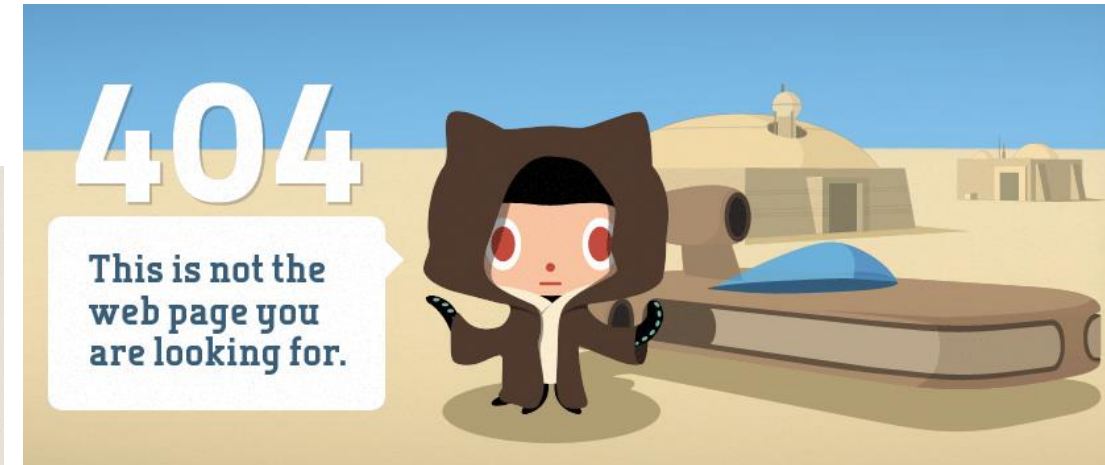
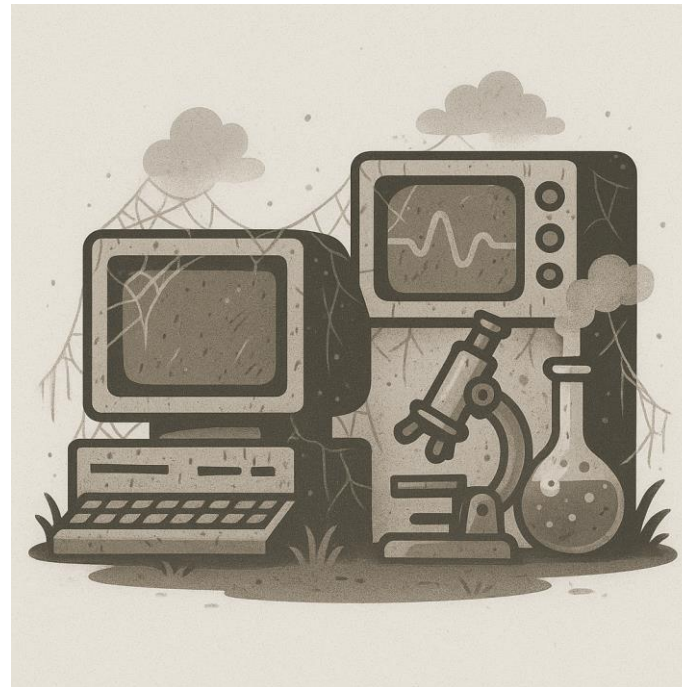
Prof. Dr. Oya Beyan
Director of the Biomedical
Informatics Institute at
University Cologne,
Faculty of Medicine
Scientific Lead



Built. Published. Forgotten.

Why brilliant visualization research stays in the lab?

This repository was archived by the owner on Jan 16, 2025. It is now read-only.



Built. Published. Forgotten.

Some statistics 1/2

- **>50% of GitHub projects “die” within 4–5 years [1]**
- **7.7% of GitHub repos were “dead” as of 2016 (~3.7M out of 48.3M) [2]**
- **5.8% of GitHub repositories in a 2024 dataset had been deleted [3]**
- **76.9% of shared research code is hosted on GitHub, only 5.9% on Zenodo [4]**
- **6.0% of repository URLs in scholarly articles are dead links [5]**
- **12.99% of referenced repos are lost entirely (not in any archive) [5]**

[1]: Ghazi, Badih, et al. "Differentially private all-pairs shortest path distances: Improved algorithms and lower bounds." *arXiv preprint arXiv:2203.16476* (2022).

[2]: <https://www.softwareheritage.org/>

[3]: He, Hao, et al. "4.5 Million (Suspected) Fake Stars in GitHub: A Growing Spiral of Popularity Contests, Scams, and Malware." *arXiv preprint arXiv:2412.13459* (2024).

[4]: Sharma, Nitesh Kumar, et al. "Analytical code sharing practices in biomedical research." *PeerJ Computer Science* 10 (2024): e2066.

[5]: Escamilla, Emily, et al. "Cited But Not Archived: Analyzing the Status of Code References in Scholarly Articles." *International Conference on Asian Digital Libraries*. Singapore: Springer Nature Singapore, 2023.

Built. Published. Forgotten.

Some statistics 2/2

- Only ~10% of shared biomedical research code is in reproducible, structured format [4]
- 74% of R code packages fail to run in clean environments
- 70% of research developers do not use automated testing [6]

[4]: Sharma, Nitesh Kumar, et al. "Analytical code sharing practices in biomedical research." *PeerJ Computer Science* 10 (2024): e2066.

[6]: Eisty, Nasir U., Upulee Kanewala, and Jeffrey C. Carver. "Testing research software: an in-depth survey of practices, methods, and tools." *Empirical Software Engineering* 30.3 (2025): 81.

The hidden cost of prototype abandonment

Some statistics

Research Impact:

- Brilliant algorithms never reach users
- Domain scientists reinvent wheels
- Innovation slows across the field

Scientific Integrity:

- "*Generalizable*" methods that can't be generalized
- Reproducibility crisis in visualization
- Trust gap between research and industry

Resource Waste:

- Millions in funding → one-time prototypes
- Researchers re-implementing existing solutions
- Domain expertise locked in abandoned code

How do we systematically fix this?

What if we had scientific principles for software?

FAIR Principales

 Access these slides



The good news: we already do

FAIR

Findable • Accessible • Interoperable • Reusable



The Role of FAIR Principles in Data Utility

What are FAIR Principles?

