FAIR Data at Synchrotron SOLEIL

E. Farhi, *EXP/GRADES*

DIADEM School 2025

La science éclaire l’avenir
Science lights up the future

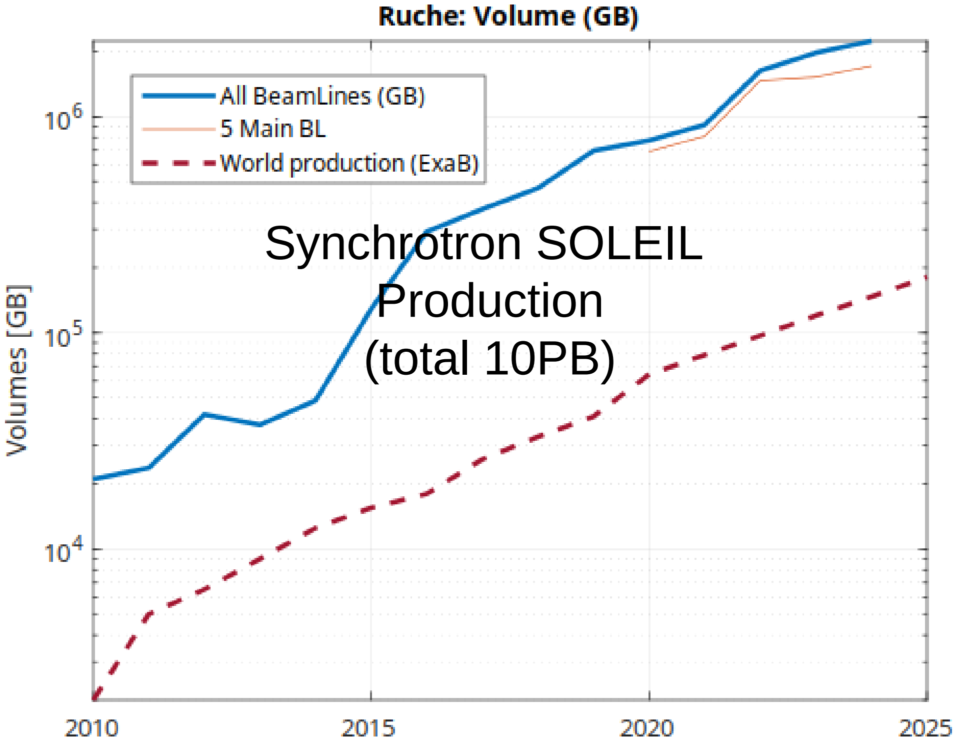
The Data Deluge

- Data Deluge aka Information Explosion
  - Raises difficulties to handle the generated amount of information.
- Can not be ignored
  - Motivation to increase resources, security, etc.
  - Produced data is assumed to be valuable, and should be kept for the future.
- Bypass humans ability to handle data
  - We now use computers everywhere for that purpose.

shutterstock.com · 86456998

The Data Deluge

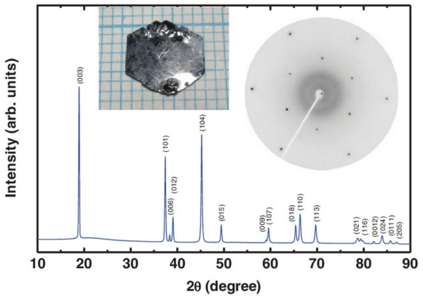
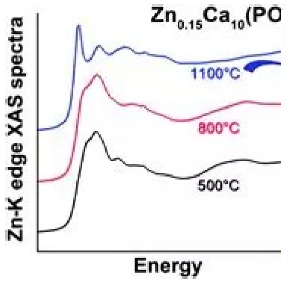
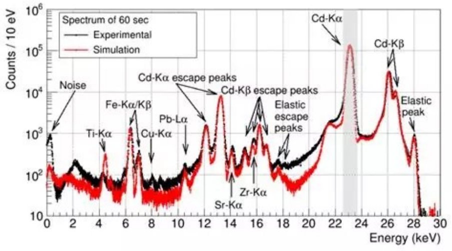
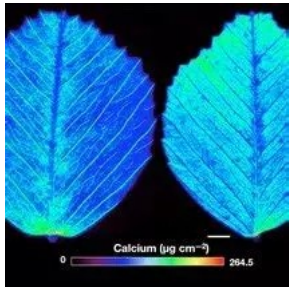
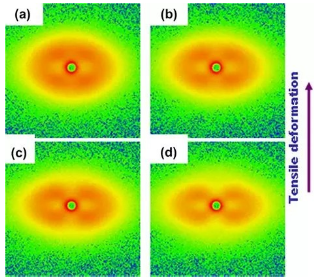
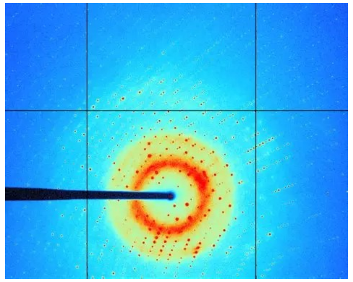
- First mention of Data Deluge: 1964 (1955)
  - Not new ! It’s been there for ever but we deal with it.
  - Probably not an issue *per se*.
- Data production mostly follows Moore’s law
  - But is now a little slower in fact (x2 every 2.4y).
  - No technological limit (except water and energy).
- The software we use can handle massive data
  - Software development follows requirements and hardware capacity.
  - We use multi CPU+GPU.
  - Storage is not expensive.



