Designing a FAIR-Compliant Repository System

Dr. Bajarang Kumbhar

M.Sc, PhD, PostDoc (IIT Bombay)

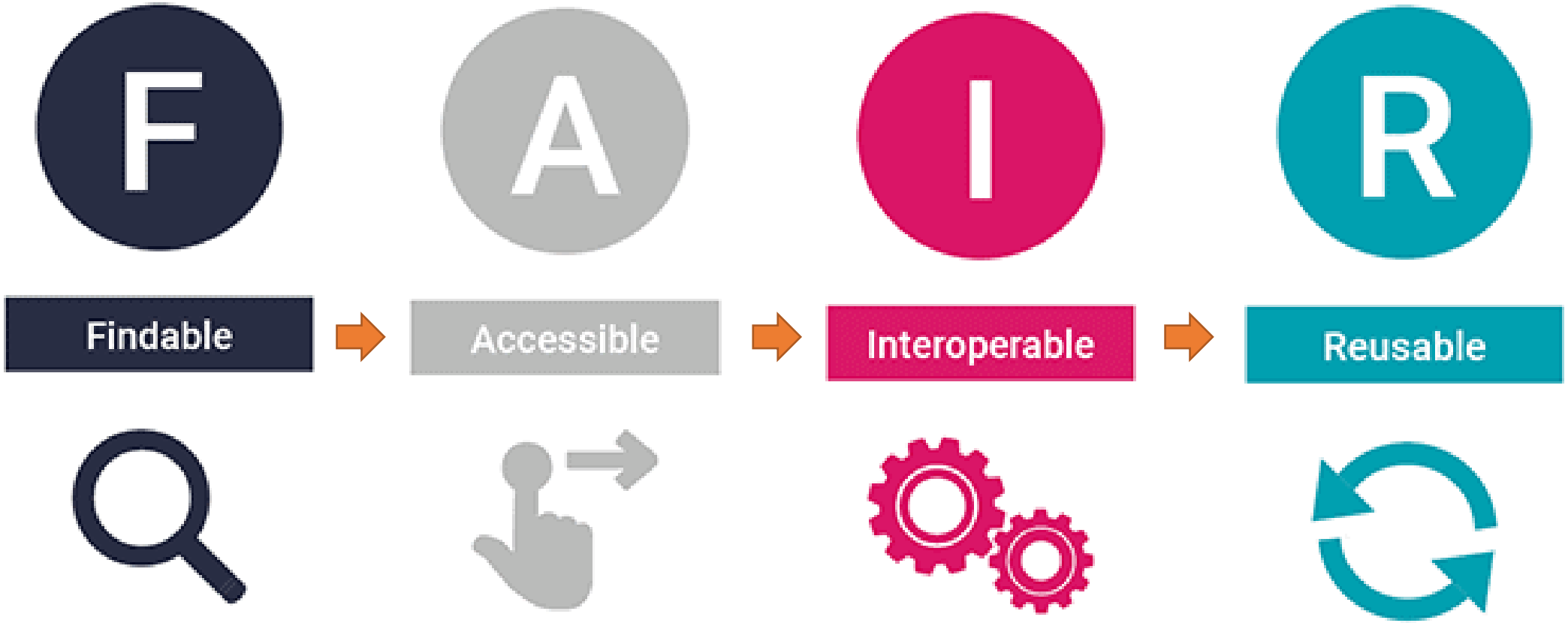
Assistant Professor

Sunandan Divatia School of Science,

SVKM's NMIMS (Deemed to-be) University,

Mumbai, Maharashtra, India

What is FAIR Data?



