FAIR Data – why and how

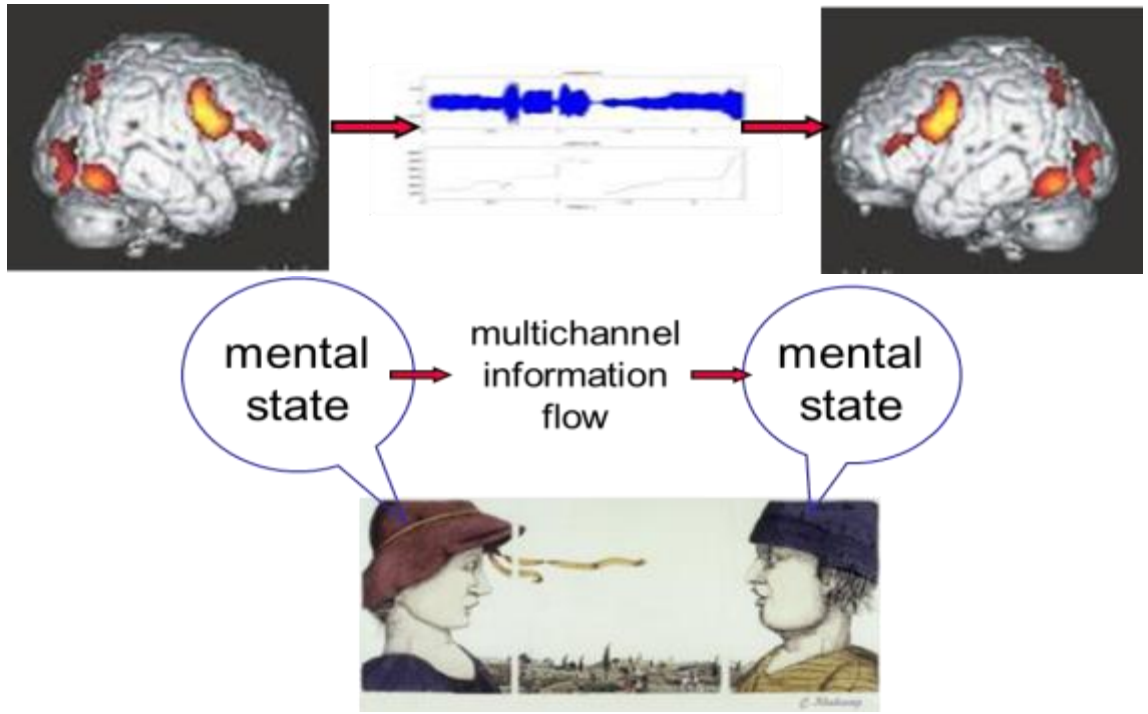
Peter Wittenburg

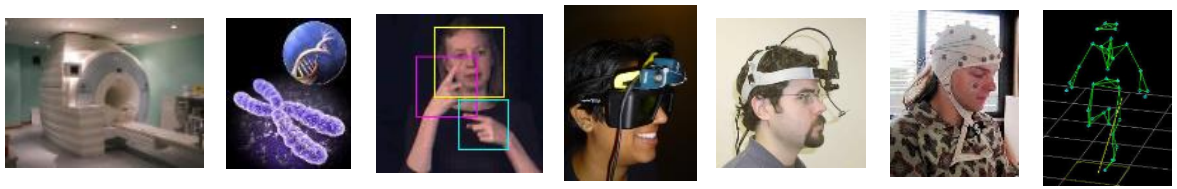
FAIR **DIGITAL OBJECTS** FORUM

1. Short bio
2. About metadata, persistent IDs and the FAIR Principles
3. State of data
4. Globally integrated dataspace and FAIR Digital Objects

Short Bio

Responsible for technology and methodology in a Max Planck Institute for understanding the human brain

