

Introduction

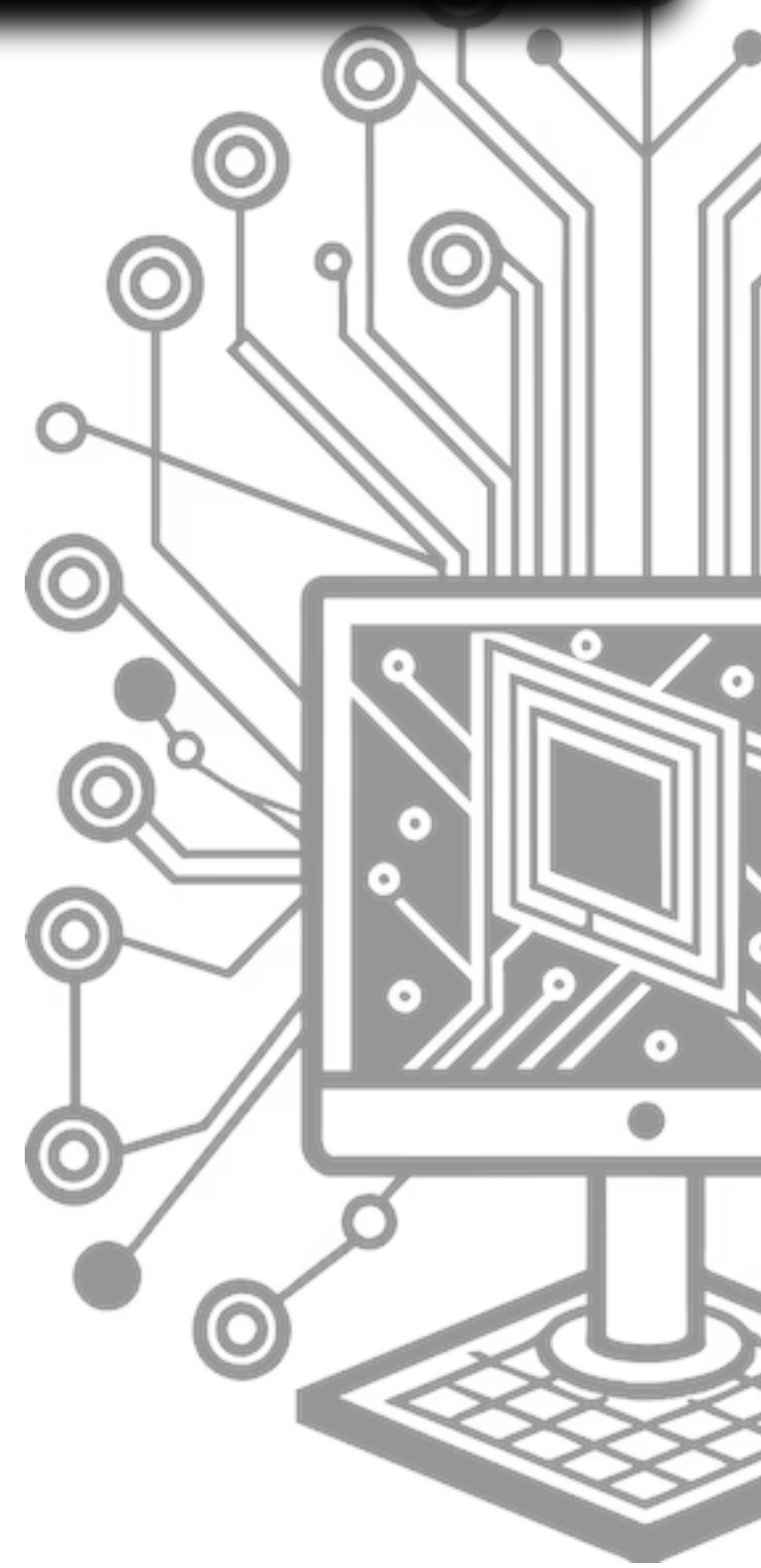


Hands-on Workshop

July 8, 2025

Credits for material : [\[H. Simma 2023\]](#) [\[D. Pleiter 2025\]](#)

... an insider user's perspective



Motivation

- ❑ Gauge configurations from Lattice simulations are **valuable** and **expensive** in terms of

- human effort
- computing resources (i.e. tax payer's money, energy, CO₂)

Example: 192×96^3 lattice, estimates for **1 config**

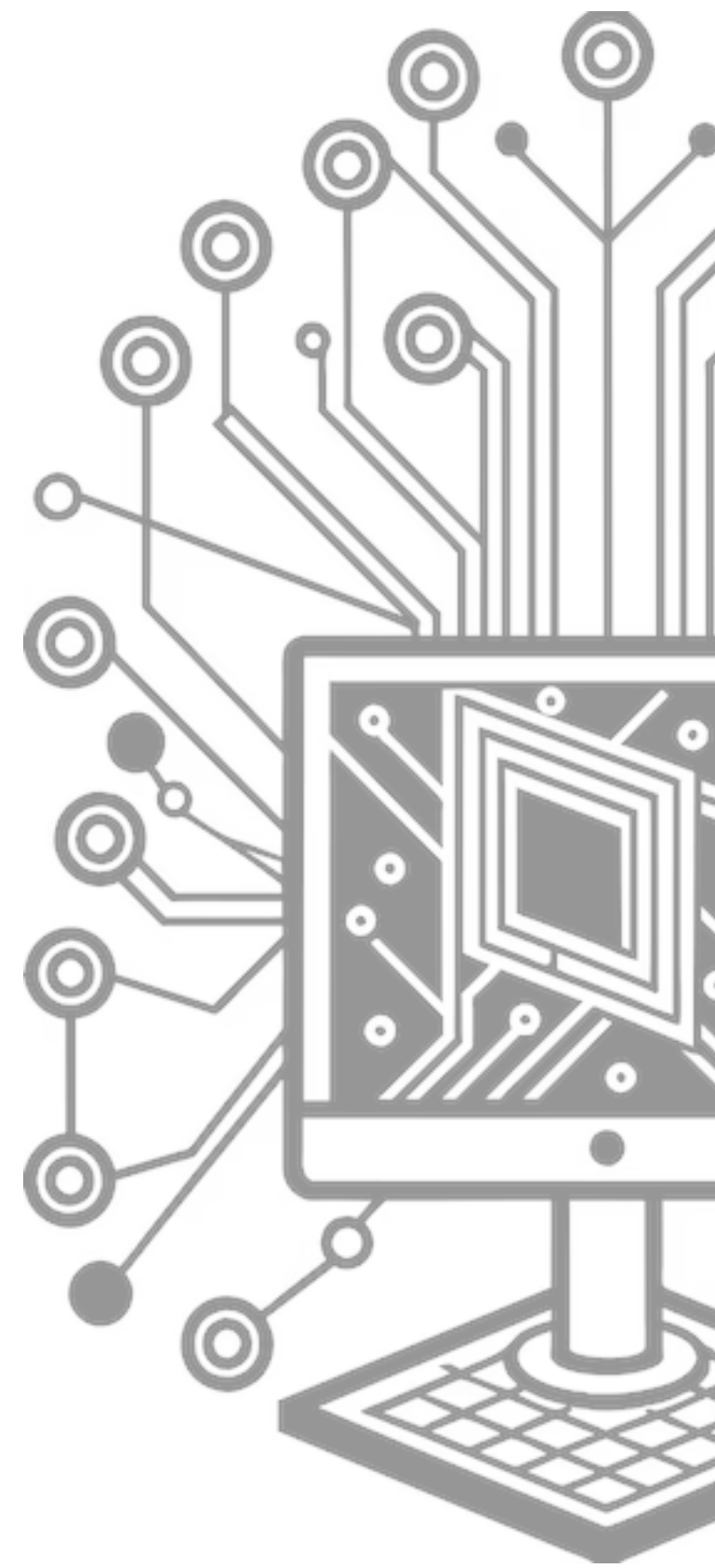
CPU	78 kch	(8192 cores × 9.5 h)	
nominal cost	780 €	(1 cent/ch)	
energy	780 kWh	(10 W/core)	
CO ₂	550 kg	(0.7 kg/kW)	[epa.gov]