- **Findable:** High-quality data is of little use if it cannot be located. Metadata and data discovery mechanisms are essential.
- **Accessible:** Once found, data needs to be accessible under well-defined conditions, ensuring both security and ease of use.
- **Interoperable:** Data must be compatible with other datasets, tools, and workflows, which is facilitated by standardized formats and vocabularies.
- **Reusable:** The ultimate goal is for data to retain its quality over time and be usable in different contexts, which requires comprehensive documentation and clear data usage licenses.



FAIR: Findability

Data is 'Findable' when it is easily searchable and accessible, where users and machines can locate needed resources without undue effort.

- **F1: Unique, durable identifiers for all data assets.**



- **Example:** Use UUIDs to distinguish between datasets
- **Benefit:** Facilitates traceability of data lineage and simplifies data management across multiple systems and projects.

- **F2: Data characterized by detailed metadata.**



- **Example:** Include all what describe the data
- **Benefit:** Enhances understanding of data context, improving the quality and speed of data integration and analytics.

- **F3: Data and metadata indexed in searchable resources.**



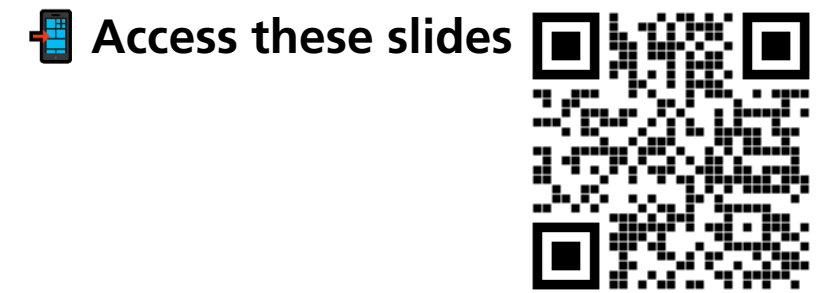
- **Example:** Implement an Elasticsearch cluster for querying vast datasets.
- **Benefit:** Streamlines data retrieval, saving time and resources, and enabling quicker decision-making processes.

- **F4: Direct correlation of metadata to data identifiers.**



- **Example:** Use a metadata management tool that automatically pairs UUIDs with their metadata entries.
- **Benefit:** Reduces ambiguity, ensuring that users always access the correct version of the dataset for their AI applications.

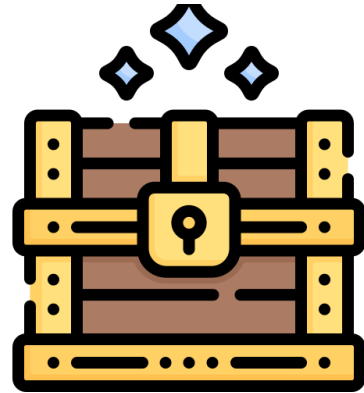
Challenges?
Solutions?
Actionable Steps?



« *Data that describe Data* »

It include:

- How the data was collected, processed and stored,
- Creator details,
- Use permissions,
- etc.



Data



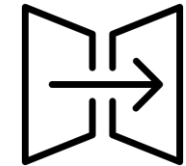
Metadata

FAIR: Accessibility

Accessibility ensures that data and metadata are retrievable and usable when needed, a vital component in the data lifecycle.

- **A1: Standardized protocol for data retrieval.**

- **Example:** HTTP, FTP, SMTP, ...
- **Benefit:** Guarantees consistent and reliable access to data for AI models across the organization.



- **A2: Metadata remains accessible even if data is not.**

- **Example:** Even when customer datasets are archived or deleted for privacy compliance, the metadata remains in a central repository, providing context for AI models trained on historical data.
- **Benefit:** Maintains knowledge and lineage of AI datasets, which is critical for long-term AI projects.



*Challenges?
Solutions?
Actionable Steps?*

FAIR: Interoperability

Interoperability ensures that data can be integrated and leveraged across various platforms and applications, crucial for collaborative AI developments.

- **I1: Adoption of standard data formats.**

- **Example:** Implementing JSON-LD for structuring and linking data, which enhances machine-readability across different systems.
- **Benefit:** Facilitates the integration of datasets from various sources and ensures seamless data exchange and processing in AI applications.

- **I2: Use of shared vocabularies and ontologies**

- **Example:** Applying schema.org standards to enrich metadata, which facilitates common understanding and data sharing.
- **Benefit:** Promotes a unified understanding of data across different domains, which enhances collaboration and reducing misinterpretation in AI modeling.

- **I3: Cross-referencing data points.**

- **Example:** Creating RDF triples that link dataset elements to external standardized data
- **Benefit:** Strengthens the richness of data context and facilitates integration with other data types and systems, ultimately optimizing data utilization.



FAIR: Reusability

Reusable data is easily repurposed for future projects, ensuring long-term value and sustainability.

- **R1: Rich and relevant attributes in metadata.**

- **Example:** A dataset from sales module, with detailed transaction histories, is reused for predictive analytics in inventory management.
- **Benefit:** Encourages diverse applications, enabling data to serve multiple purposes and AI projects.



- **R1.1: Data accompanied by clear licensing.**

- **Example:** A HANA database is shared across different business units under a specific license that details permissible uses and modifications.
- **Benefit:** Clarifies usage rights and restrictions, reducing legal barriers to data utilization.



- **R1.2: Traceable data provenance.**

- **Example:** A sales dataset is reused to enhance a forecasting AI, with its edit trail clarifying data transformations.
- **Benefit:** Fosters transparency for data's history and usage rights, essential for informed reuse.



- **R1.3: Adherence to community and domain standards**

- **Example:** Standardized HR data easily integrates with SAP SuccessFactors for AI-driven talent insights.
- **Benefit:** Ensures data is compliant and readily integrated into industry-specific AI applications.



FAIR Principales: Summary

- **Findable**

- Persistent ID
- Metadata online

- **Accessible**

- Data online
- Restrictions where needed

- **Interoperable**

- Use standards, controlled vocabs
- Common (open) formats

- **Reusable**

- Rich documentation
- Clear usage licence

- FAIR data does not have to be open
- Data can be shared under restrictions & still be FAIR
- Making data FAIR ensures it can be found, understood and reused
- Open data is a subset of all the data shared

"As open as possible, as closed as necessary"

FAIR ≠ Open

Source: The FAIR Data Concept by « Sarah Jones », <https://www.slideshare.net/sjDCC/fair-data>

Implementing FAIR

- **How can one rigorously adhere to FAIR principles?**
 - FAIR principles serve as guideline, so they flexible rather than rigid, encourage aligning data practices with their core values for better management and integration.
- **What technology is essential to learn before implementing FAIR?**
 - FAIR doesn't require specific tech but benefits from knowing data tools and standards like JSON, XML, and APIs, plus semantic web concepts.
- **Does the implementation of FAIR principles necessitate substantial investment?**
 - Adopting FAIR principles requires some level of investment in systems and training but pays off with greater data efficiency and innovation potential.