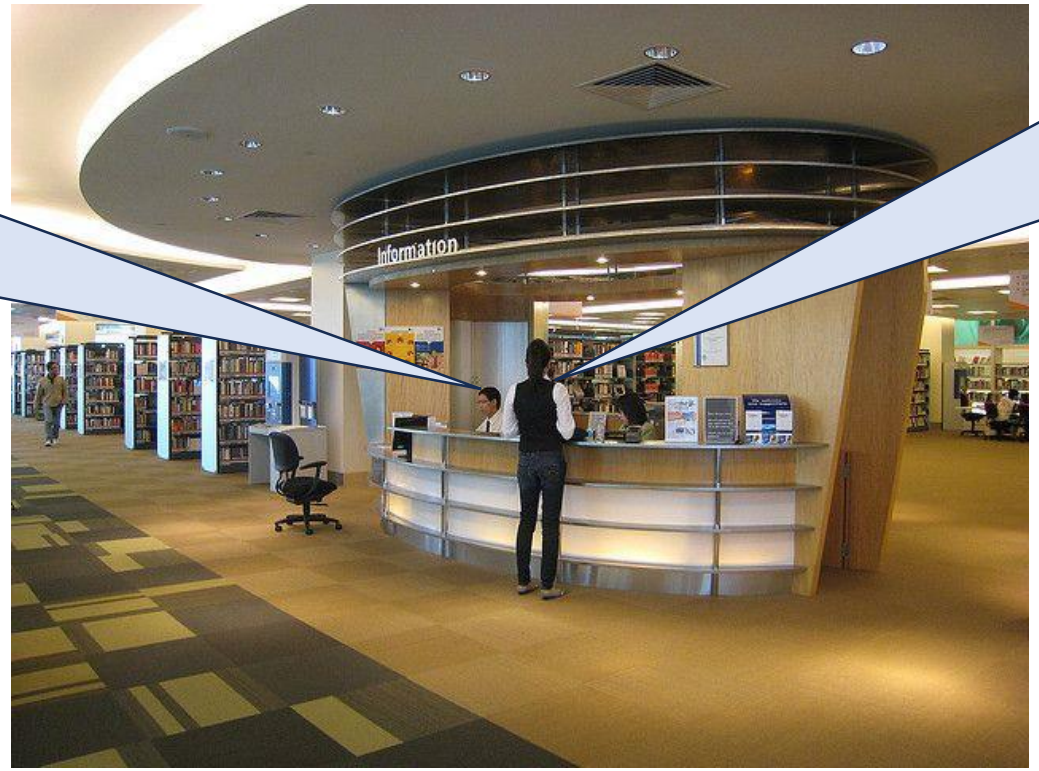




- Modern science is data driven
- Applied advanced pattern recognition using data hungry algorithms
- Started building data infrastructures to improve access to data from other institutes
- Responsible for DOBES, CLARIN and EUDAT research/data infrastructures
- Started Research Data Alliance 2013
- Started FAIR DO Forum 2019

Library Scene

I am afraid I can‘t help you. Do you have the ISBN number, author name, etc.?

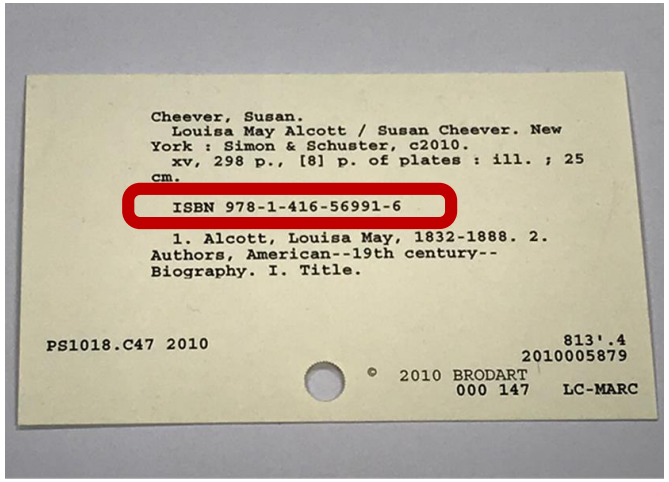


I need the book that was in row 20/ shelf 11/layer3 about 5 years ago.