- ❑ Proper data management is an important aspect of **good scientific practice**

- FAIR principles
- Data Management Plan

- ❑ Making the data **useful** — for us and others — does **not** come for free!

data size	97 GB		
1 data transfer	7.8 kWh	(0.08 kWh/GB)	[iea.org]
10 y repository	320 €	(3500 \$/TB)	[figshare.com]
10 y tape	70 €	(70 \$/TB)	[fujifilm.com]



International Lattice Data Grid (ILDG)

Community effort to share expensive primary data:

- proposed at Lattice conference [2002](#)
- community-wide agreed **metadata schema** [2004](#)
- first services operational $\approx$ 2007
- start of efforts to modernize ILDG [2022](#)

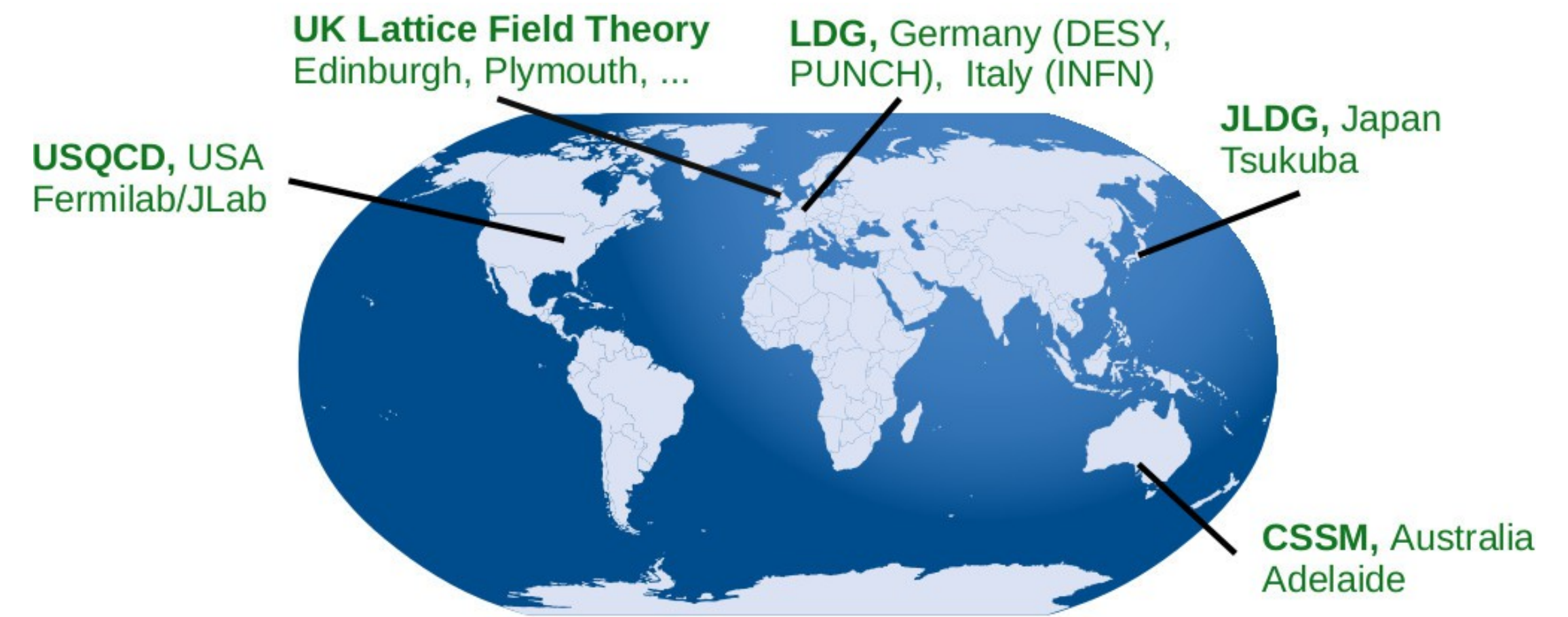
Organization:

- federation of autonomous “Regional Grids”
- forming a single [Virtual Organization \(VO\)](#)
- 2 Working Groups (metadata and middleware) + Board

Basic Concepts:

- ILDG defines standardized metadata schema, file-format, API
- Regional grids (with specific policies, technologies, resources, ...) provide catalogue services + storage

E.g. LDG in Europe makes to a large extent also use of WLCG technology and services



[\[hpc.desy.de/ildg\]](http://hpc.desy.de/ildg)

Virtual Organisations (VO)

What is it?

A set of individuals and/or institutions defined by sharing rules [\[I. Foster et al.; 2001\]](#)