Enhancing Research and Development with FAIR Principles

- **Data Accessibility and Discovery:**
 - Rapid discovery and use of data via advanced tools → enhancing meta-analyses and benchmarking for scientific discovery.
 - Promotes industry engagement and publisher involvement to adopt FAIR principles → increasing innovation and data sharing.
- **Reproducibility and Reporting:**
 - Ensures reproducibility of AI models → establishing trust and validity.
 - Facilitates effective reporting and application → aiding in clear methodological explanations and reducing biases.
- **Infrastructure and Model Standardization:**
 - Supports development of data infrastructures that accept and store FAIR and AI-ready data → enabling seamless integration.
 - Encourages the creation of interoperable data infrastructures and standardized AI models for uniform performance assessments.

Source: Huerta, E. A., et al. "FAIR for AI: An interdisciplinary and international community building perspective." Scientific data 10.1 (2023): 487.

Setting New Standards in AI Performance and Ethics

- **Research and Architecture Innovation:**

- Drives computer science and AI research on efficient surrogate architectures and models, leveraging FAIR data for reliable performance predictions.
- Enhances generalization capabilities of biomedical AI models by training on diverse, FAIR datasets, improving health outcomes.

- **Scientific Integrity and Bias Reduction:**

- Establishes scientific correctness where full reproducibility is not possible, ensuring the integrity of research findings.
- Aids in identifying and mitigating potential biases in AI models, particularly in life sciences and healthcare, fostering fair and unbiased applications.

Source: Huerta, E. A., et al. "FAIR for AI: An interdisciplinary and international community building perspective." Scientific data 10.1 (2023): 487.

Comparative Impact of FAIR Principles in AI Systems

AI Eco-System not adhering to FAIR Principles

Challenges:

- Difficulty sourcing data due to poor findability
- Challenges in accessing and integrating heterogeneous data
- Limited data reusability across projects
- Case example: An AI system that struggled with dataset integration leading to delayed project timelines.

AI Eco-System adhering to FAIR Principles

Benefits:

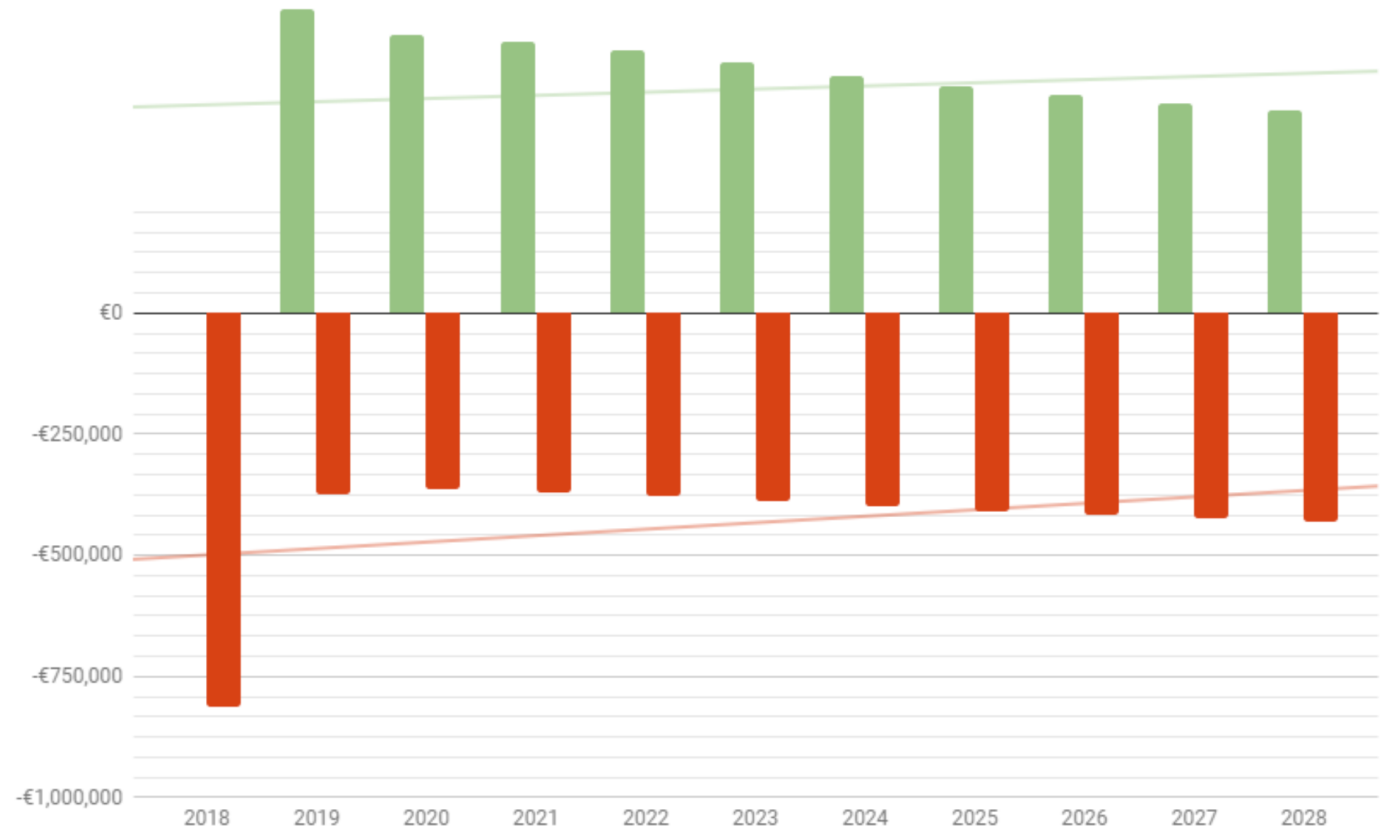
- Quick data discovery and streamlined access
- Smooth integration with existing data systems and standards
- Enhanced ability to repurpose datasets, reducing redundancies
- Case example: An AI system that benefited from interoperable datasets, resulting in faster development and collaboration.

FAIR vs. Non FAIR

- “In a study which preceded this report, the cost of not having FAIR data for the EU-28 has been estimated at EUR 10,2 bn per year, and this is bound to grow unless action is taken.”