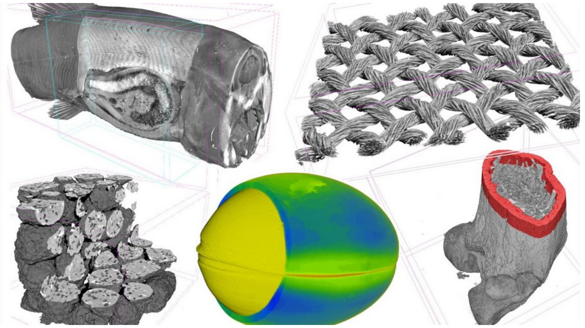
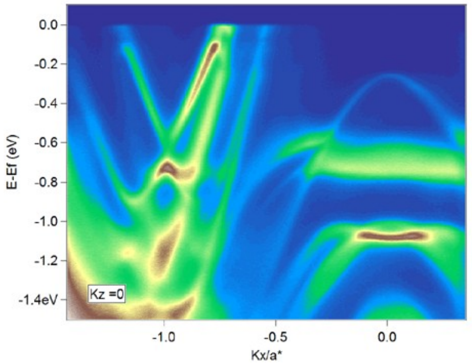
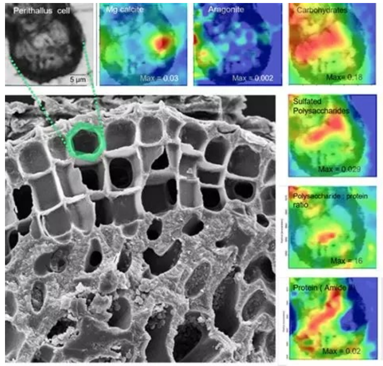
Synchrotron SOLEIL Production (total 10PB)

There is NO data deluge (tech follows)

Typical data at SOLEIL

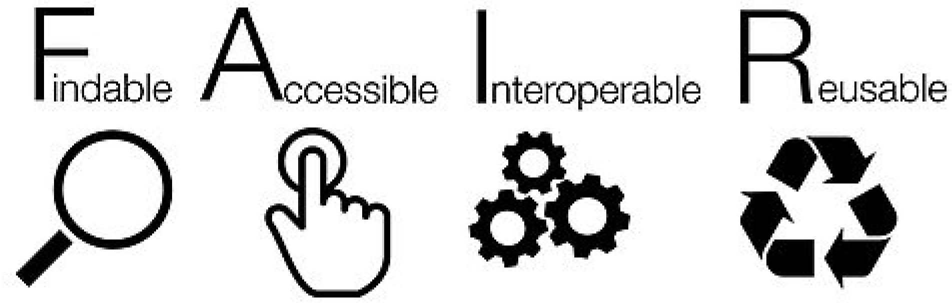
Dimension	Experiment	Examples
1D	Powder Diffraction Spectroscopy (absorption, fluorescence, XPS, RIXS/IXS)	  
2D	Texture diffraction, small angle scattering, protein crystallography, microscopy, ptychography	  

Typical data at SOLEIL

Dimension	Experiment	Examples
3D	tomography, ptychography, ARPES	  0.0 -0.2 -0.4 -0.6 -0.8 -1.0 -1.2 -1.4eV E-Ef (eV) Kz =0 -1.0 -0.5 0.0 Kx/a*
hyperspectral	microscopy-fluo/abs/diff Time-resolved 2D	 Perithallus cell 5 µm Mg calcite Max = 0.03 Aragonite Max = 0.002 Carbohydrates Max= 0.18 Sulfated Polysaccharides Max = 0.029 Polysaccharide : protein ratio Max = 16 Protein (Amide Max = 0.02

What is FAIR Data ?