Of course, the place in a shelf is NOT a persistent identifier of a book.

The library might have undergone a change, might have several copies, etc.

Library PIDs and Metadata

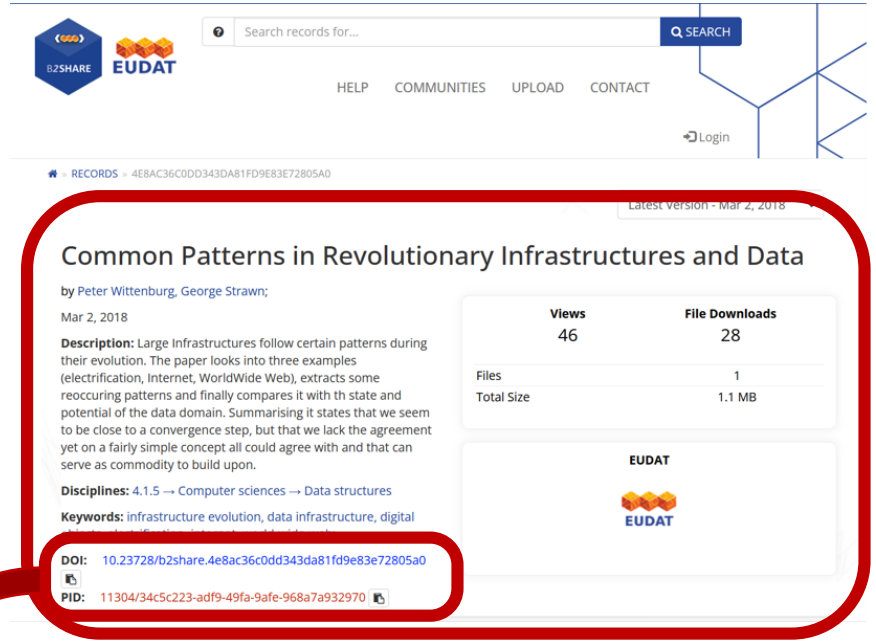


A book has a Persistent ID (ISBN) and a Metadata Card.

It will direct you where you can find the location of the printed copy.

Transformation to an all digital world

EUDAT B2Share Repository



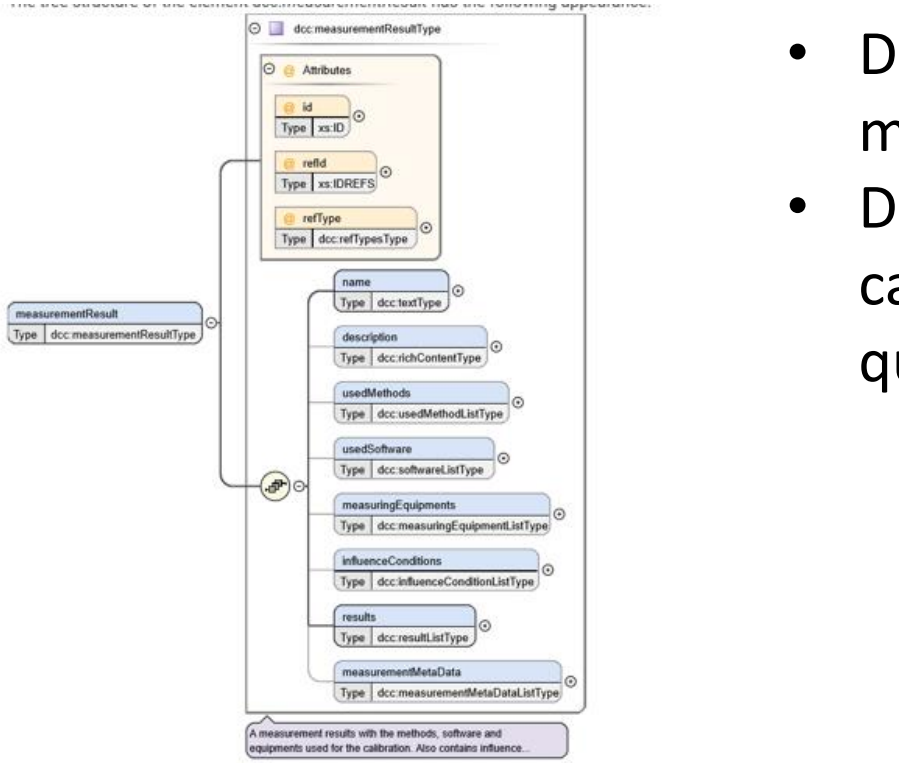
Metadata according to a simple schema (structure & vocabulary)