Cost-Benefit analysis for FAIR research data – Policy recommendations, EU Commission 2018

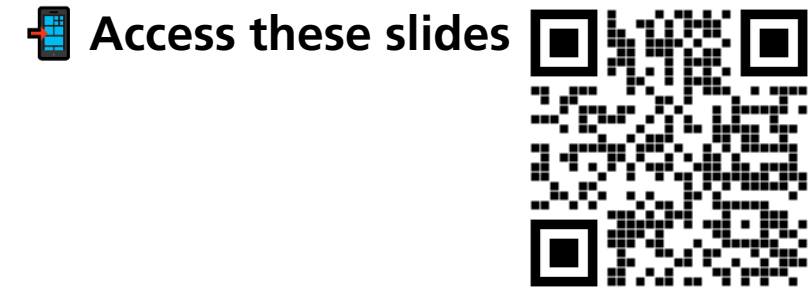
Benefits and costs of applying the FAIR principles (€)



Source: Cost-Benefit analysis for FAIR research data – Policy recommendations, EU Commission 2018

FAIR Data

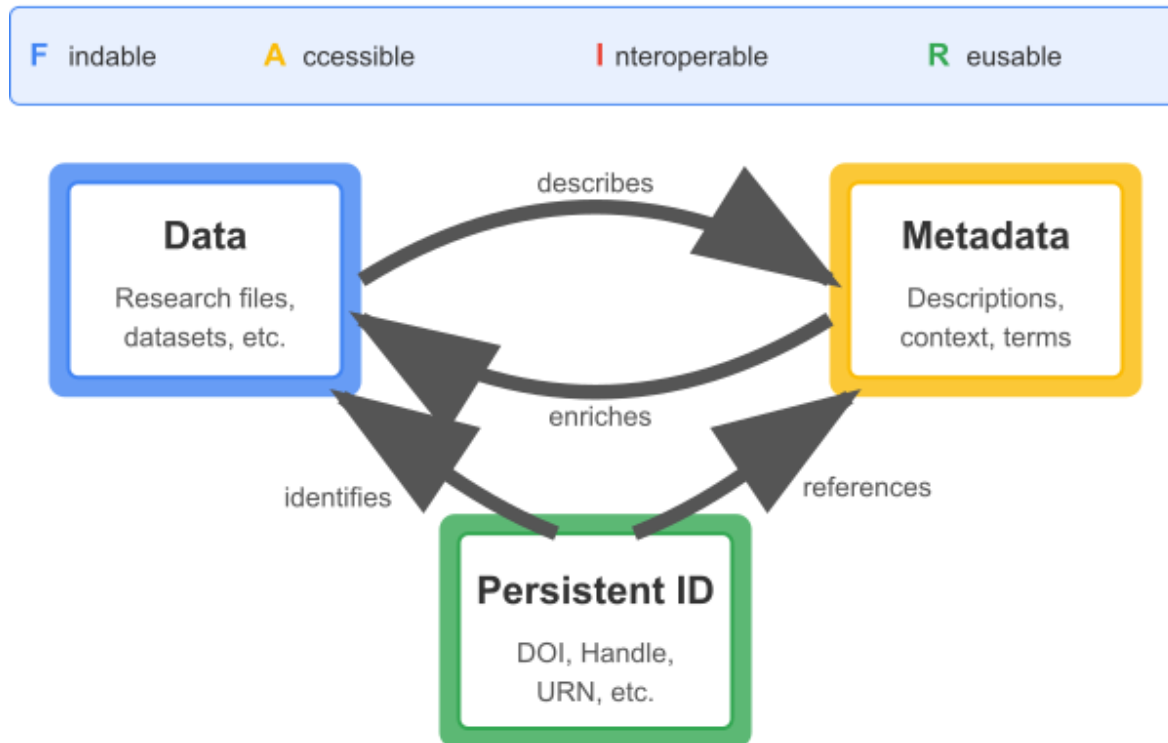
Starting point



- **FAIR Principles Reminder:**

- **F**indable: Data has rich metadata and persistent identifiers
- **A**ccessible: Data can be retrieved using standardized protocols
- **I**nteroperable: Data uses formal, accessible vocabularies
- **R**eusable: Data has clear usage licenses and provenance information