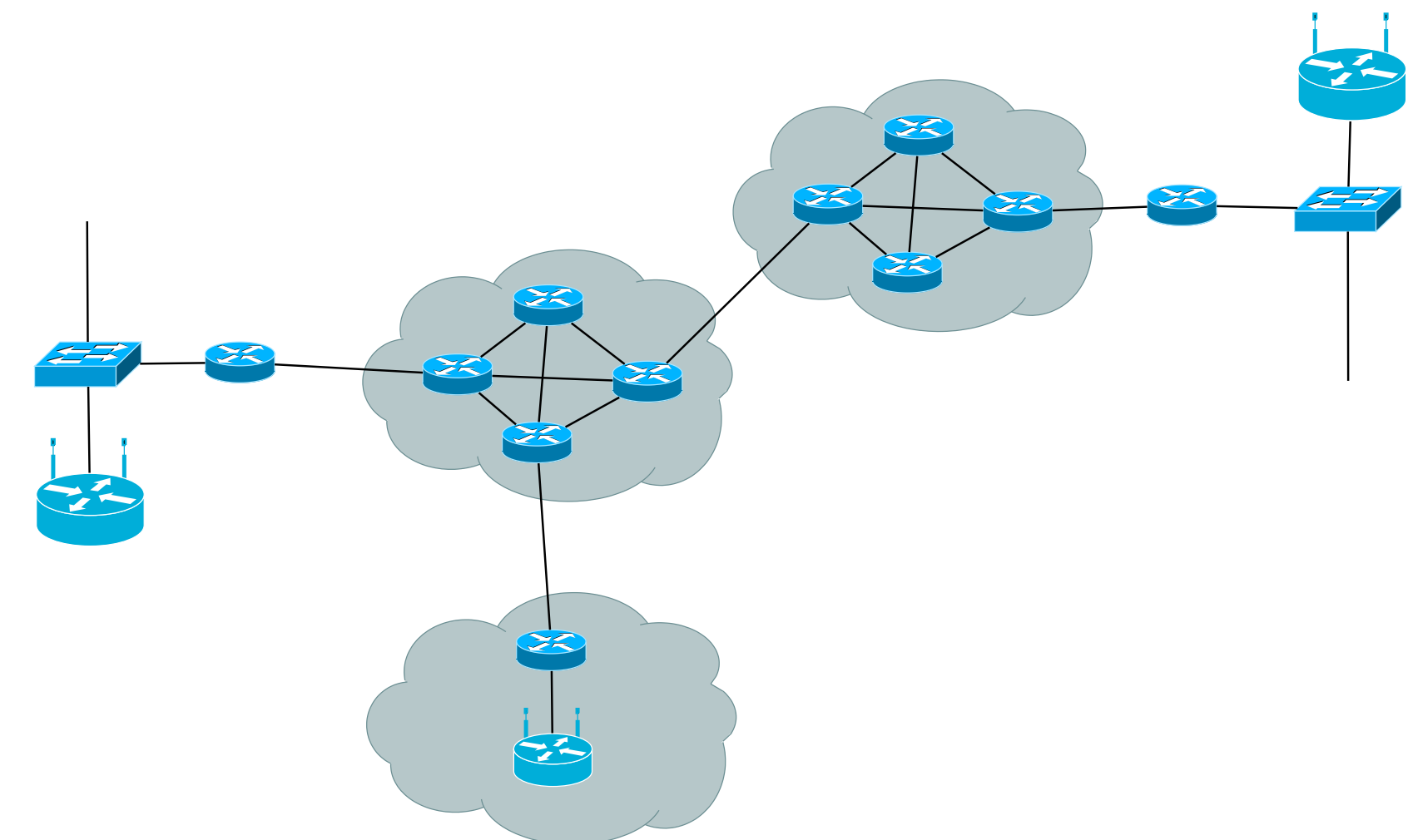
VOs vary tremendously in their purpose, scope, size, duration, structure, community, and sociology

What about us?

It is a model that has for long be successful for grid infrastructures

Generally:

- A (research) community organises itself as VO
- Digital infrastructures provides resources and services to the VO
- The VO manages the use of these resources and services



Organizational Structure of ILDG

❑ Metadata Working Group (MDWG)

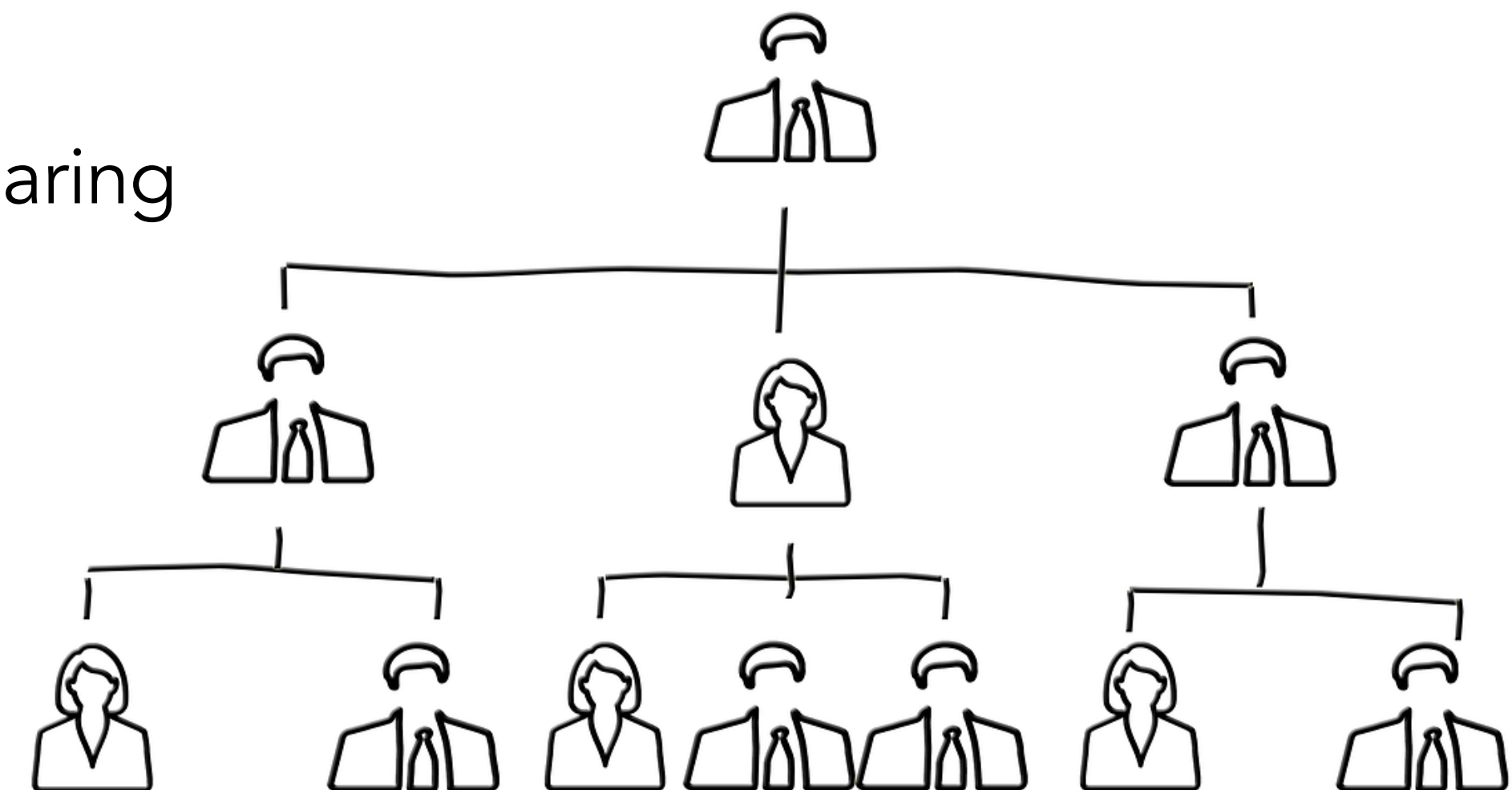
- Agrees on community-wide standards for the description of the data
- Specifies metadata schema (QCDml) and data formats

❑ Middleware Working Group (MWWG)

- Specifies interfaces of services to ensure interoperable regional grids
- Supports technical implementation of regional grids
- Suggests or develops prototypes of user tools

