Digital Transformation towards Interoperability

Peter Wittenburg

Max Planck Society – Max Planck Computing & Data Facility

acknowledgements to many distinguished experts:

Giridhar Manepalli, George Strawn, Bob Kahn, Larry Lannom, Ulrich Schwardmann, Alex Hardisty, Dimitris Koureas, Maggie Hellström, Koenraad de Smedt, Carlo Zwölf, etc.

Big Challenges

1. Stability of Scientific Memory in Digital Times - preventing the Dark Digital Age

- paper, ISBN and copies in national libraries were a promise to society
- need a promise for the persistence of relevant data, relations and code (Scientific Memory) in the Digital Era

2. Making Data Management & Processing Efficient and Cost Effective

- we spend about 80 % of time & costs in data projects with data wrangling (**across sectors**)
- in US >500 billion $ per year lost in health sector due to non-FAIR data (WEF Davos)
- 10.000s of tools, 1.000s of repositories, 100s of standards & protocols to cope with and all are different with respect to the way they deal with data – fragmentation and heterogeneity
- as with TCP/IP: looking for a universally accepted & patent-free integration standard that does not hamper dynamic developments and “competition” (Cloud Systems to not address the FAIR layer)

3. Enabling Semantic Cross-walk (-> Daan)

4. Maintaining Data Sovereignty in a Global Integrated Data Space (GIDS)

5. Keeping Europe in a Competitive Position in Data Use/Analysis

not explicitly in my talk

Phenomena in Data Science

(deep analysis of 75 research infrastructure reports/plans in Spring + Summer 2020)

FAIRness Paradox

- almost all researchers have heard about FAIR and support the idea
- but daily practices in the labs did hardly change in the last 5 years
- FAIRness shifted to **FAIRness by Publication (FbP)** instead of **FAIRness by Design (FbD)**

Production Chains

- creating research results includes **chains of specialised actors and labs**
- FAIRness is shifted to the next actor in the chain – finally no one does it

Phases: where are we?

- George Strawn (Internet Pioneer, Advisor to US Administration)
  - 1950s: many computers + many data sets
  - 1990s: one computer + many data sets (the network is the computer)
  - 2030s: one computer + one data set (will have the Global Integrated Data Space – GIDS)

- EOSC/NFDI/EDI/etc. are about building the basis for new types of data/research infrastructures and EU/MS are investing much money! (US may follow in 3 years)
- Building efficiently usable infrastructures costs decades and involves politics/sociology, economics and technology!

- Thus, we need to anticipate the structures to be established in 10+ years complementing the evolutionary work already being done!
- Thus, will EU maintain the current advantage and generalise from our great experience with a decade of piloting with discipline data spaces?

From Connecting to the Internet towards Connecting to the Global Interoperable Data Space