How Data, Metadata, and Identifiers Relate

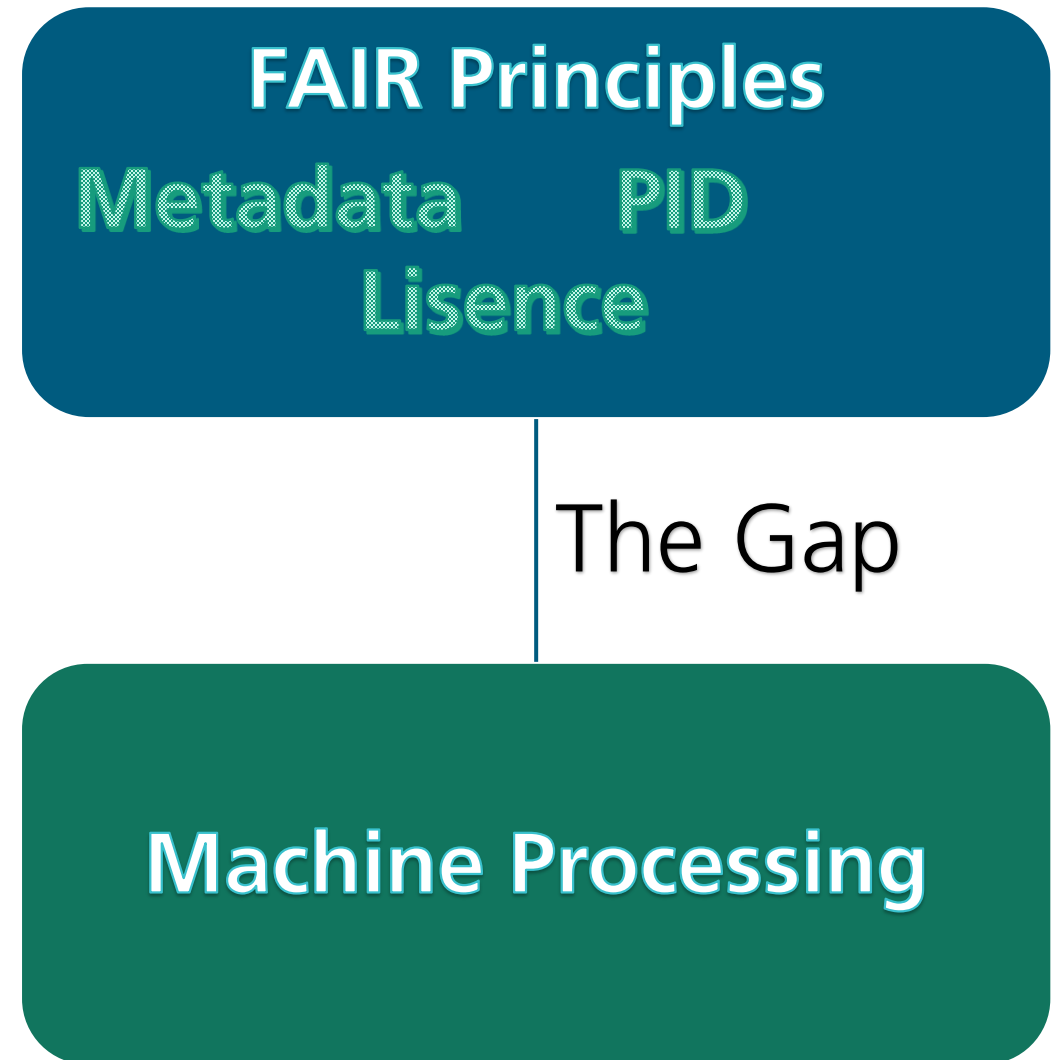


FAIR principles provide flexible guidelines rather than rigid technical requirements
FAIR ≠ Machine actionable by default

Why Machine Actionability Matters

The gap

- Repositories have metadata, but it is not usable by software without custom logic,
- Research tools cannot discover or act on datasets autonomously.
- Rich semantics and operations are not encoded or discoverable.



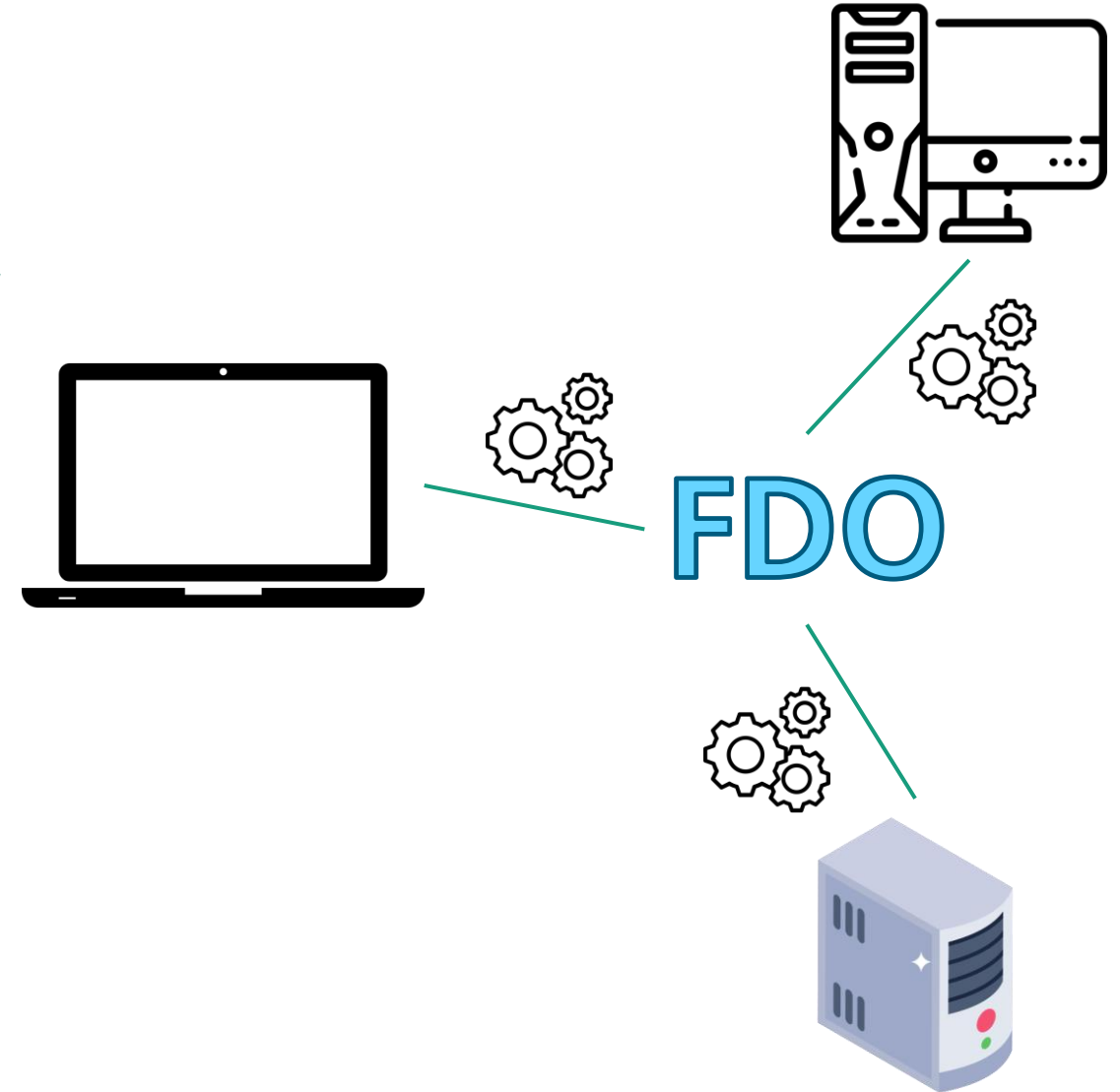
Machines can read FAIR data, but they cannot act on it!

What is a FAIR Digital Object (FDO)?