A digital object has a DOI/Handle as Persistent ID.

A URL is a location in a store!

Many Types of Metadata

Digital Calibration Certificate

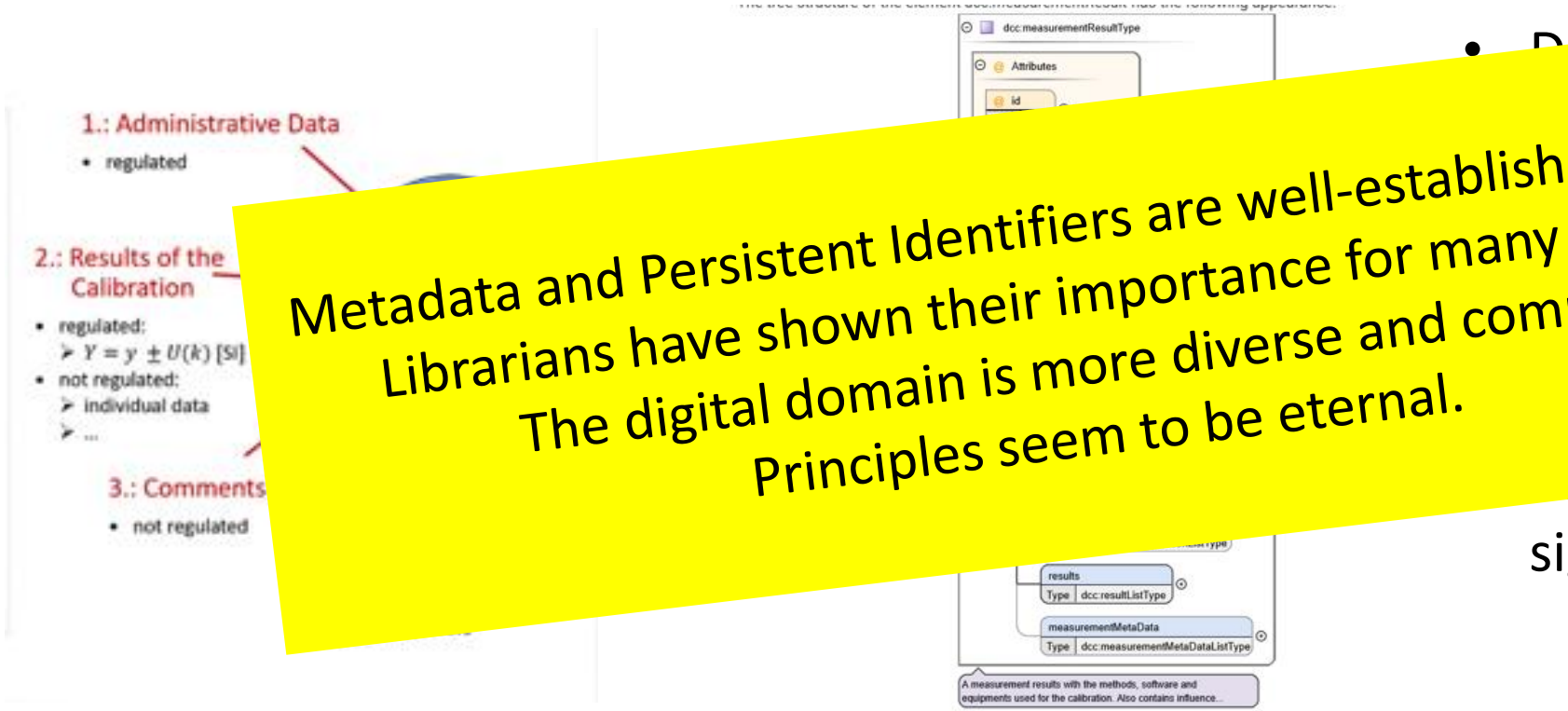


Extract from the XML schema: type definition "measurementResultListType"

- DCC is a typical complex metadata schema
- Describes the act of a calibration to document quality etc.

Many Types of Metadata

Digital Calibration Certificate



Extract from the XML schema: type definition "measurementResultListType"

Metadata and Persistent Identifiers are well-established Concepts.
Librarians have shown their importance for many activities.
The digital domain is more diverse and complex.
Principles seem to be eternal.

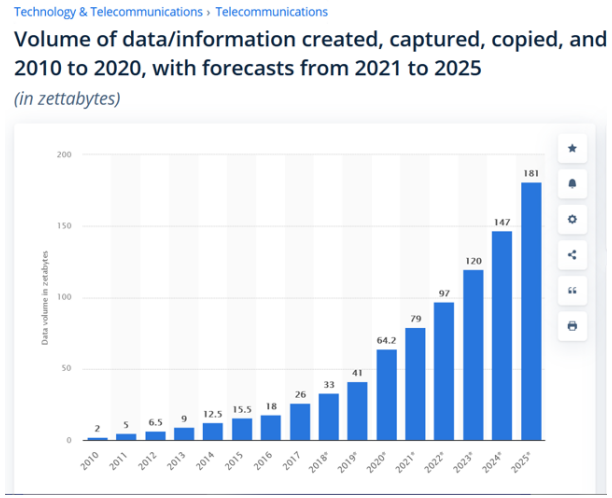
- DCC … complex
  … na
  … t of a
  … cument
  … ecification
  … access rights and requires signatures

Books and Data

- Libraries are about books, papers and media (publications)
  - there are about 150 million books in total
  - books are human readable and in libraries (findable, accessible, interoperable, reusable)

- Research repositories are about reusable data in varying circumstances
  - Our small MPI data repository had already ~1 million digital data objects in 2012
  - Metadata: Searching for data with same brain diseases, same age, etc.
  - Climate simulation people create about 100.000 DOs in a simulation run
- The number of Digital Objects cannot be counted anymore
  - Quantity requires new mechanisms

- Of course, want to access all research artefacts
  - Open Science in 16xx: make results available to peers
  - Open Science in 202x: make research artefacts available



State of Data


- Surveys: 80% of time and money in data-projects is wasted with data munging
- P. Heidorn (U California): in 20 years > 80% of our data will be lost
- V. CERF (Internet pioneer): we are risking the Digital Dark Age
- R. Kahn (Internet pioneer): we lack a persistent, stabile & global data management solution
- T. B. Lee (Web pioneer): we need a radical change – a new architecture


- Digital data to become findable
  - metadata is required - supporting domain vocabularies
- Digital data to become accessible
  - Needs a general access protocol
- Digital data to become interoperable (not only human readable)
  - software is required – needs structure and semantic definitions
- Digital data to become reusable
  - Metadata is mostly poor and rights are not clear

FAIR ≠ Open Data

State of Data


- Surveys: 80% of time and money in data-projects is wasted with data munging
- P. Heidorn (U California): in 20 years > 80% of our data will be lost
- V. CERF (Internet pioneer): we are risking the D
- R. Kahn (Internet pi ent solution
- T. B


The state of data is bad.

