FAIR Data

Working Groups

Hands-on Workshop

June 14, 2023

Overview

- Background to storing data
- Introducing FAIR
- Lattice data goes FAIR

Introducing databases

- In the 1960s companies such as IBM started to develop databases to store information in a systematic way
- Sabre (travel reservation system) was the first commercial large scale database
- After some experimentation, relational database were developed where the information is stored in tables and a query language called SQL was used to extract information from the tables.
- There is a schema that defines the column names and data types in the table.

FarmID	Name	Acres
1	Black Hallow Farm	500
2	Robinwood Farm	4000

Table: Farm table

The evolution of databases

From the 1970s to the late 1990s relational databases were dominant.

However

- Relational databases didn't map well to object oriented languages such as c++, so object orientated databases were developed (but not widely used.)
- The development of the original web motivated researchers to develop a **semantic web**. There is a lot of information stored in a web page. Tim Berners-Lee wanted it be possible to use this information in a systematic way. The semantic web didn't take off, but one of the technologies developed was XML.
- The storage of large quantities of social media posts and images motivated researchers to develop databases which were not based on tables. Examples are graph and document databases. Generically these are known as **NoSQL databases**.

What is the dream?

From https://www.physics.nat.fau.eu/2021/07/05/fairmat-lifting-the-treasure-trove-of-materials-data/ Using optoelectronics as an example, the goal is to discover and investigate highly efficient, low-cost, and nontoxic semiconductors with optimal properties for devices and systems for renewable energy generation and conversion.

- If scientific data from many different sources (experiments or simulations) can be searched, then this could speed up scientific discovery
- Data from the different sources need to be combined.
- The data needs to be properly described.
- The data needs to be readable by both humans and **machines**.
- Ideally the data needs to be persistent.

Between 2006 and 2016 Google hosted google code project. Archive still exists.

FAIR Principles

Findable
Accessible
Interoperable
Reusable

force11.org
⋮
Wilkinson 2016
⋮
go-fair.org

- It is becoming a mandatory requirement by funding agencies
  "The [European] Commission will work with global policy and research partners to foster cooperation and to create a level playing field in scientific data sharing and data-driven science."
  EU Commission, COM(2016)178
- provides guiding principles, not an implementation
- conceptually refers to three types of entities:
  - data = any digital object
  - metadata (MD) = information about digital object
  - infrastructure
- requires machine actionable (meta)data

What does “findable” mean?

Findable

- F1 globally unique and persistent ID assigned to (M)D
- F2 data described with rich MD
- F3 MD includes data ID of data
- F4 (M)D registered or indexed in a searchable resource

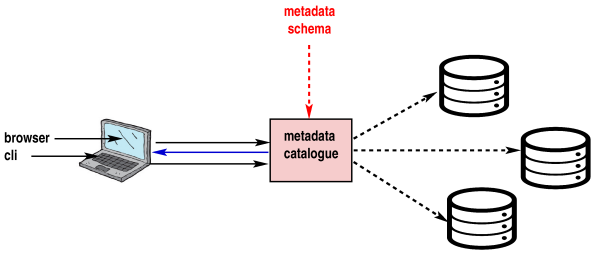
Metadata includes information on

- content (general and domain-specific vocabulary)
- provenance (who, when, where, how?)
- access (format, path, license, . . .)
- . . .

How does ILDG address “findable”?

Metadata

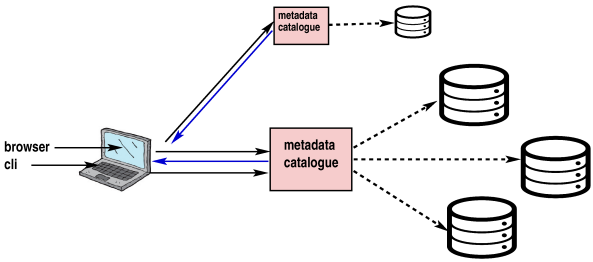
- follows a well-defined and rich schema
- stored separately from data (big)
- searchable in central catalog of each RG (regional grid)



How does ILDG address “findable”?