The gap

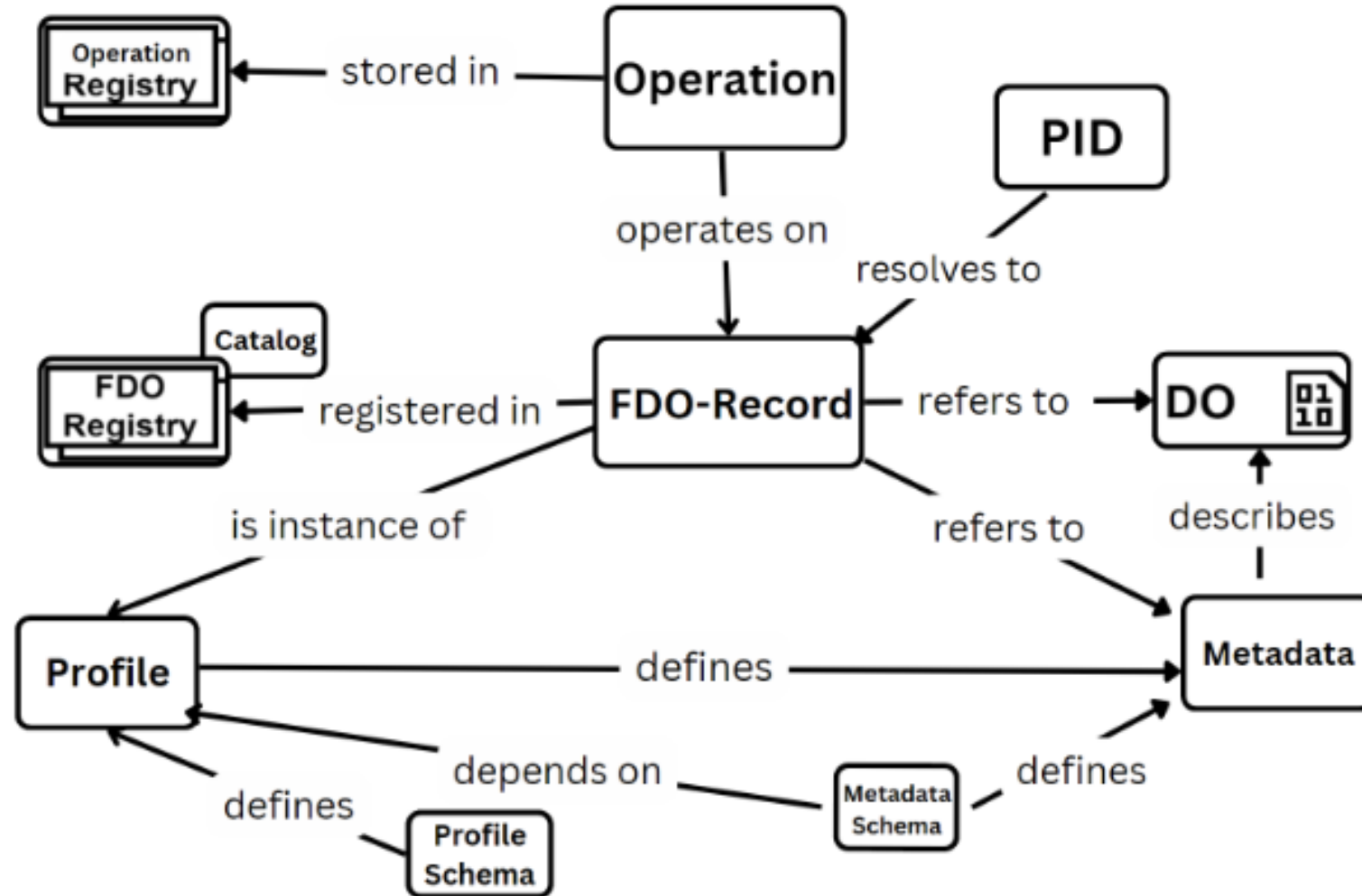
A FAIR Digital Object is a typed, persistently identified, structured unit that contains or references data and metadata in a way that is machine-actionable

Machine-actionable means: machines can discover, interpret, verify, and operate on the object without human guidance.



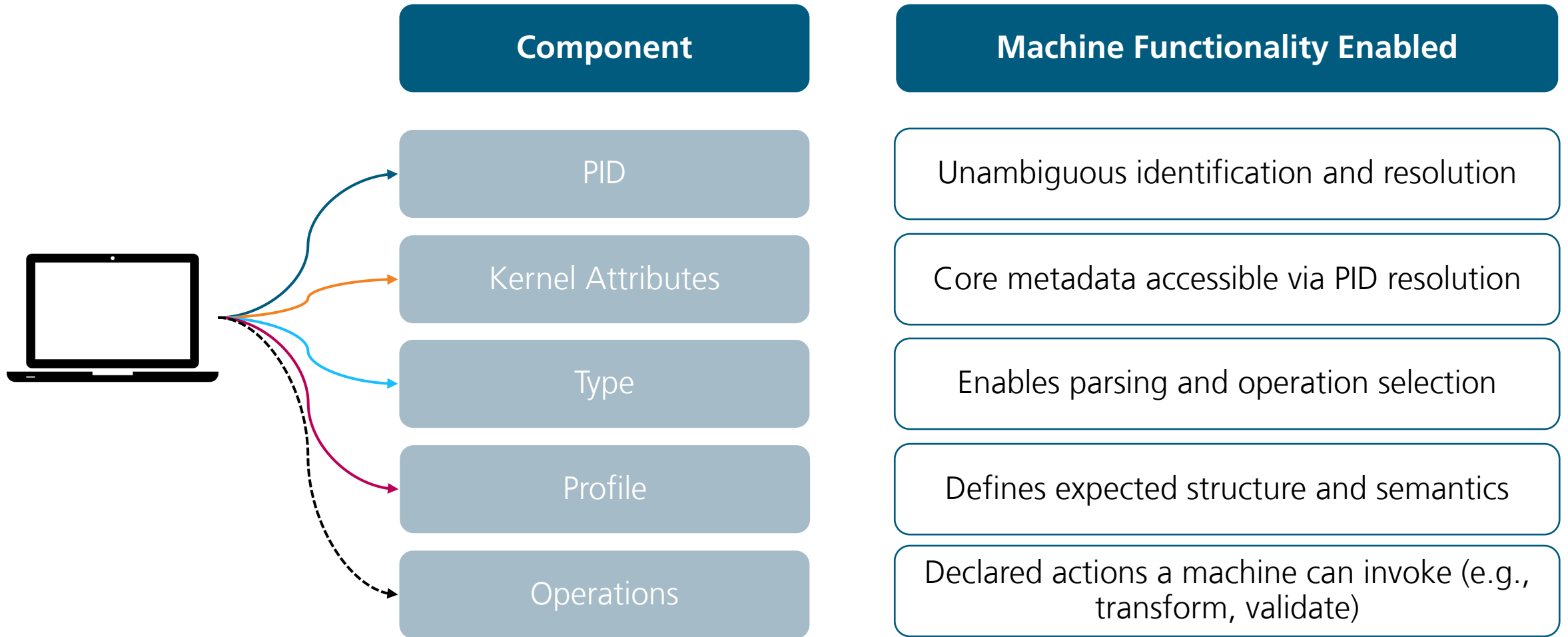
What Makes It Machine-Actionable

Inside an FDO



What Makes It Machine-Actionable

Inside an FDO



Manuscript Annotation Data

Practical Example

Traditional FAIR Approach

Canterbury Tales Manuscript

(Image file)