Building a system where data is **easy to** **Search, Access, Share, and Reuse**

Importance of FAIR System

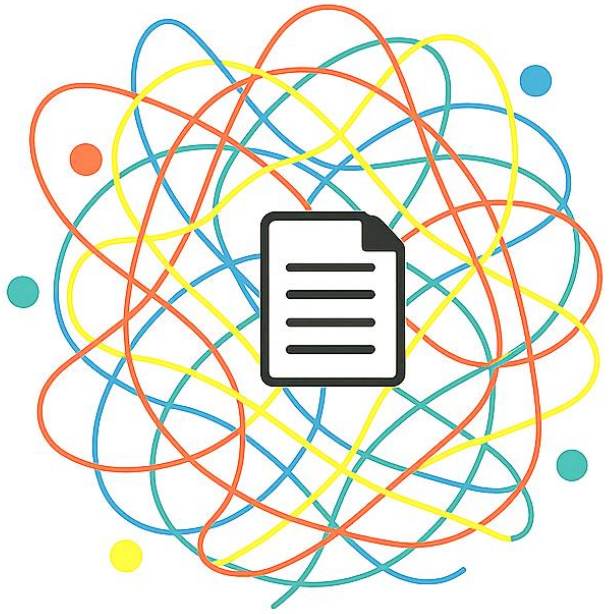
Data Silos

Lack of Discoverability

Poor MetaData Quality

Importance of FAIR System

BEFORE



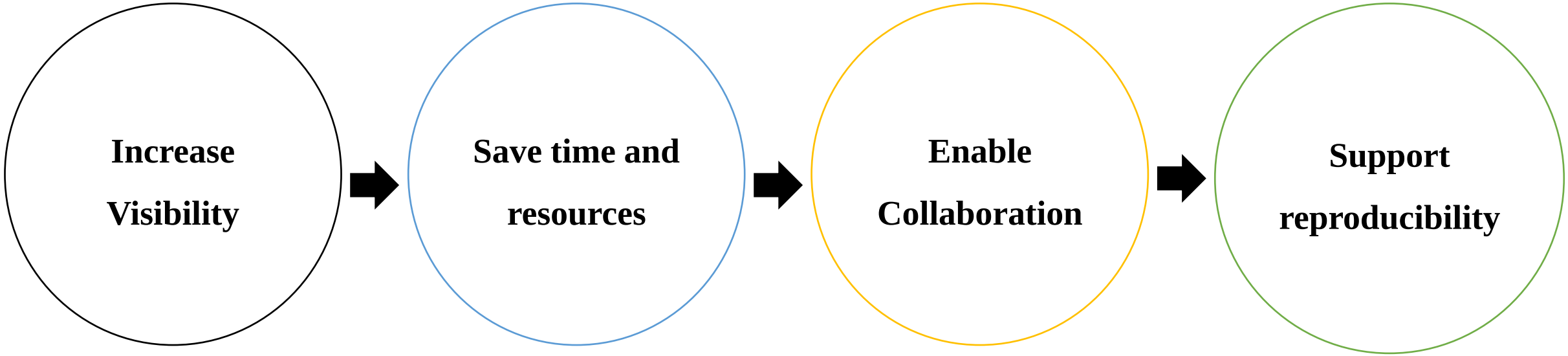
MESSY DATA

AFTER

CONNECTED, DISCOVERABLE DATA

Importance of FAIR System

FAIR Principles - Ensuring data is usable by humans and machines.



FAIR Principles

Principal 1: Findable

- Persistent Identifiers e.g. DOI
- Easily available metadata e.g. Zenodo

Principal 2: Accessible

- Standard protocols e.g. HTTP
- Metadata available

Principal 3: Interoperable

- Standard data format (CSV, JSON etc)
- Linking to related datasets & external repositories

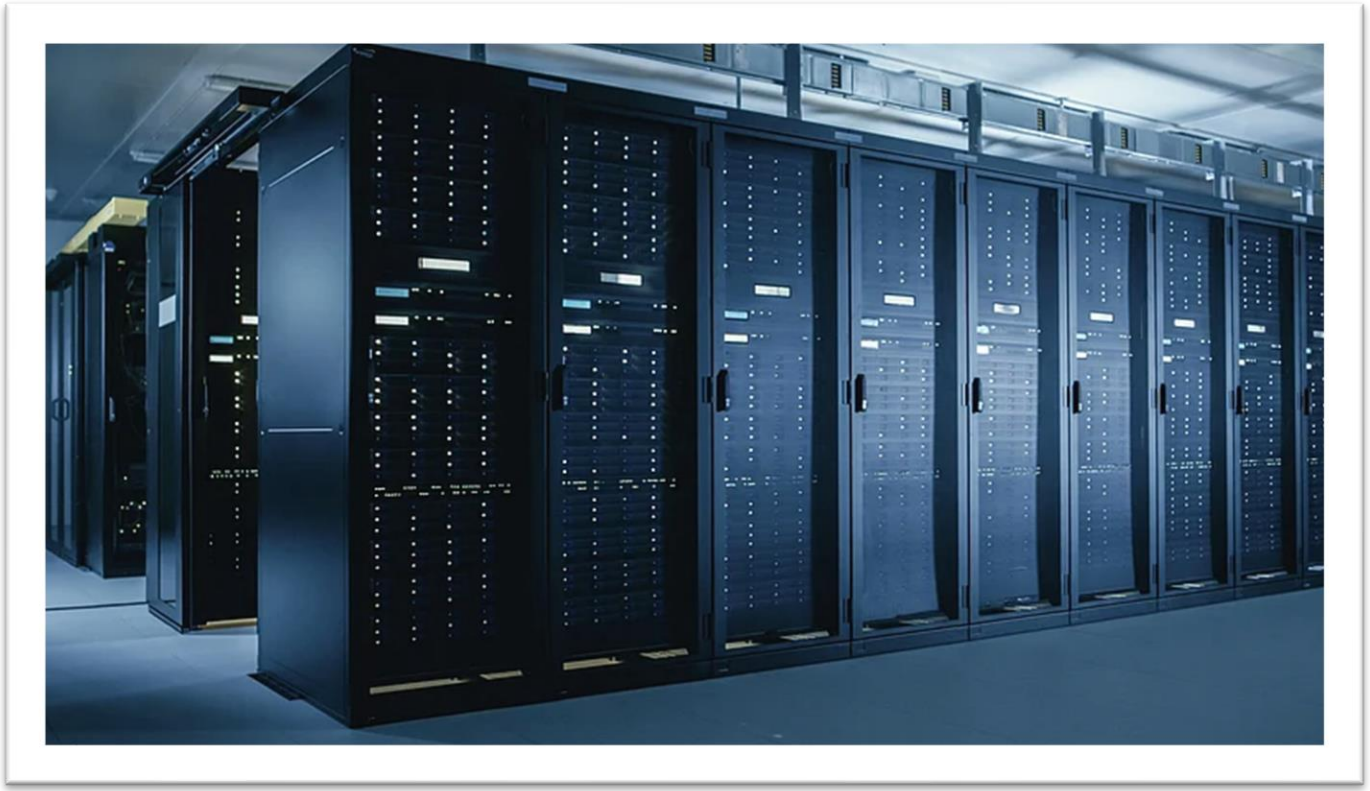
Principal 4: Accessible

- Clear licenses (CC-BY, CC0)
- Community-agreed metadata standards

What we Require for FAIR ?