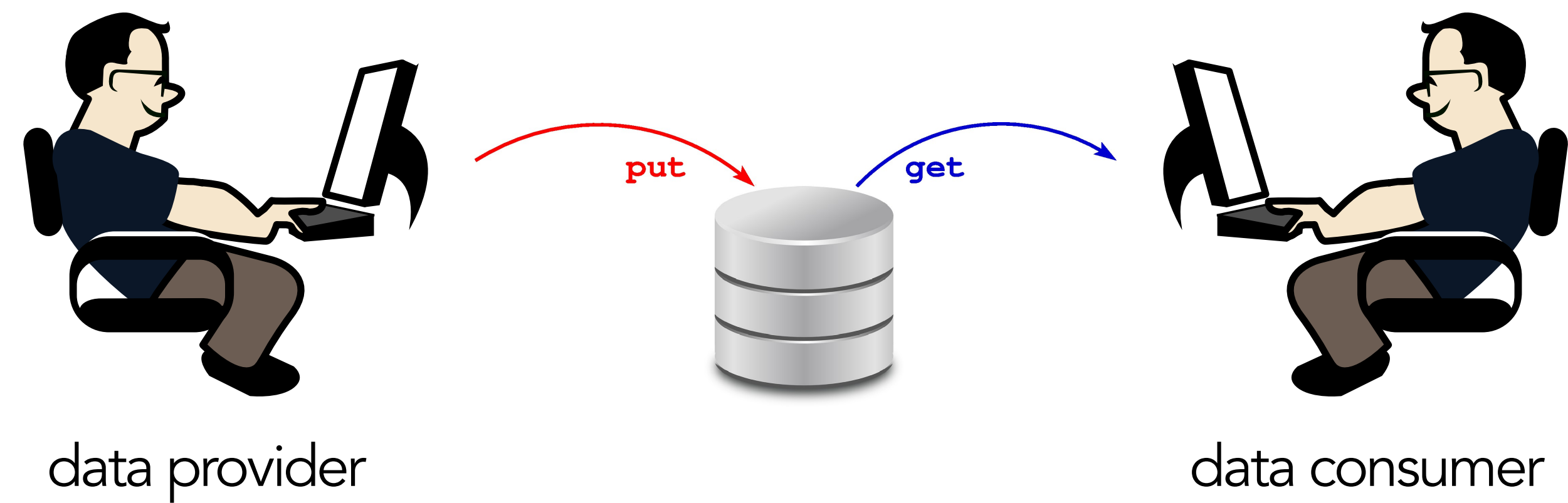
❑ Board

- Represents ILDG towards community and service providers
- Decides on policies and guidelines for membership and data sharing
- Supports regional grids in applying for resources
- Oversees working groups



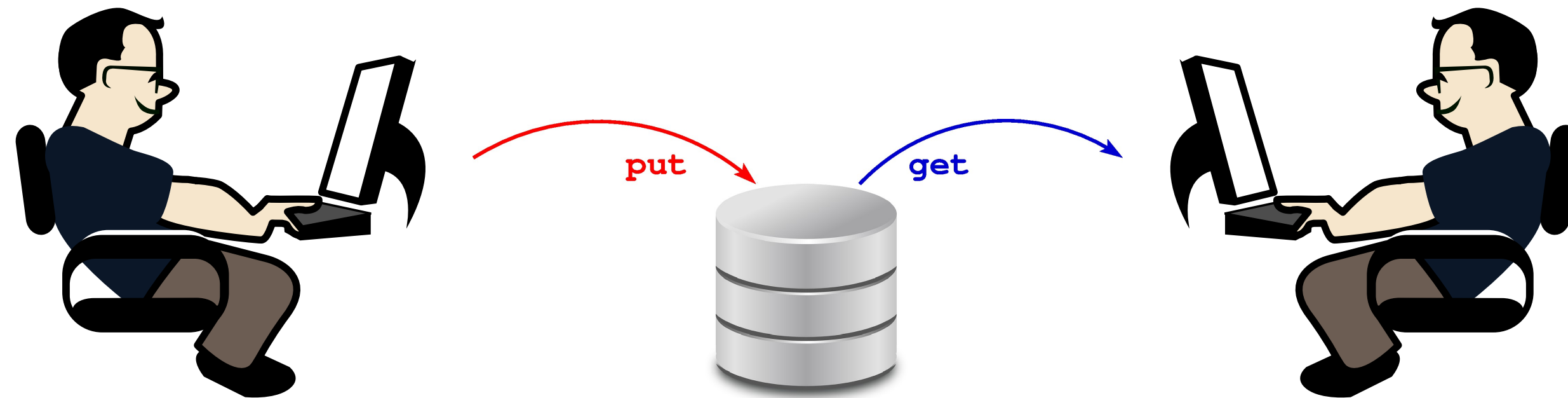
Use Cases

ILDG needs to support 4 different use cases (and user requirements) for sharing and exchanging of gauge configurations:



	data consumer	data provider
collaboration-internal sharing (initial "embargo" restrictions)	✓	✓
community-wide sharing ("public" or "published"?)	✓	✓

Benefits for Data Providers and Consumers



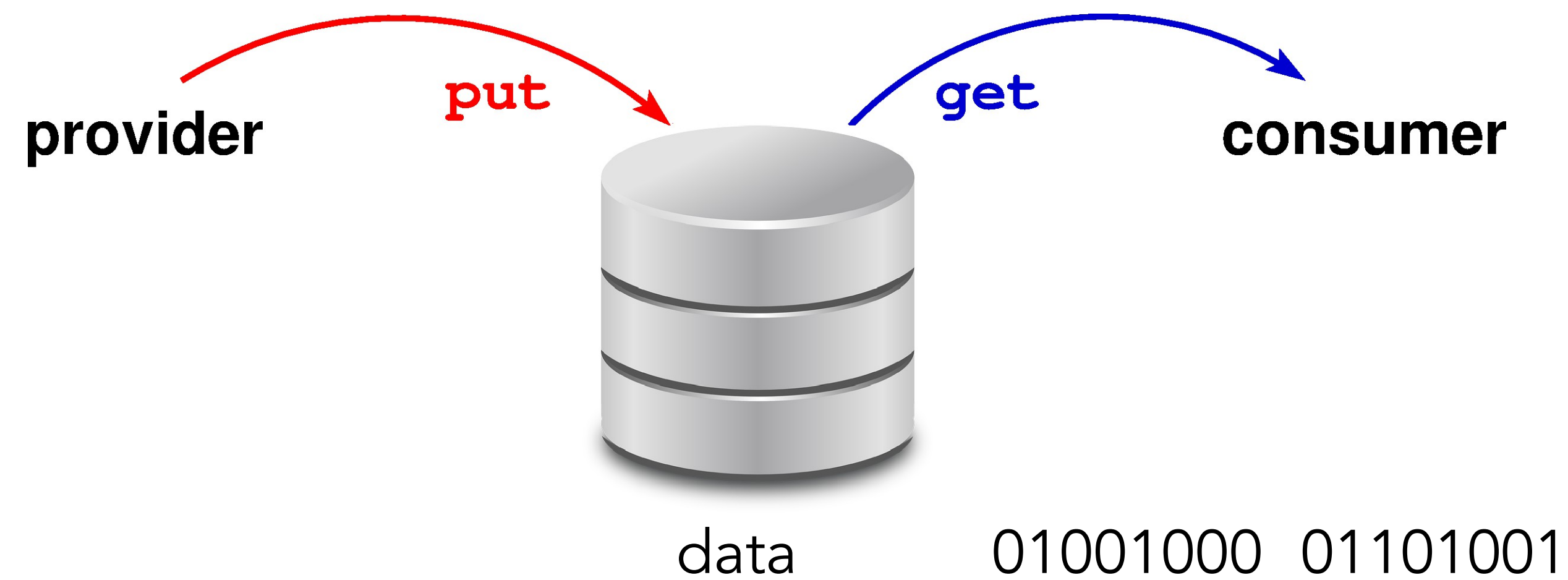
Data providers:

- ✓ have a clear data management plan
- ✓ can follow a well-defined workflow
- ✓ public and internal data can be handled in the same way (no extra efforts at end of embargo period)
- ✓ data is citable and may get published
- ✓ efforts are rewarded by funding agencies

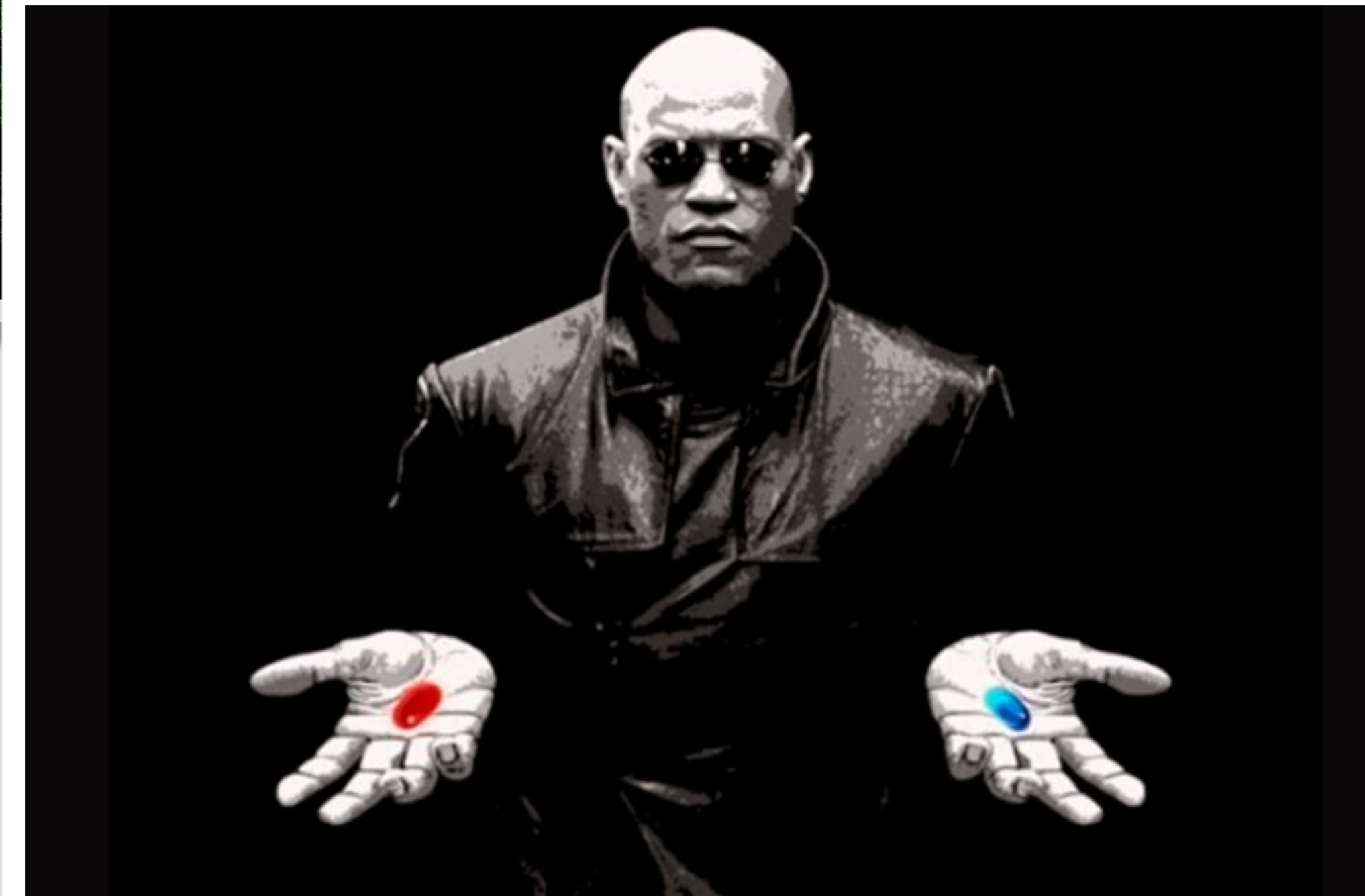
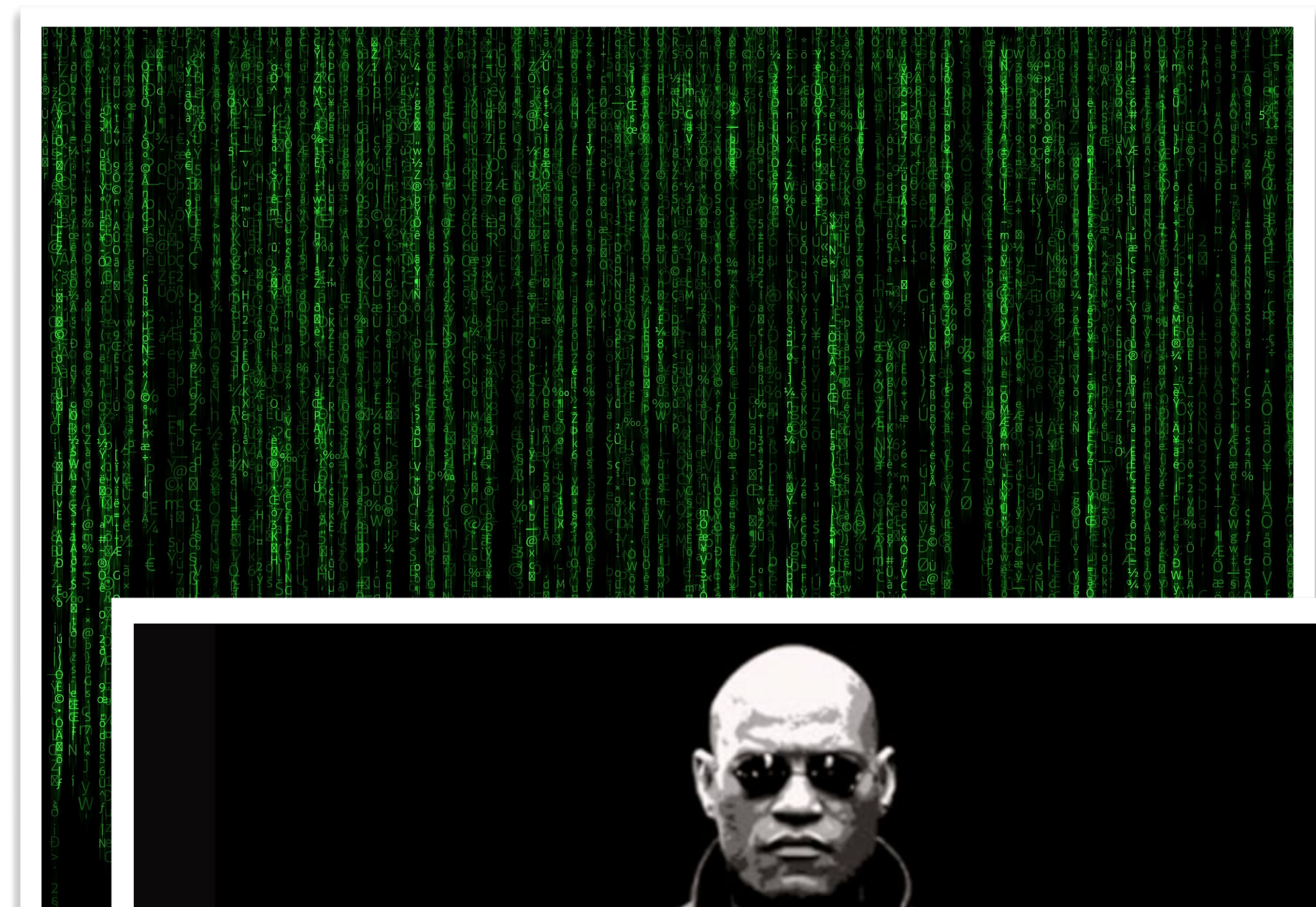
Data consumers:

- ✓ know about existence of interesting and useful data
- ✓ receive all relevant info on configs (location, tracking, reliability, ...)
- ✓ can do high quality research at less cost
- ✓ have well-defined and known usage rules

Naive Data Sharing

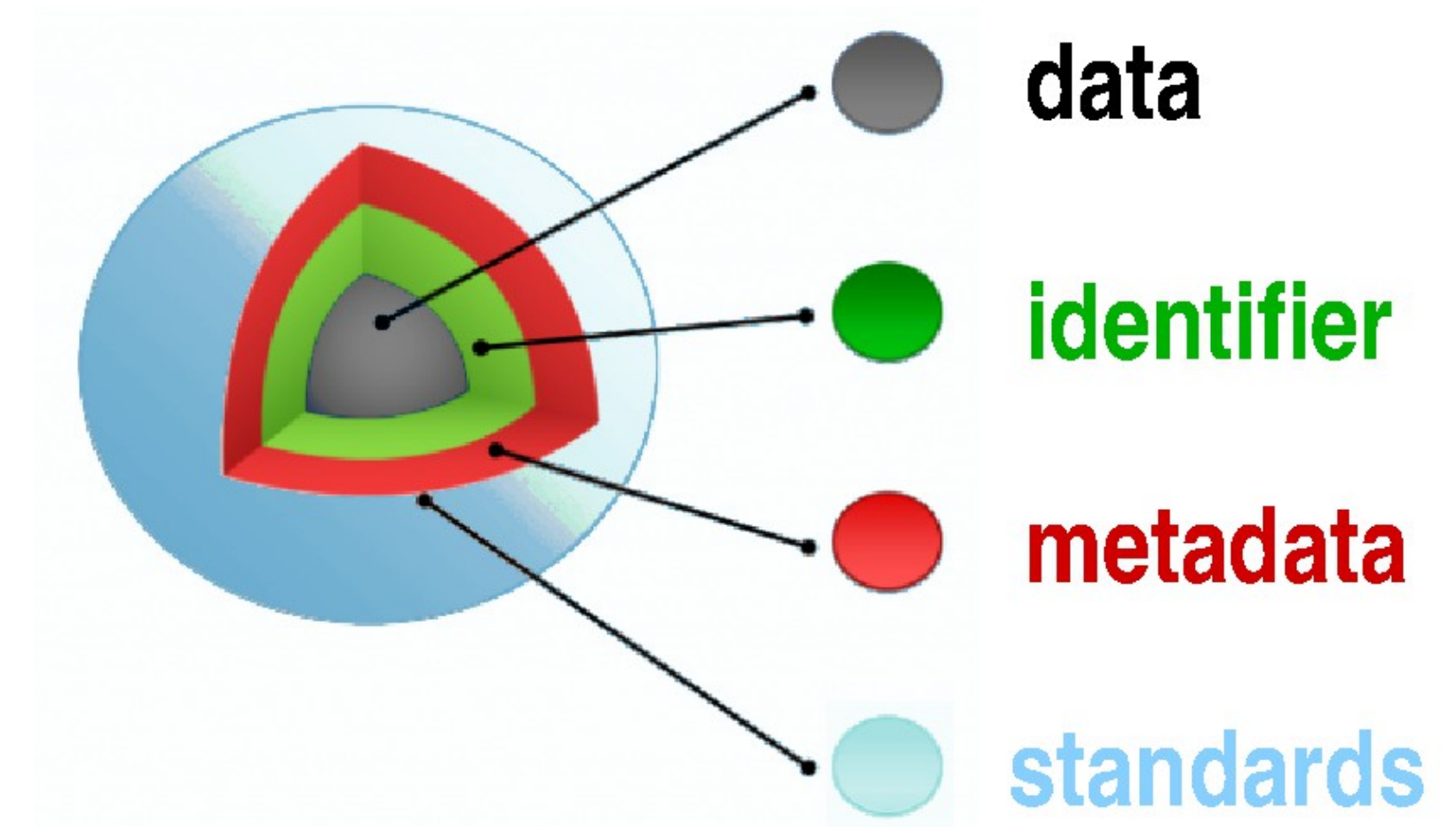


Data (bits) **without** meta data
(= information about a digital object) is **useless!**



FAIR principles for scientific data management and stewardship

Findable
Accessible
Interoperable
Reusable



References:

- The FAIR Guiding Principles for scientific data management and stewardship: [Wilkinson 2016](#)
- European Cloud Initiative - Building a competitive data and knowledge economy in Europe: [EU Commission 2016](#)
- GO FAIR is a bottom-up initiative that aims to implement the FAIR data principles: [go-fair.org](#)
- The Future of Research Communications and e-Scholarship: [force11.org](#)

FAIR principles for scientific data management and stewardship

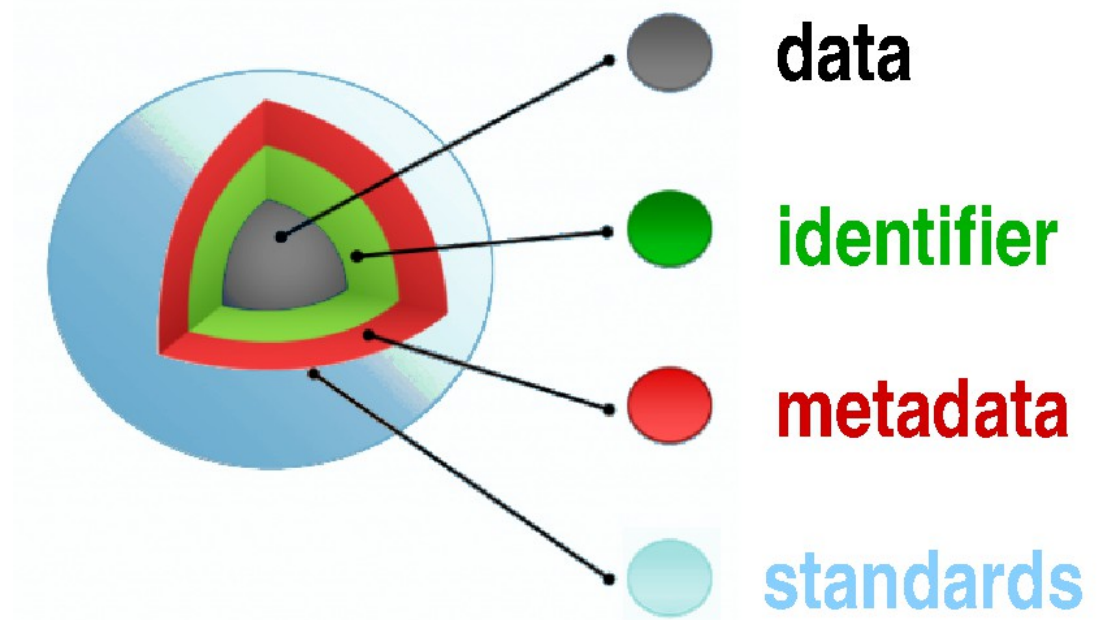
Findable

Accessible

Interoperable

Reusable

- ❑ required by funding agencies
- ❑ 15 concise principles formulated in [Wilkinson 2016](#)
41 detailed indicators in [FAIR Data Maturity Model](#)
- ❑ guiding principles (not implementation)