Dream in 1970ies

a user simply plugs-in a computer
and is part of the Internet

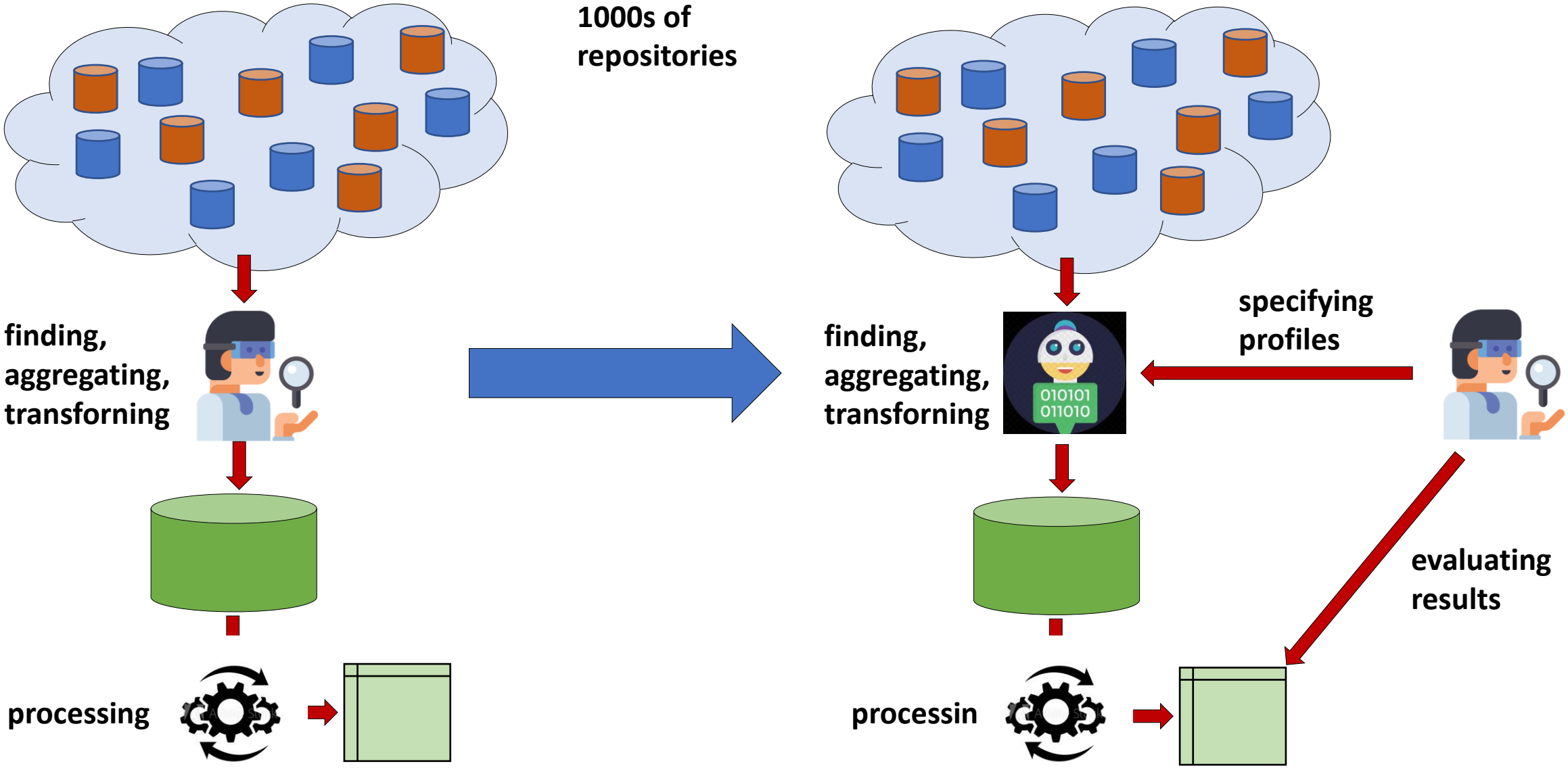
Dream in 2020ies

a community simply plugs-in a repository
and is part of the GIDs

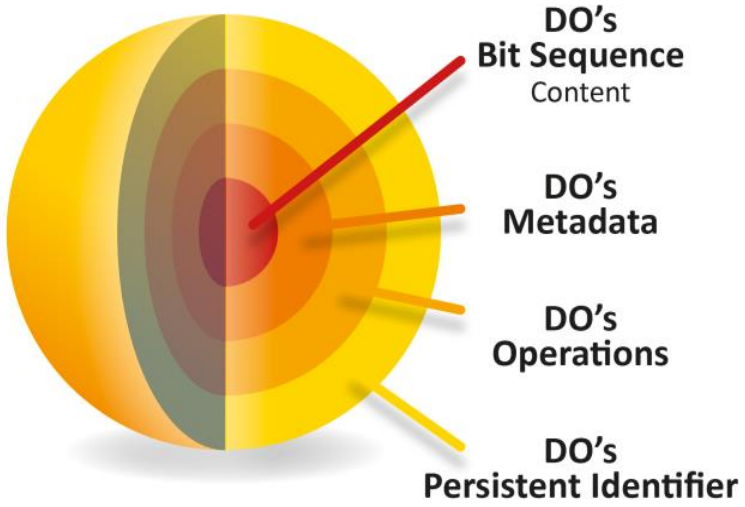
What does „plugging-into GIDS“ mean?