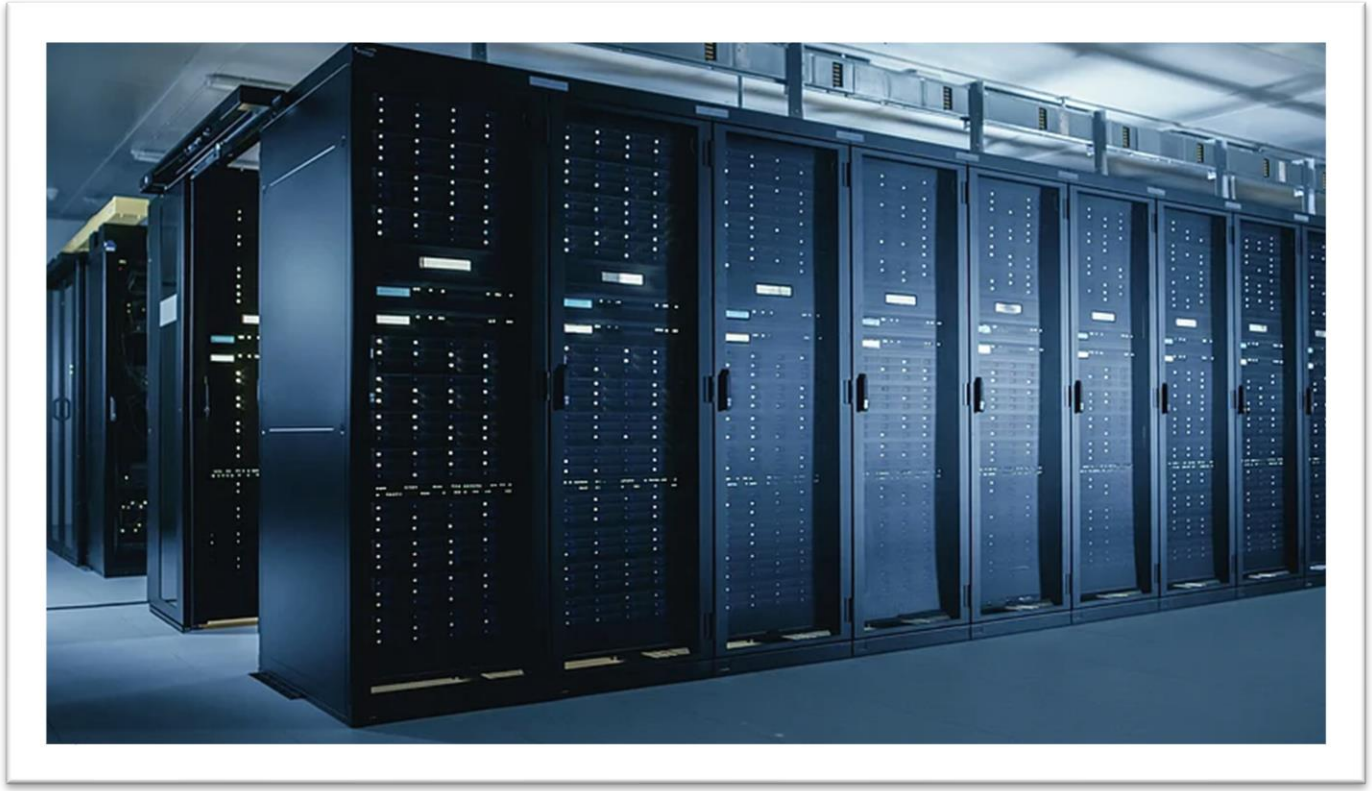
Repository

What is a Repository System?



A digital platform for storing, organizing, and sharing datasets, publications, software's and code

What is a Repository System?



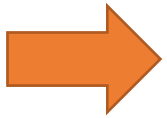
- ✓ **Centralized Access**
- ✓ **Long-term Preservation of Data**
- ✓ **Supports Open Science**

Dedicated Space will help to solve the data storage issue and Support the Open Science

What is required for Repository System?

Computational Resources

- **Processing Power: GPU/CPU**
- **Databases**
- **Machine Learning / NLP Tools**



Security & Access Control

- **Data stewards & curators for quality metadata.**
- **Training & Documentation for researchers and developers.**
- **Governance policies**

What is required for Repository System?