Findable

F1 (Meta)data are assigned a globally unique and persistent identifier

F2 Data are described with rich metadata

F3 Metadata clearly and explicitly include the identifier of the data they describe

F4 (Meta)data are registered or indexed in a searchable resource

Interoperable

I1 (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2 (Meta)data use vocabularies that follow FAIR principles

I3 (Meta)data include qualified references to other (meta)data

Accessible

A1 (Meta)data are retrievable by their identifier using a standardised communications protocol

A2 Metadata are accessible, even when the data are no longer available

Reusable

R1 (Meta-)data richly described with plurality of accurate and relevant attributes

R1.1 (Meta-)data released with clear and accessible data usage license

R1.2 (Meta-)data associated with detailed provenance

R1.3 (Meta-)data meet domain-relevant community standard

What does “findable” mean?

Findable

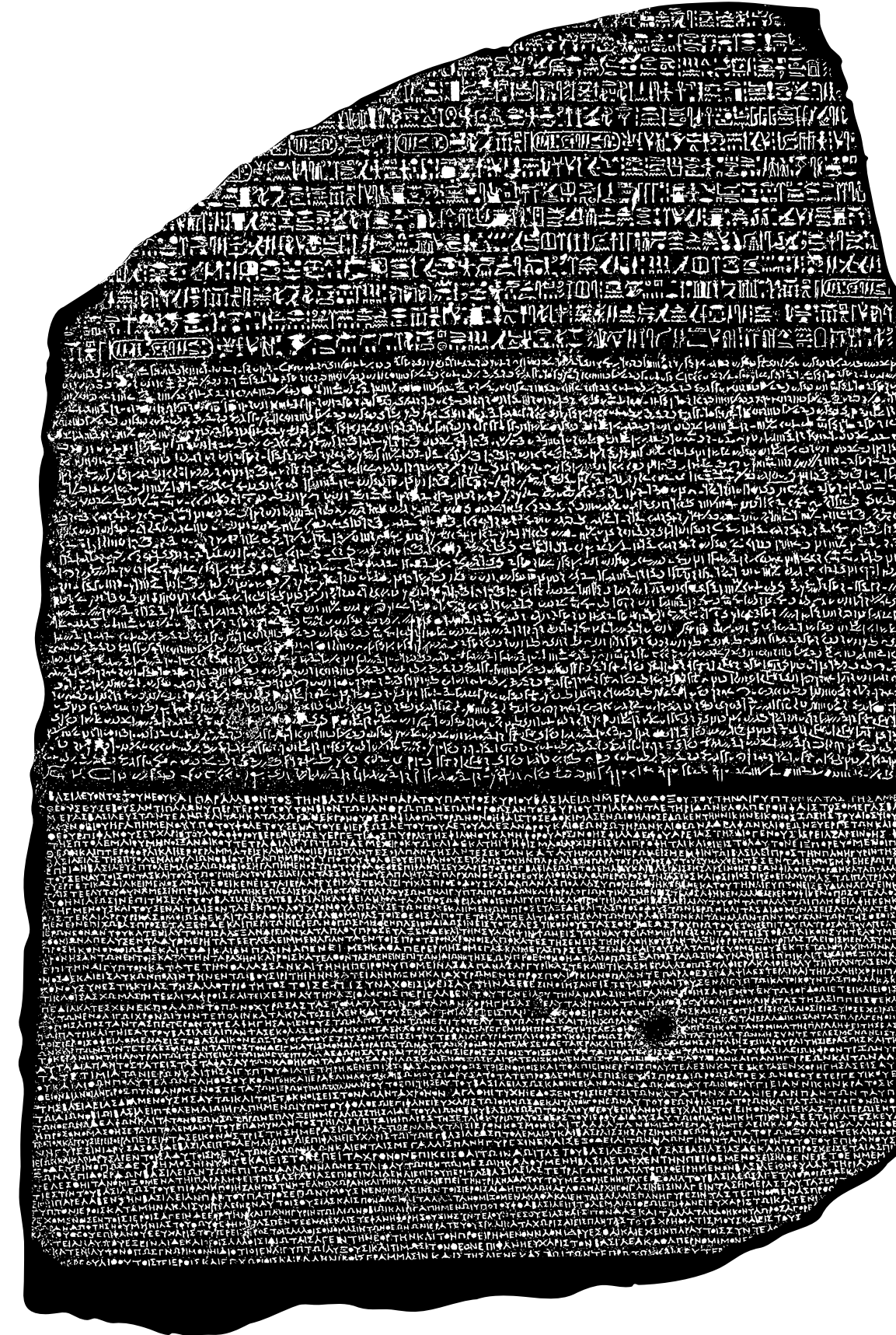
- F1 (Meta)data are assigned a globally unique and persistent identifier
- F2 Data are described with rich metadata
- F3 Metadata clearly and explicitly include the identifier of the data they describe
- F4 (Meta)data are registered or indexed in a searchable resource

- Persistent ID is essential concept (F1)
- Rich Metadata (MD) includes many kinds of information, e.g.
 - content (using a general and domain-specific vocabulary)
 - provenance (who, when, where, how?)
 - access (format, path, license, ...)
 - ...

```
-r w -r -r - - Apr 1 2025 09:45 README
```

```
-r w - - - - - Apr 1 2025 10:11 b56k137n4
```

- Registered MD can be searched and harvested (F4)



What does “accessible” mean?