- need to have an entity that is clearly identifiable, self-contained and traceable
- need to have an entity that binds all relevant information persistently (type, metadata, rights, licenses, etc.) for reuse
- when a repository is added to the GIDS: it offers its holding (data, all metadata) to all interested crawlers by a DO Interface Protocol (DOIP) to update collections, registries & portals automatically
- when a user adds data to a repository: the repository updates its offers enabling crawlers to harvest
- managing trust relationships will be a challenge – PIDs help

Change of Researchers‘ Role



FAIR Digital Objects – just in time

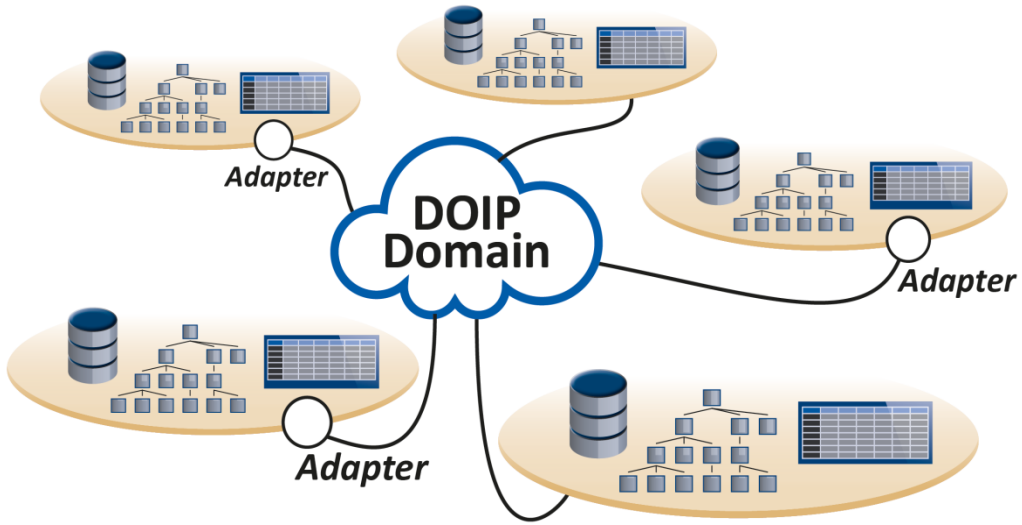


FAIR Digital Objects (FDO)

- are atomic and self-standing by bundling all relevant information to process digital concent (FAIR, machine actionable)
- have a Globally Unique, Persistent and Resolvable Identifier (-> Maggie)
- Identifier System (Handles/DOIs) is globally administered, distributed, secure, redundant, free of patents and is owned by the Swiss non-profit DONA Foundation
- DOIP acts comparable to TCP/IP: helps reducing complexity to 1*N in a complex, heterogeneous and fragmented landscape