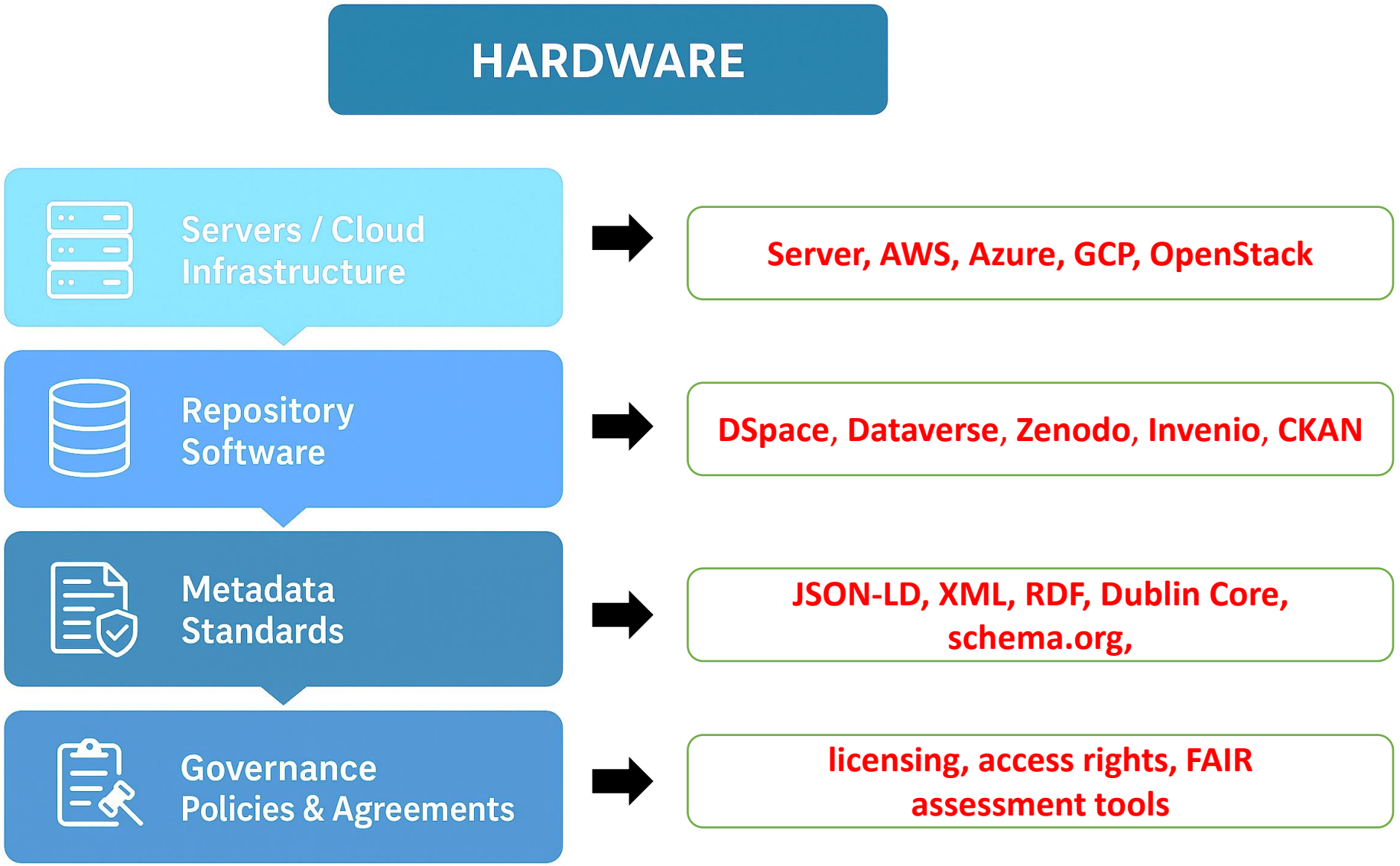
Hardware & Infrastructure

- Servers / Cloud Infrastructure
- High-performance storage
- Networking
- Backup & Redundancy

Software & Platforms

- Repository Software
- Persistent Identifier (PID) Systems
- Metadata Standards & Ontologies
- Interoperability Tools

What is required for Repository System?



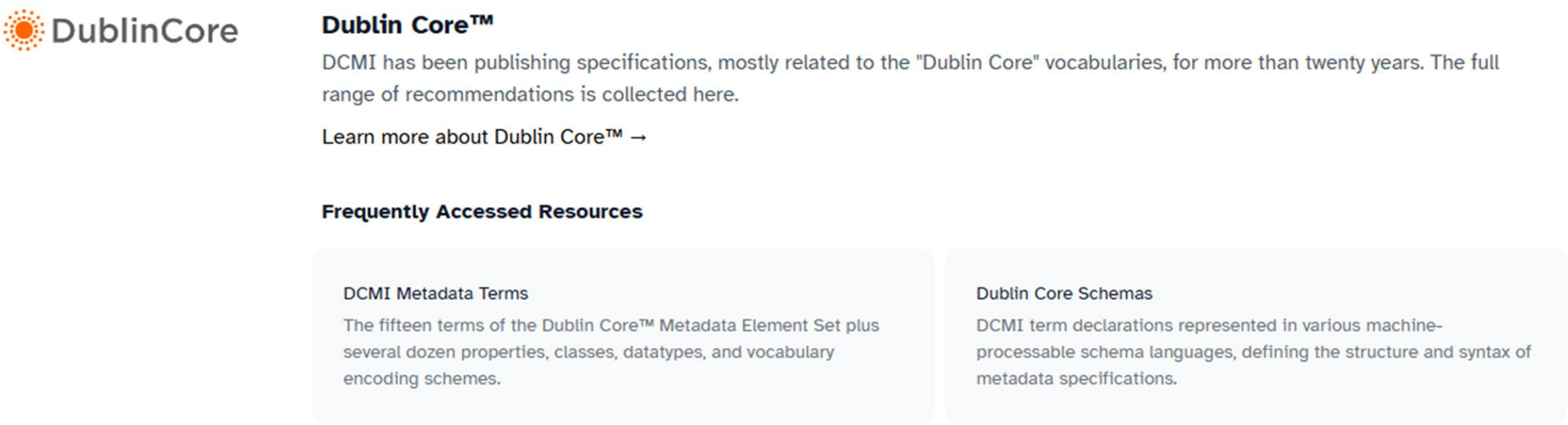
Components of a FAIR-Compliant Repository

- ✓ **Metadata Standards:**
- ✓ **Persistent Identifiers:**
- ✓ **Interoperability:**
- ✓ **Access Controls:**
- ✓ **Easy to search:**
- ✓ **Track update:**

Components of a FAIR-Compliant Repository

- **Metadata Standards:**

  Clear description of data (author, date, keywords, format).



Components of a FAIR-Compliant Repository

- **Metadata Standards: Examples**

Clear description of data (author, date, keywords, format).

nature › scientific data › articles › article

Article | Open access | Published: 22 May 2024

From Planning Stage Towards FAIR Data: A Practical Metadatasheet For Biomedical Scientists

Lea Seep, Stephan Grein, Iva Splichalova, Danli Ran, Mickel Mikhael, Staffan Hildebrand, Mario Lauterbach, Karsten Hiller, Dalila Juliana Silva Ribeiro, Katharina Sieckmann, Ronja Kardinal, Hao Huang, Jiangyan Yu, Sebastian Kallabis, Janina Behrens, Andreas Till, Viktoriya Peeva, Akim Strohmeyer, Johanna Bruder, Tobias Blum, Ana Soriano-Arroquia, Dominik Tischer, Katharina Kuellmer, Yuanfang Li, ... Jan Hasenauer