- We produce lots of data (“*the data deluge*”).
- Initiated by a consortium in 2016 (https://doi.org/10.1038/sdata.2016.18).
- The goal of FAIR principles is to ensure a long life time of the data and knowledge, especially using computers.



Could also be F.A.R$_H$ I 🙂

FAIR principle

- **Findable**:

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

  F2. Data are described with rich metadata (defined by R1 below)

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

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

FAIR principle

- **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

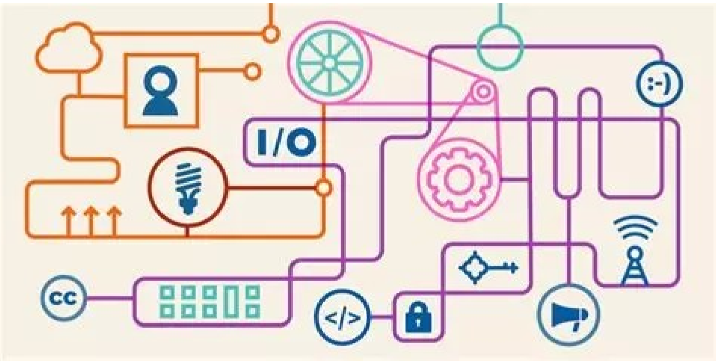
FAIR principle

- **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



The “*ontology*”
= hierarchical nomenclature

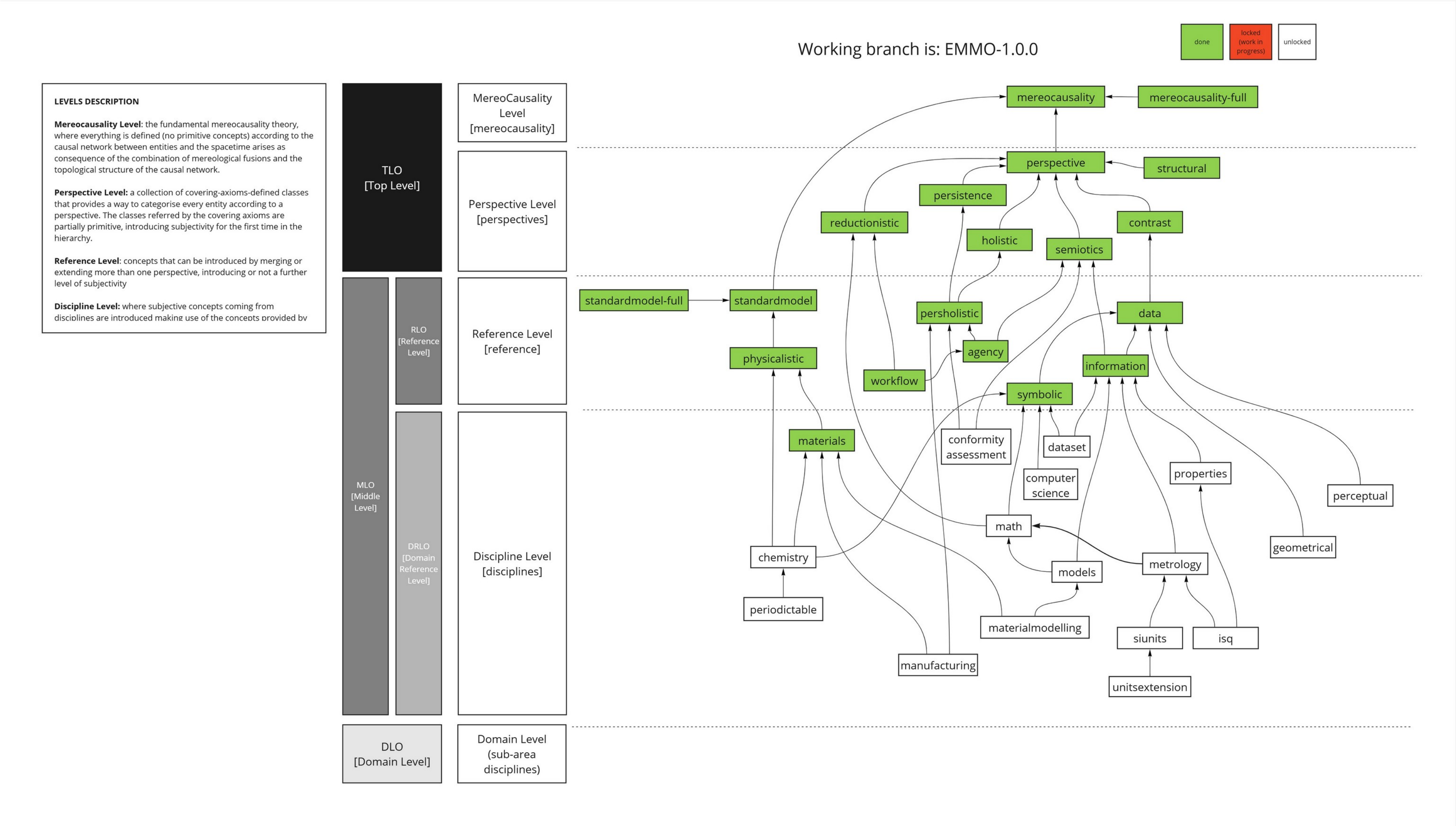
FAIR principle

- **Reusable**:

  R1. (Meta)data are richly described with a plurality of accurate and relevant attributes

An example of “ontology”

See : https://github.com/emmo-repo/EMMO



LEVELS DESCRIPTION

Mereocausality Level: the fundamental mereocausality theory, where everything is defined (no primitive concepts) according to the causal network between entities and the spacetime arises as consequence of the combination of mereological fusions and the topological structure of the causal network.

Perspective Level: a collection of covering-axioms-defined classes that provides a way to categorise every entity according to a perspective. The classes referred by the covering axioms are partially primitive, introducing subjectivity for the first time in the hierarchy.

Reference Level: concepts that can be introduced by merging or extending more than one perspective, introducing or not a further level of subjectivity

Discipline Level: where subjective concepts coming from disciplines are introduced making use of the concepts provided by

FAIR principles at SOLEIL

- **Findable**
  - ☑ Data files are stored in HDF5/NeXus with associated metadata from the community (*ontologies*).