Measures of improvements are urgently required to improve the state of data.

Need to urgently reduce costs to make data projects feasible for many.

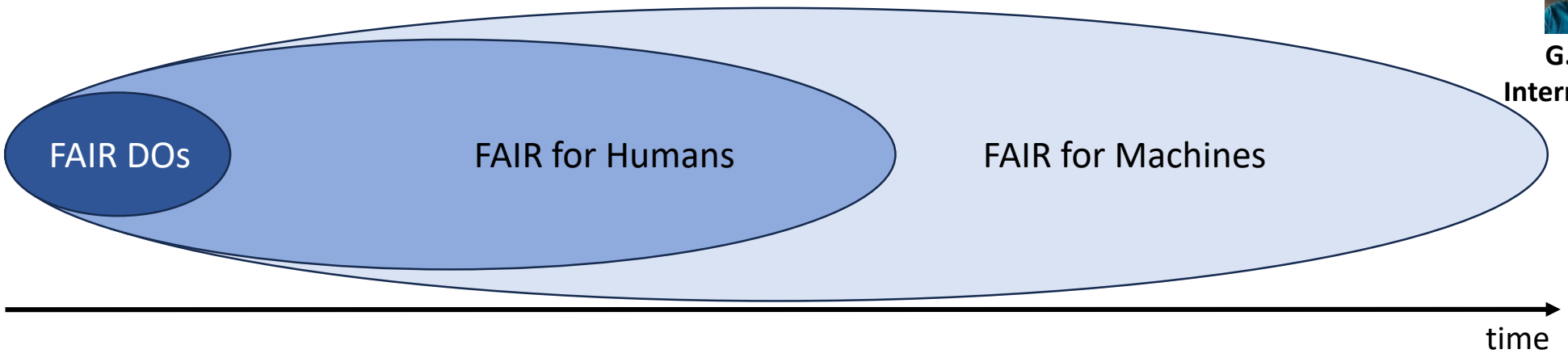
- Digital d
  - meta
- Digital d
  - Needs
- Digital da perable (not only human readable)
  - softwa s required – needs structure and semantic definitions
- Digital data to become reusable
  - Metadata is mostly poor and rights are not clear

FAIR ≠ Open Data

Feasibility

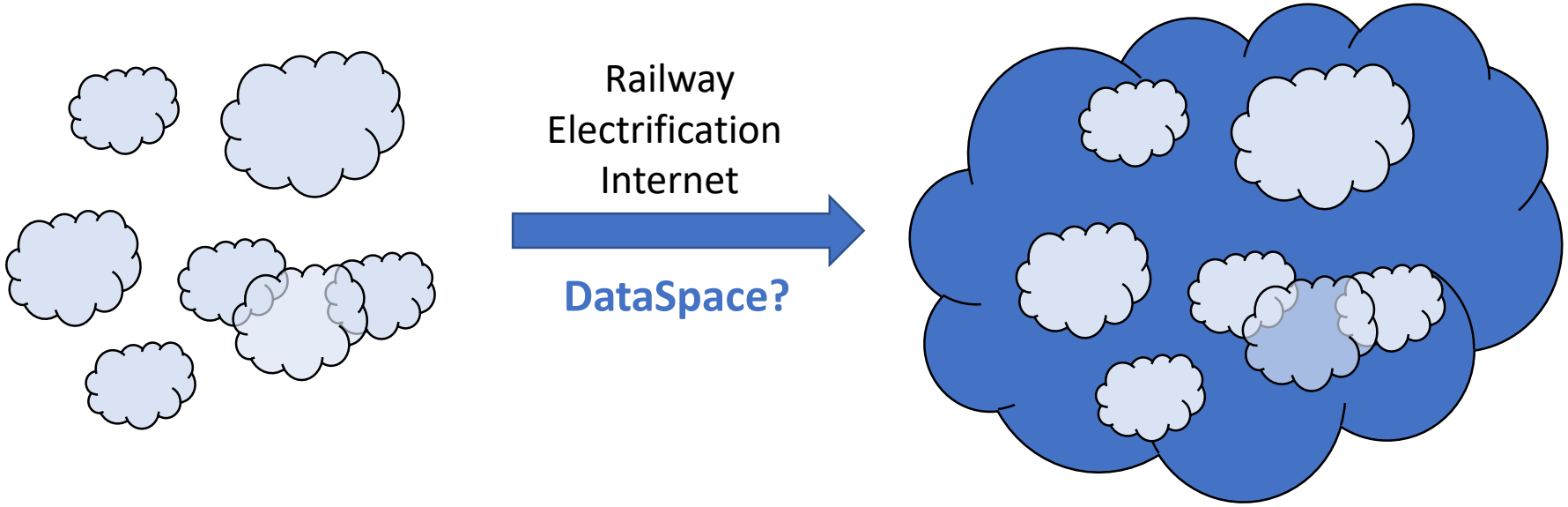
- 100 % FAIR implies a huge transformation of repository strategies
  - Globally we have thousands of repositories – mostly underfunded
  - This transformation will cost much money and time
  - A fast integration as a first step is, however, possible for far less money