DOI: 10.1234/manuscript.5678

Repository Landing Page

Title: Canterbury Tales MS

Creator: British Library

Download ZIP

FDO Approach

Canterbury Tales Manuscript

(same Image file)

DOI: 21.14123/fdo.tc-ms-01

PID Resolution Returns:

contentType: image/tiff+tei

type: MedievalManuscript

profile: tei-transcription-v1.2

Access Data

Manuscript Annotation Data

Practical Example

Traditional FAIR Approach

What Machines can « see »

Just a download link

Some human-readable metadata

NO processing instructions

FDO Approach

What Machines can « see »

This is a TEI-encoded manuscript

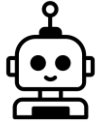
Use tei-transcription-v1.2 parser

FDO for a Visualization Tool

VolumeRenderer3D



Access these slides



Machine Agent

Follows

PID/DOI: 21.14123/fdo.vol-ren3D-01

Resolves to

FDO Record

PID: 21.14123/fdo.vol-ren3D-01

type: 21.14123/type:Software

profile: 21.14123/profile:VisTool

metadata: 21.14123/metadata:vol-ren

Bit sequence: 21.14123/tool:vol-ren

Defines

Operations

Op1: 12.000/op.Install

Op2: 12.000/op.Execute

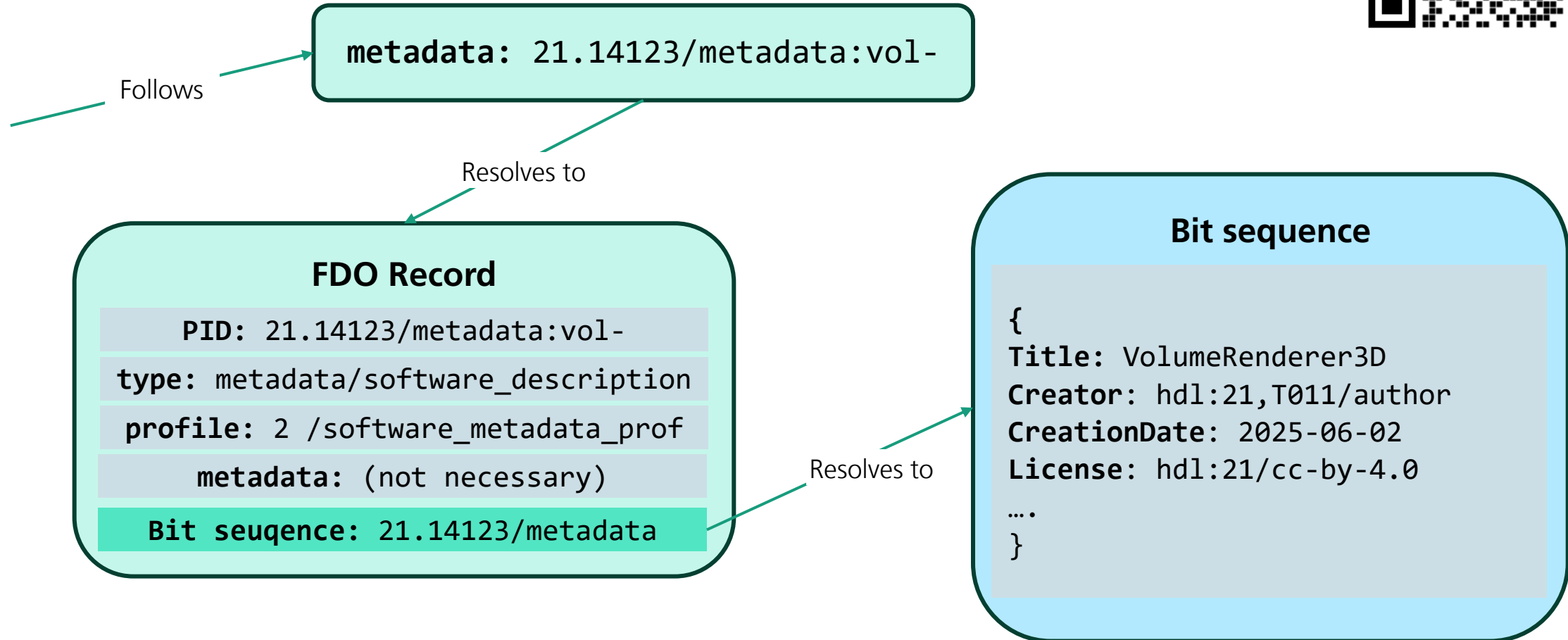
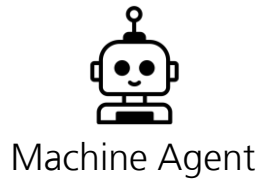
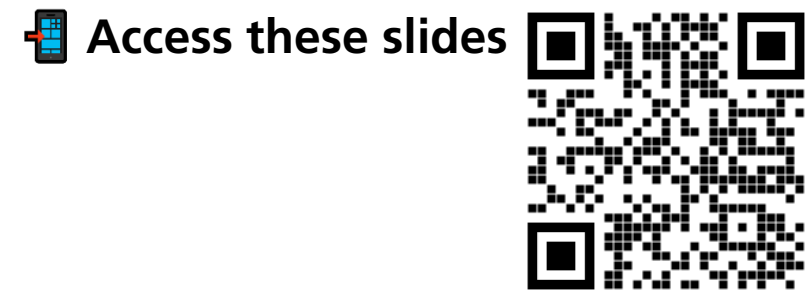
Op3: 12.000/op.Validate

metadataSchema:
21.14123/schema-software

Follows

FDO for a Visualization Tool

VolumeRenderer3D



How FDOs solve the problem

Problem	FDO Solution
Prototype dies when student graduates	FDO persists via institutional PID
"Works on my machine"	FDO specifies environment and dependencies
Unknown input/output formats	FDO defines expected data structure via typed metadata
Broken demo/download links	FDO ensures persistent resolution of content and metadata
Can't integrate with other tools	FDOs support typed operations for workflow composition

FDOs turn fragile prototypes into reusable, interoperable, machine-actionable and persistent digital assets!

FDOs and Agentic AI

Why agentic AI needs FDOs?