  - ☒ Not yet: SciCAT is an interface that will allow to search for data sets. It will add DOIs and an API.
  - Alternative: Zenodo (DOI, URI).
- **Accessible**
  - ☑ Globus is used to retrieve data, but is rather complex for users.
  - ☒ Not yet: No browser based data broker. SciCAT will provide it. A 3-years embargo will be enforced.
  - Alternative: Zenodo (DOI+API+browser+storage, 30 years persistence). Our group provides a JupyterHub data browser for assigned experiments.

FAIR principles at SOLEIL

- **Interoperable**
  - ☒ Not yet: The SOLEIL dataset metadata is not yet fully compatible with adopted standards (NeXus, NFFA/NFDI, Big-Map, PanOSC).
  - Alternative: ZARR format, PhySci ontology.
- **Reusable**
  - ☒ Not yet: The metadata is not yet complete. Needs more items, in connection with electronic logbooks and experiment configurations.
  - Alternative: License (CC0), XAS and Ptycho community.

FAIR tools chosen at SOLEIL

- **SciCAT <https://www.scicatproject.org/>**
  - An EU catalog for synchrotron data, as a web-service. Handles DOI’s.
  - Alternative: iCAT, Zenodo, HAL.
- **Globus <https://www.globus.org/>**
  - A US UChicago non-profit, but commercial service.
  - Alternative: SFTP, GridFTP, NextCloud, CopyParty.
- **NeXus/HDF5 <https://www.nexusformat.org/>**
  - A non-profit, but commercial company. Used since 2006 at SOLEIL. Most data are stored in this format. Smaller data producers use *e.g.* text, Igor, TIFF, etc.
  - Alternative: ZARR.

low-cost FAIR ?

- **With limited resources, you may comply with FAIR principles by adopting the following rules and tools:**
  - Specify **metadata** for your data (e.g. identification, configuration, abstract, key results, etc).
  - Comply with existing nomenclatures, if any. Keep it simple.
  - Specify a **license** for property and reuse (CC0, CC-BY...).
  - Choose a **hierarchical data format** (a directory, ZARR, HDF5, …).
  - Store metadata with data (embedded, JSON, YAML, XML).
  - Store your data locally (store metadata on Zenodo, and ensure access to your storage) or send it all to Zenodo.

Job done

FAIR-ly simple, hey ?