+ Show authors

Scientific Data **11**, Article number: 524 (2024) | Cite this article

5162 Accesses | **5** Citations | **3** Altmetric | Metrics

Components of a FAIR-Compliant Repository

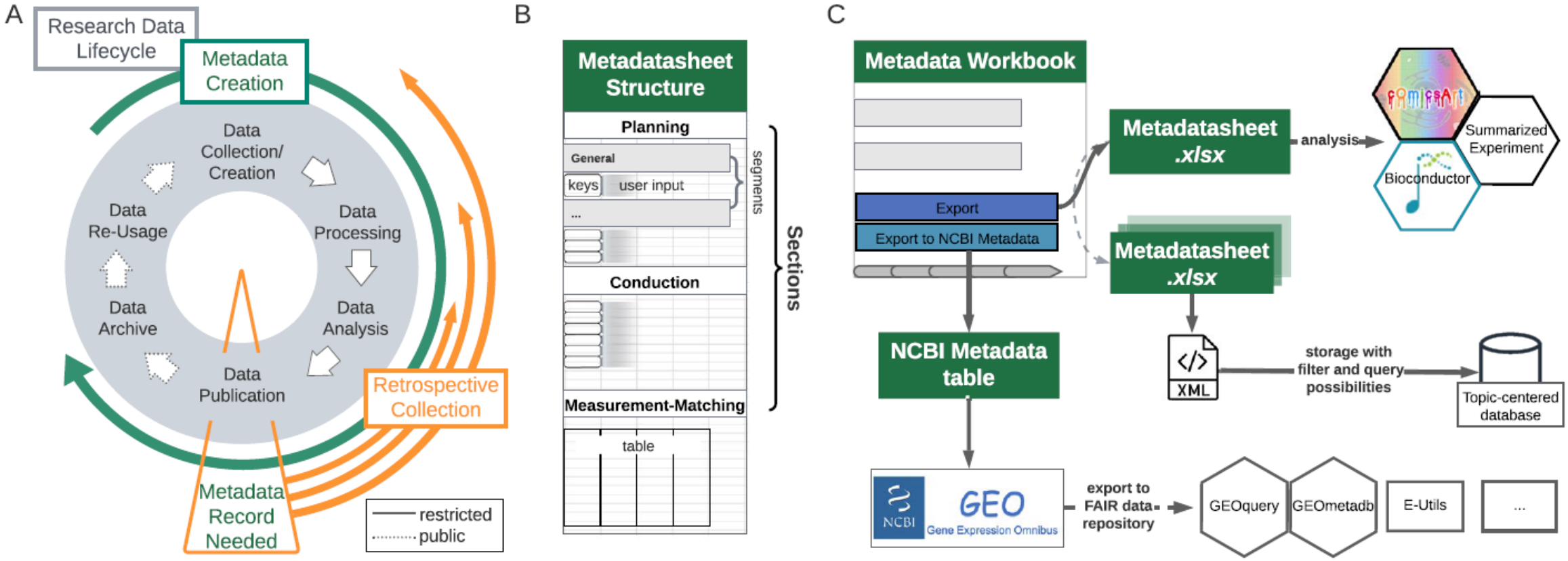


Fig. 1 Alignment of Metadata Lifecycle with the Research Data-Lifecycle.

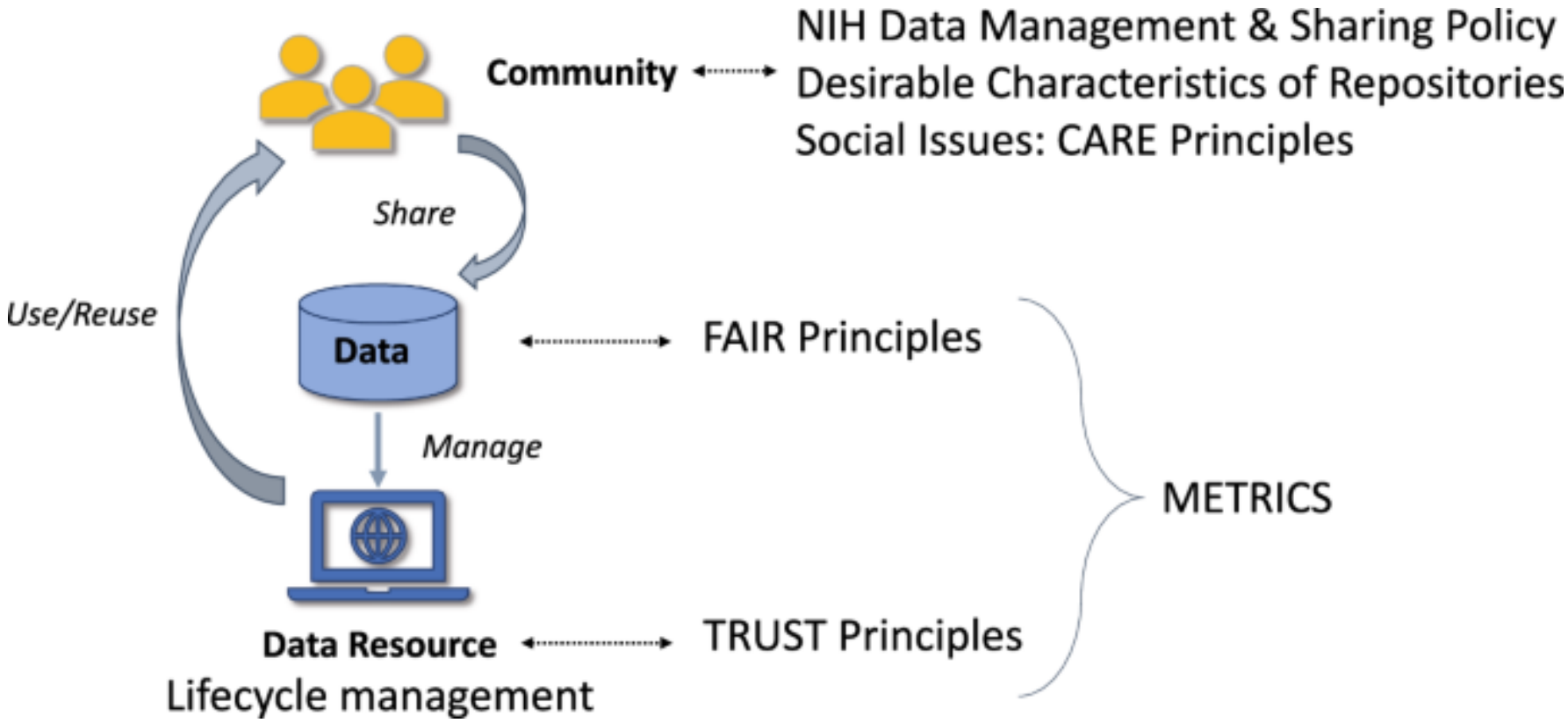
Components of a FAIR-Compliant Repository