Agentic AI can autonomously make decisions and take actions in dynamic environments. It is capable of:

- Perceives context
- Understands options
- Takes meaningful, autonomous actions

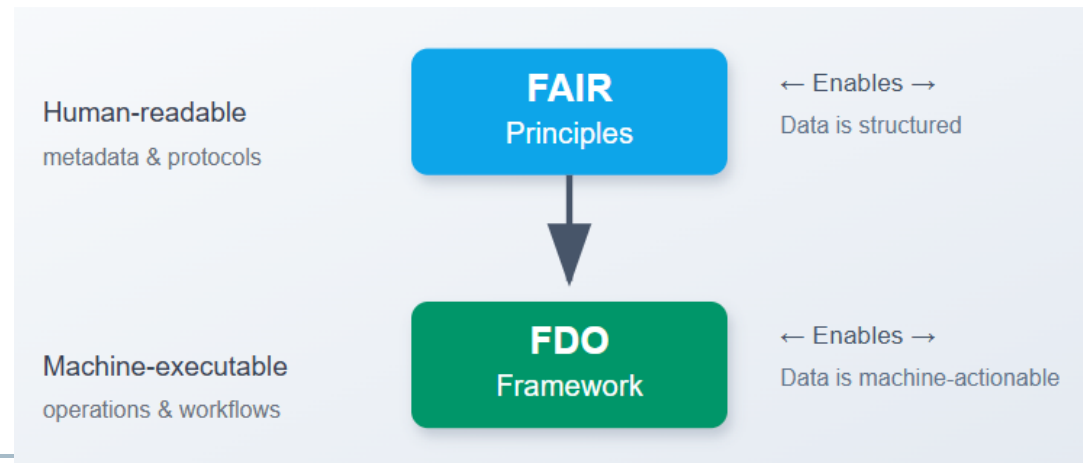
FDOs provide:

- Standardized structure + semantics
- Actionable records
- Discoverable types and operations

FAIR vs. FDO

What is the difference?

	FAIR	FDO
Objective	Make data understandable by humans	Make data processable by machines
Focus	Findability, Accessibility, Interoperability, Reusability	Machine-actionable interpretation and automation readiness
Limitation	Human mediation required for use and integration	Requires orchestration but enables automation

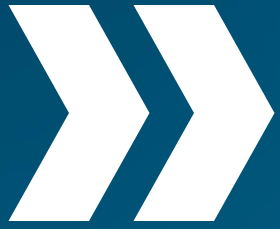


Digital Objects: Where We Are & What's Next

Why agentic AI needs FDOs?

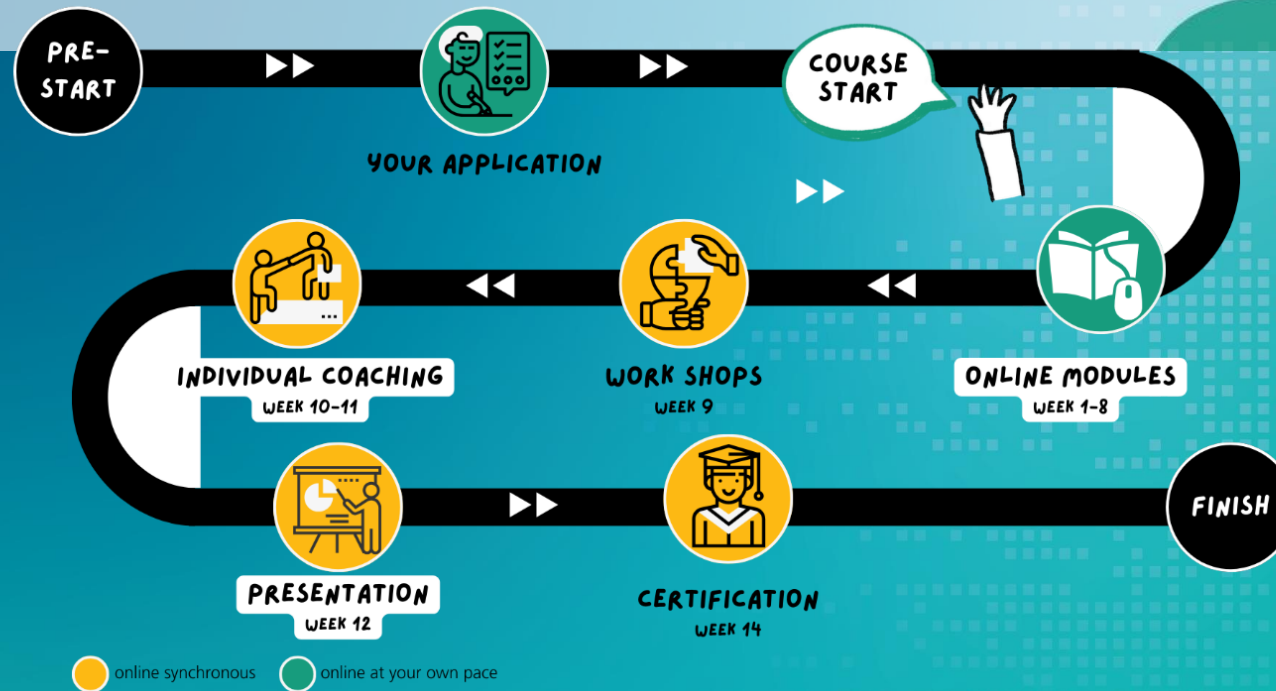
Current Developments

- *FDO Forum*: an international consortium of experts from research institutions and infrastructure initiatives dedicated to specifying and promoting FAIR Digital Objects.
- *NFDI4DS*: integrates FDOs into research software and data management practices, emphasizing digital sovereignty and sustainability.
- *FDO One*: a testbed for implementing FDOs across various disciplines, focusing on harmonizing persistent identifier systems and enhancing machine actionability.
- *FDO Connect*: aims to establish a robust framework for managing and linking research artefacts—such as datasets, publications, and software—by representing them as FAIR Digital Objects.



FAIR Data

Join the community of
FAIR Data Ambassadors



Kontakt

Dr. Zeyd Boukhers
Lead of the research group „FAIR Data & Distributed Analytics“
Department „Data Science and Artificial Intelligence“
Tel. +49 2241 143-735
Fax +49 2241 143-735
zeyd.boukhers@fit.fraunhofer.de
<https://fit.fraunhofer.de/fdda>

Fraunhofer FIT
Schloss Birlinghoven
53757 Sankt Augustin
www.fit.fraunhofer.de



Fraunhofer Institute for Applied
Information Technology FIT

Thank you for your attention!