G. Strawn
Internet Pioneer



- For a basic integration we need to repeat the Internet success story
  - Creating a minimal Interop standard to connect all existing and emerging islands
  - Not trying to change everything at once

Globally Integrated Dataspace

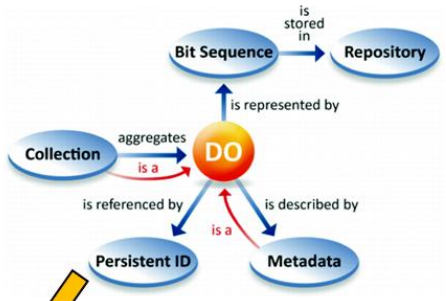


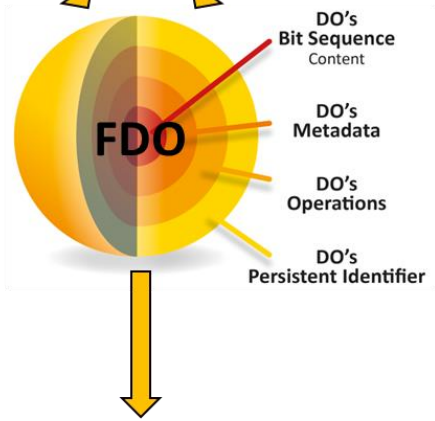
- Needs to be based on FAIRness and increase TRUST
  - Mechanisms to be based on Responsibility, Accountability and Persistency
  - Mechanisms should be distributed and support global governance

FAIR Digital Objects similar to Internet Basics

FAIR Principles

Digital Objects





2023: FDO Requirements Specifications ready
March 2024: FDO Summit with many applications