Article | Open access | Published: 13 June 2024

Biomedical Data Repository Concepts and Management Principles

Dawei Lin, Matthew McAuliffe, Kim D. Pruitt, Anupama Gururaj, Christine Melchior, Charles Schmitt & Susan N. Wright

Scientific Data **11**, Article number: 622 (2024) | Cite this article



Components of a FAIR-Compliant Repository

Guideline	Based on[1]
F1. A workflow is assigned a globally unique and persistent identifier.	D-F1 and S-F1
F1.1. Components of the workflow representing levels of granularity are assigned distinct identifiers.	S-F1.1
F1.2. Different versions of the workflow are assigned distinct identifiers.	S-F1.2
F2. A workflow and its components are described with rich metadata.	D-F2 and S-F2
F3. Metadata clearly and explicitly include the identifier of the workflow, and workflow versions, that they describe.	D-F3 and S-F3
F4. Metadata and workflow are registered or indexed in a searchable FAIR resource.	D-F4 and S-F4
A1. Workflow and its components are retrievable by their identifiers using a standardized communications protocol.	D-A1 and S-A1
A1.1. The protocol is open, free, and universally implementable.	D-A1.1 and S-A1.1
A1.2. The protocol allows for an authentication and authorization procedure, when necessary.	D-A1.2 and S-A1.2
A2. Metadata are accessible, even when the workflow is no longer available.	D-A2 and S-A2
I1. Workflow and its metadata (including workflow run provenance) use a formal, accessible, shared, transparent, and broadly applicable language for knowledge representation.	D-I1 and S-R1.2
I2. Metadata and workflow use vocabularies that follow FAIR principles.	D-I2
I3. Workflow is specified in a way that allows its components to read, write, and exchange data (including intermediate data), in a way that meets domain-relevant standards.	D-R3 and S-I1
I4. Workflow and its metadata (including workflow run provenance) include qualified references to other objects and the workflow's components.	D-I3, S-I2, and S-R1.2
R1. Workflow is described with a plurality of accurate and relevant attributes.	D-R1 and S-R1
R1.1. Workflow is released with a clear and accessible license.	D-R1.1 and S-R1.1
R1.2. Components of the workflow representing levels of granularity are given clear and accessible licenses.	D-R1.1 and S-R1.1
R1.3. Workflow is associated with detailed provenance of the workflow and of the products of the workflow.	D-R1.2 and S-R1.2
R2. Workflow includes qualified references to other workflows.	D-I3 and S-R2
R3. Workflow meets domain-relevant community standards.	D-R1.3 and S-R3

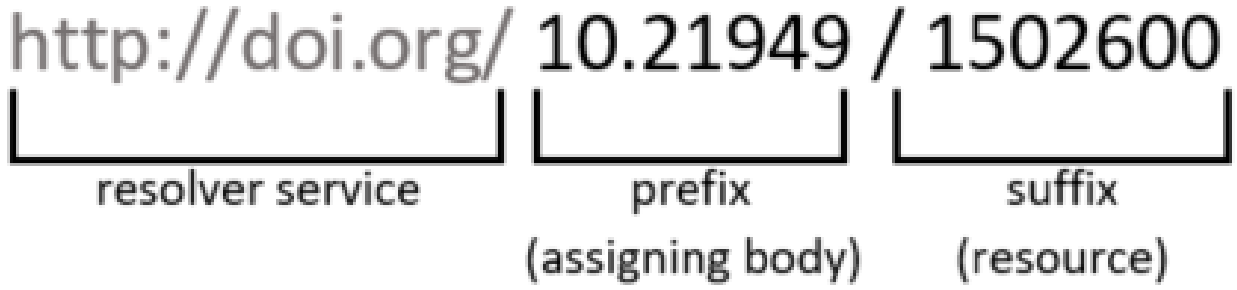
Applying the FAIR Principles to computational workflows

Wilkinson et. al., *Sci Data* 12, 328 (2025).