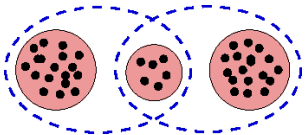
Metadata

- follows a well-defined and rich schema
- stored separately from data (big)
- searchable in central catalog of each RG (regional grid)



Unique identifiers

- Ensembles: have only MD (content, access permissions, ...)
  `mc://⟨rg⟩/⟨collab⟩/⟨proj⟩/...`
- Configurations: MD (related ensemble, provenance info) **and** actual data
  `lfn://⟨rg⟩/⟨collab⟩/⟨proj⟩/...`



ID	entity	relation	content	data storage	access control
`lfn`	config	mc	yes	yes	no
	↓				↑
`mc`	ensemble	—	yes	no	yes
	↑↑↑				
∗)	publication	set of mc	yes	no	no

∗) ILDG 1.0 has no official registration of IDs or publication metadata yet!

Introducing DOI

Example of DOI:
https://doi.org/10.22323/1.430.0203

- A digital object identifier (DOI) is a persistent identifier or handle used to uniquely identify various objects, standardized by the International Organization for Standardization (ISO).
- DOIs are widely used to identify academic, professional, and government information, such as journal articles, research reports, data sets, and official publications.

https://en.wikipedia.org/wiki/Digital_object_identifier

DOI and Data Publishing

Data Publishing

- Registration of persistent identifier (DOI)
- Metadata for registration (DataCite)
- Landing Page (hosting and automatic generation)
- Harvesting of metadata

Exploratory setups by JLDG and USQCD

- using national registration authorities (JaLC, OSTI)
- workflow and metadata for registration and generation of landing pages

Possible directions in ILDG 2.0

- establish workflow for registration, generation and hosting of landing pages (e.g. Zenodo)
- extended metadata support
- dedicated metadata harvesting (e.g. by INSPIRE)
- common registration authority

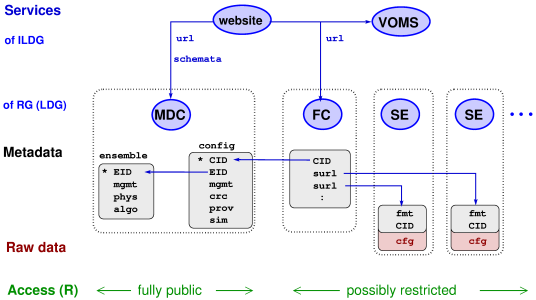
What does “accessible” mean?

Accessible

- A1 (M)D retrievable by ID using standardized protocols
- A1.1 protocol is open, free, and universally implementable
- A1.2 protocol allows authentication/authorization procedure where necessary
- A2 MD accessible even if data is no longer available

- A1 can be achieved e.g. by a File Catalog: ID $\mapsto$ storage location(s)
- Accessible does not imply (unrealistic) public access without authentication
- MD is precious even without the associated data

How does ILDG address “accessible”?



- all metadata is publicly accessible (from MDC)
- well-defined community-wide metadata schema
- metadata available in a standard markup language
- standardized protocols and API of services for access to data and metadata

What does “interoperable” mean?

Interoperable

- I1 (M)D use a formal, accessible, shared, and broadly applicable language
- I2 (M)D use vocabularies that follow FAIR principles
- I3 (M)D include qualified references to other (M)D

- ability of data (or tools) from non-cooperating resources to integrate (or work together) with minimal effort

How does ILDG address “interoperable”?

Common standards for

- Metadata schema
- Data format
- API and URL for web services of regional grids

New directions:

- Extend ILDG format to include support for HDF5
  - definition of ILDG packing rules
  - convenient tools for packing and conversion
- Token-based authentication
- REST API

What does “reusable” mean?

Reusable

R1	(M)D richly described with plurality of accurate and relevant attributes
R1.1	(M)D released with clear and accessible data usage license
R1.2	(M)D associated with detailed provenance
R1.3	(M)D meet domain-relevant community standards

- reference to a paper may not be sufficient
- good scientific practice $\leftrightarrow$ FAIR
- also related to verifiable invariance of results (see presentation by Ed Bennett)
  - reproducibility: same data + same analysis
  - replicability: new data + same analysis
  - robustness: same data + new analysis

How does ILDG address “reusable”?

Ensemble MD

- Physics
- Algorithm
- Management

Config MD

- Markov step
- Implementation (machine, code)
- Management (creator, date, checksum)

But no license aspects! (cf. R1.1)