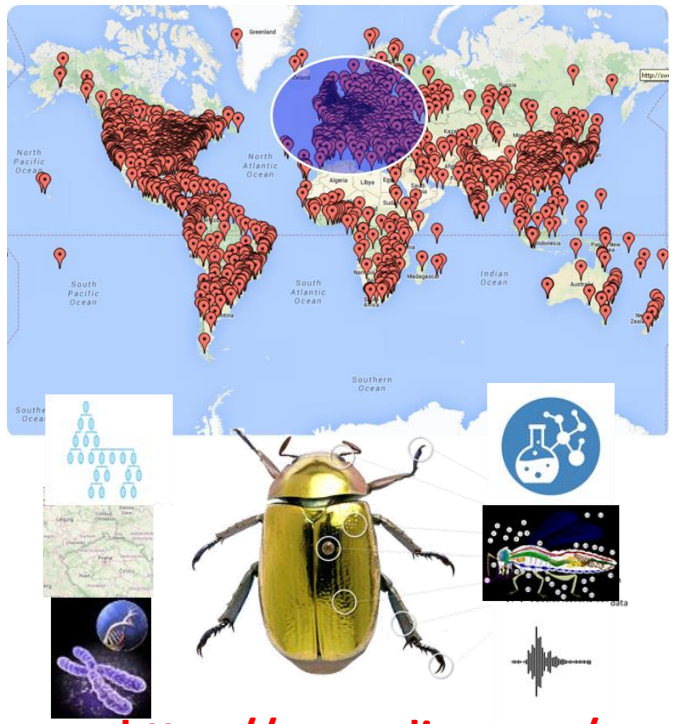
FDOs

- are persistent atomic entities of FAIR information identified by a PID
- follow minimal specifications about persistently bundling all information to enable processing of an object
- do not require to have FAIR repositories right away
- increase the protection of data where necessary

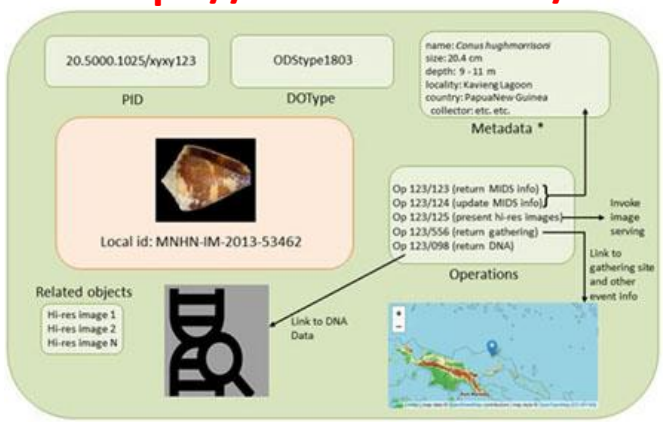
FDO Forum

- Independent
- No patents etc. involved
- participate

FDOs to structure the complex Data Landscape



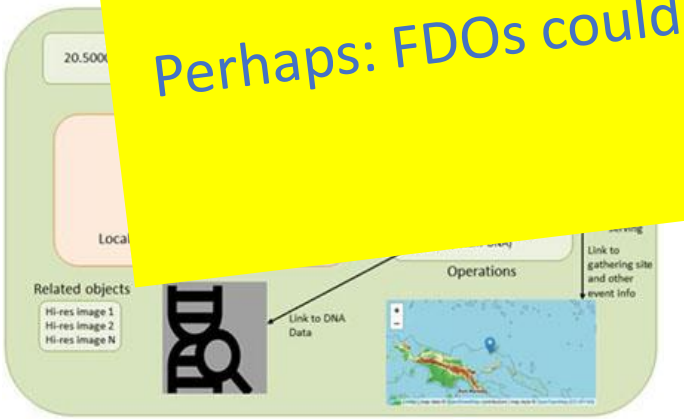
https://www.dissco.eu/



- global Biodiversity-Infrastructure
  - 120 Natural Science Institutes in EU
  - 1000 collections, 3 billion objects, 2 million standards
  - trillions of relations (storing knowledge)
  - Scattered information about specimen

- Task1: build on integrated dataspace and virtually bundle all information belonging to one specimen
- Task2: enable the execution of specific software on the digital twin of an object (which is an FDO)

FDOs to structure the complex Data Landscape



FDOs can be used NOW to build a stable global data infrastructure.
FDOs can be used NOW to structure complex dataspaces and thus create knowledge structures for coming generations.
FDOs can be combined with data protection technology (EDC).

Perhaps: FDOs could form the basis for a global quality infrastructure for metrology.

- global Bi

rds

ally

hen

e on the

FBO)

Thanks a lot for your attention.

FAIR **DIGITAL OBJECTS** FORUM

https://fairdo.org/