Components of a FAIR-Compliant Repository

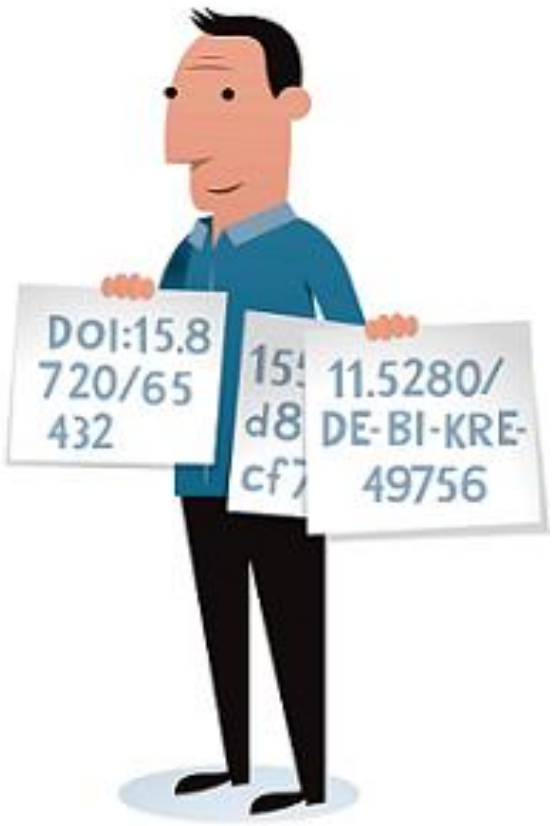
- **Persistent Identifiers:**

Digital Object Identifiers (DOI) for datasets, to ensure they remain findable and citable over time, even if the object's location changes



Components of a FAIR-Compliant Repository

- **Persistent Identifiers:**



✓ Handle

✓ **Archival Resource Key (ARK)**

✓ **Persistent Uniform Resource Locator (PURL)**

✓ **Universal Resource Name (URN)**

Components of a FAIR-Compliant Repository

- **Interoperability:**

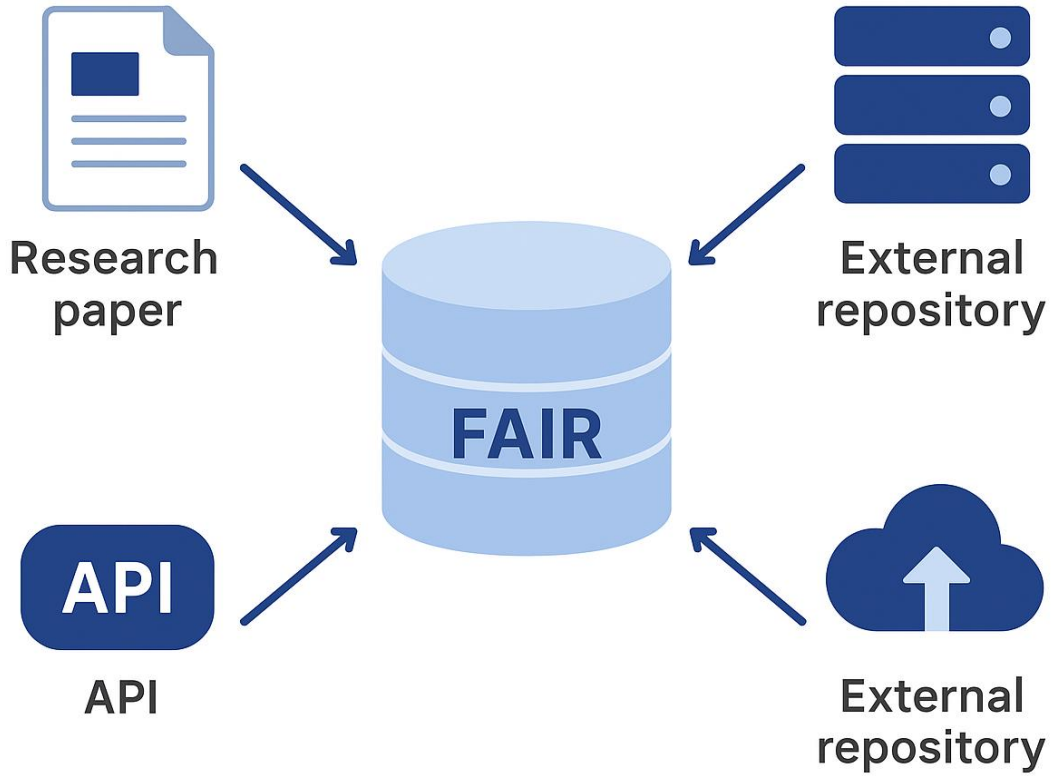
Support for standard file formats (CSV, JSON, XML, etc.).

Components of a FAIR-Compliant Repository

Data Type	Preferred Formats	Why Use It
Tabular data	CSV (comma-separated), TSV	Simple, universal, human- & machine-readable
Structured data	JSON, JSON-LD	Lightweight, API-friendly, supports nested structures
Hierarchical data	XML	Good for schemas (XSD) and metadata exchange
Metadata	Dublin Core, Schema.org, DataCite XML/JSON	Ensures semantic interoperability
Images	PNG, TIFF (non-lossy)	Open, widely supported
Documents	PDF/A, TXT, Markdown	For archival and accessibility

Components of a FAIR-Compliant Repository

- **Interoperability:** Linking to related datasets & external repositories



Components of a FAIR-Compliant Repository

- **Interoperability:** Linking to related datasets & external repositories
  - Metadata Linking: DataCite Metadata Schema
  - Cross-Repository Interlinking: NCBI, Zenodo, Figshare, Dryad, PDB, etc.

    GenBank IDs, PDB IDs
  - Human-Readable Linking:
  - Versioning & Provenance Tracking:
  - Compliance with Standards: ORCHID ID

Components of a FAIR-Compliant Repository

- **Access Controls:** FAIR does not mean "free for all" - it means as open as possible, as closed as necessary.

 **Lock icon for controlled data**

 **Globe icon for open data**

FAIR: As open as Possible, as closed as necessary

Components of a FAIR-Compliant Repository

- **Access Controls:**
- **Open Data:**
  - No privacy, security, or IP issues exist, Supports transparency & reproducibility
- **Access Controls:**
  - For sensitive data, with authentication & authorization
  - Important: Metadata should always remain visible, even if data access is restricted

Components of a FAIR-Compliant Repository

- **Easy to search: Findable = Human-friendly + Machine-friendly search.**

- **Rich Metadata:** title, author, keywords, description, DOI
- **Searchable Indexing:** Full-text search + faceted filters
- **Machine-Readable:** Metadata formats like JSON-LD, RDF

Steps to Design the System

- Requirement Analysis: Who will use it? What kind of data?