Accessible

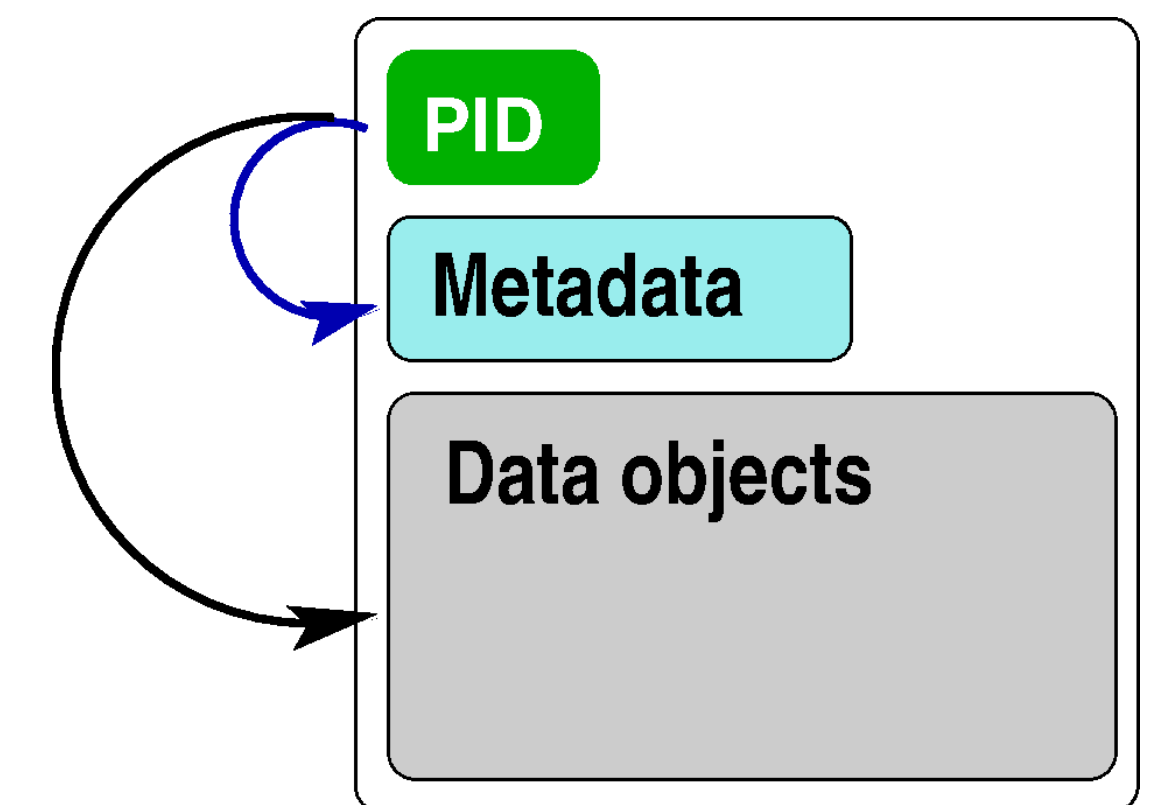
A1 (Meta)data are retrievable by their identifier using a standardised communications protocol

A1.1 protocol is open, free, and universally implementable

A1.2 protocol allows authentication/authorization procedure where necessary

A2 Metadata are accessible, even when the data are no longer available Interoperable

- A1 can be achieved e.g. by Metadata and File Catalogues
 - ID → metadata document(s)
 - ID → storage location(s)
- Accessible does not mandate public access without restrictions
- Metadata is precious even without the associated data

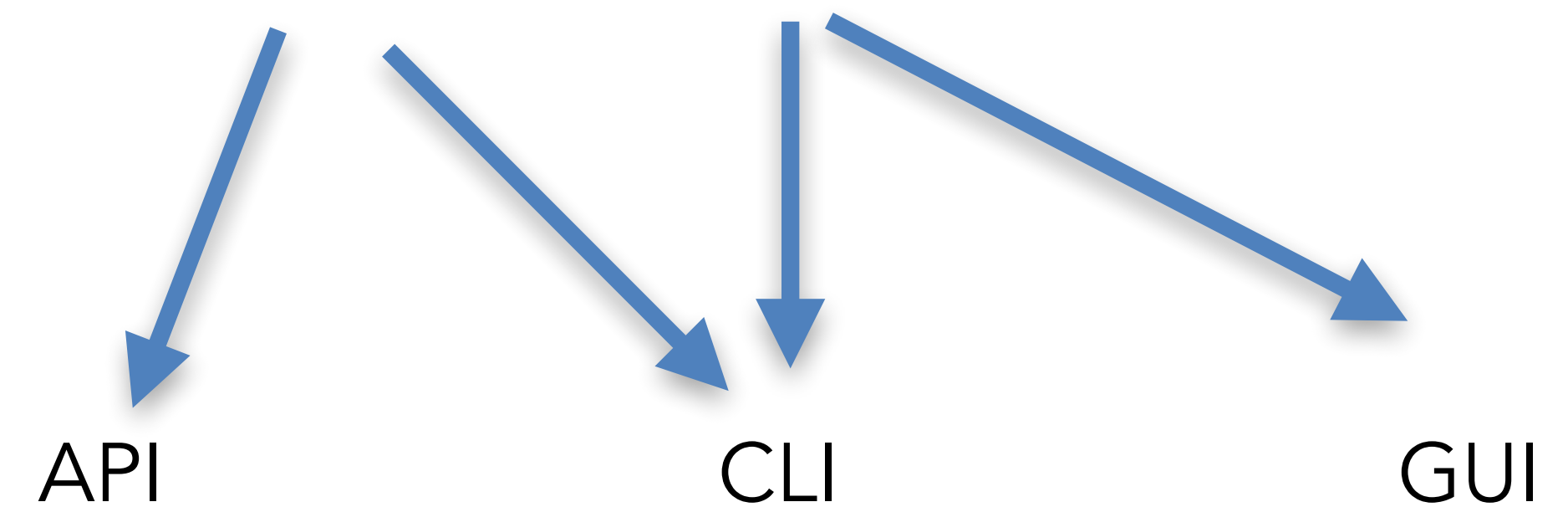


What does “interoperable” mean?

Interoperable

- I1 (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2 (Meta)data use vocabularies that follow FAIR principles
- I3 (Meta)data include qualified references to other (meta)data

- ability of data (or tools) from non-cooperating resources to integrate (or work together) with minimal effort
- reference to a paper may not be sufficient
- FAIR requires machine actionable (meta)data
- data must be FAIR for **machines** and **humans**



What does “reusable” mean?

Reusable

R1 (Meta-)data richly described with plurality of accurate and relevant attributes

R1.1 (Meta-)data released with clear and accessible data usage license

R1.2 (Meta-)data associated with detailed provenance

R1.3 (Meta-)data meet domain-relevant community standard



- Metadata in ILDG is always public (CC0)
- Data itself can use e.g. [Creative Commons](#) license:
 - CC BY credit must be given to the creator
(must own or control copyright in work, licenses can not be revoked)
 - Extensions **SA** and **NC** are **not recommended**
 - Extension **ND** (no derivatives or adaptations of the work are permitted) is **unsuitable**
- Proper data publishing is more than public data (DOI, landing page, MD harvesting, ...)

Aims and Features of ILDG 2.0

Done:

- ✓ revised and extended QCDml Metadata Schema
- ✓ new Identity and Access Management (grid certificate → token, support for embargo periods)
- ✓ re-factored Metadata and File Catalogues

In progress:

- ✓ further modernizations and extensions (HDF5, FTS, tools, GUI)
- ✓ data publishing (with registration of DOI)

ILDG . . .

- ! Is not intended as a free solution when running out of disk space
- ! Does not come for free. It requires effort and resources on users' side
- ! It is a work in progress

Plan of the Workshop

	Lectures / Exercises	tasks of participants
Today	Services and Middleware Intro to homework	download and upload
Wed	Metadata Ex1: basic steps and tools	↓ discussion, markup and packing
Thu	Technical details (optional) Ex2: full workflow	↓ collaboration-specific aspects
Fri	Feedback + wrap-up	present 1 slide per collaboration