- Architecture Planning: Choose repository software (DSpace, Dataverse, Zenodo).

Steps to Design the System

Type

Open-source software (self-hosted or cloud-hosted)

Cost

Free to download and use

Best for

Institutional repositories (universities, research institutes)

Features

- ✔ Handles large collections (theses, papers, data)
- ✔ Supports DOIs, handles, metadata standards
- ✔ Customizable interface
- ✘ Needs technical expertise to install & maintain

Steps to Design the System

- Dataverse:
  - **Easy to manage –**
  - **Easy to secure –**
  - **Access your Dynamics 365 Data -**
  - **Rich metadata –**
  - **Productivity tools –**

For more details: https://learn.microsoft.com/en-us/power-apps/maker/data-platform/data-platform-intro

Steps to Design the System

Zenodo

Cloud-hosted repository run by CERN

Cost:

Completely free
(no hosting required)

Best for:

- ✔ Individual researchers or small groups

Features:

- ✔ Simple, no setup needed — just register and upload
- ✔ DOI assigned automatically for every upload
- ✘ Supports datasets, software, papers
- ✘ Limited customization & branding (shared platform)
- ✘ File size limit (50GB per upload, request needed for more)

Highly recommended for the storing the data for the publication purpose.

Steps to Design the System

Zenodo

Cloud-hosted repository run by CERN

Cost:

Completely free
(no hosting required)

Best for:

- ✔ Individual researchers or small groups

Features:

- ✔ Simple, no setup needed — just register and upload
- ✔ DOI assigned automatically for every upload
- ✘ Supports datasets, software, papers
- ✘ Limited customization & branding (shared platform)
- ✘ File size limit (50GB per upload, request needed for more)

Highly recommended for the storing the data for the publication purpose.

Zenodo for to store and share a metadata

Article | Open access | Published: 10 January 2023

Resolving colistin resistance and heteroresistance in *Enterobacter* species

Swapnil Prakash Doijad, Nicolas Gisch, Renate Frantz, Bajarang Vasant Kumbhar, Jane Falgenhauer, Can Imirzalioglu, Linda Falgenhauer, Alexander Mischnik, Jan Rupp, Michael Behnke, Michael Buhl, Simone Eisenbeis, Petra Gastmeier, Hanna Gölz, Georg Alexander Häcker, Nadja Käding, Winfried V. Kern, Axel Kola, Evelyn Kramme, Silke Peter, Anna M. Rohde, Harald Seifert, Evelina Tacconelli, Maria J. G. T. Vehreschild, DZIF R-Net Study Group & ... Trinad Chakraborty + Show authors

Nature Communications **14**, Article number: 140 (2023) | Cite this article

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The genome sequencing data generated in this study have been deposited in the NCBI database under the BioProject accession number PRJNA622426. Experimental data for antibiotic resistance phenotype, minimum inhibitory concentration (MIC) measurements and mass spectrometry are provided in Supplementary Material, Supplementary, or Source data files. Fluorescence microscopy raw images are submitted as a dataset on Zenodo with DOI: 10.5281/zenodo.7405690. All data are available without any restriction. Source data are provided with this paper.

Challenges & Solutions

- **Challenge:** Lack of discoverability
  - **Solution: Good metadata + search engine optimization.**
- **Challenge:** Data format mismatch
  - **Solution: Use standard interoperable formats.**
- **Challenge:** Access restrictions

Solution: Controlled but documented access policies.

Best Examples of FAIR repositories

- **Zenodo** (general-purpose research repository).
- **NCBI GeneBank** (biological data). www.ncbi.nih.nil.gov
- **PDB** (Protein Data Bank). www.rcsb.org

Benefits & Future Scope

- Promotes open science and innovation.
- Easier AI/ML integration due to structured data.

Transparency Makes research methods, data, and results openly available, increasing trust.

Collaboration Enables global sharing and collective problem-solving across borders and disciplines

Innovation Open, FAIR data accelerates discovery by letting others test, validate, or build upon work

Equity Democratizes access to knowledge, especially for researchers in resource-limited settings

Long-term Preservation FAIR data ensures knowledge remains usable even after projects end

Conclusion

- “Let’s make data more findable, accessible, interoperable, and reusable!”
- Open FAIR repository need to develop
- FAIR is the “How” **(Practical steps to manage data)** and Open Science is the “Why” **(bigger goal of sharing knowledge fairly)**.


Terima-kasih
Kyaayyjuutainpartaal
dhanyavaad
Dhan'yabāda
merci
shukraan-lak
efcharistó
Ďakujem
gracias
Děkuji
Ačiū
D'akuju
고맙습니다
Köszönöm
Obrigado
Grazzi
Hvala-vam
Shukria
spasibo
Tänan
ကျေးဇူးတင်ပါတယ်
ขอบคุณค่ะ
Дзякуй
Paldies
Gamsahamnida
Arigatou
Hvala-ti
धन्यवाद
Kiitos
tak
多謝
Teşekkür-ederim
thank-you
Tack
谢谢
Dank-je
Danke
ευχαριστώ
Dzaikuj
motshakeram
asante
Дякую
Khob-khun-ka
Спасибо
Blagodarya-ti
Aitäh
Khop-chai
doh-je
Faleminderit
dhanwaad
ຂອບໃຈ
cảm-ơn
Благодаря-ти
Salamat
Dziękuję
شکریہ
Dankie
Xiè-xiè
Grazie
Mulțumesc
Хвала-вам
ありがとうございます
takk
धन्यवाद