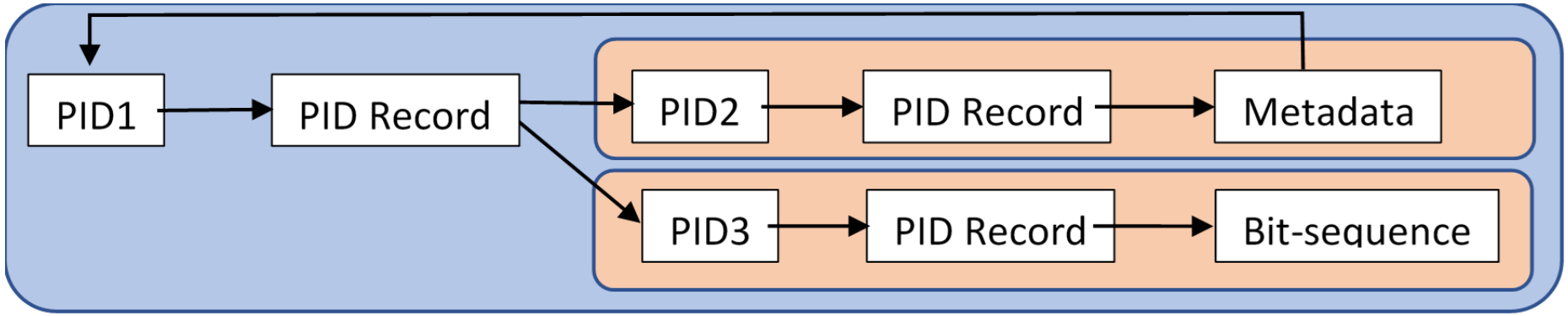
FDO characteristics

- abstraction (any content)
- persistent binding
- encapsulation
- a single protocol needed (DOIP)



Are FAIR Digital Objects FAIR?

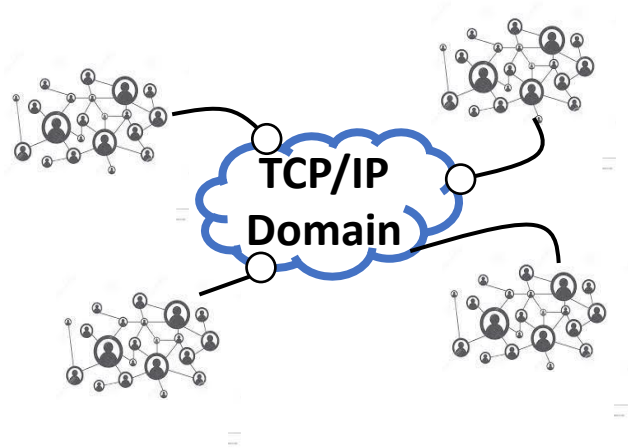
Typical Canonical FDO Example: FDO with metadata and two bit-sequences, themselves being FDOs.

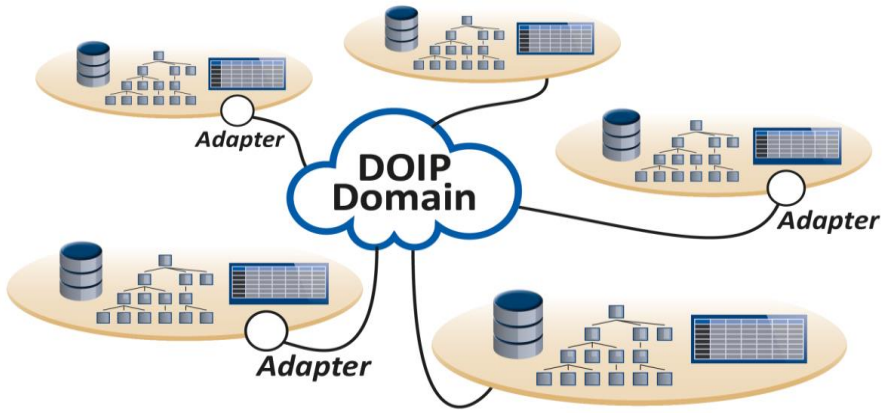


PID needs to resolve in predictable resolution result according to a registered profile	**Handle/DOI ok**	**URL?**
Attributes in PID record need to be defined & registered & thus machine actionable	**DTR ok**	**URL?**
Attributes that include references to MD and Bit-Sequences to be machine actionable	**DTR ok**	**URL?**
Metadata to be accessible, interpretable (mostly not machine actionable)	**ESFRIs ok**	**ESFRIs ok**
Metadata elements that refer to FDO need to be machine actionable	**?**	**?**
Bit-Sequences need to be accessible, interpretable (compliant with Type)	**in general ok**	**in general ok**

Making metadata provided by communities FAIR will be the challenge.

A Domain of FAIR Digital Objects





	Internet	Integrated Data Space
challenge	creating an integrated computer network	creating an integrated global data space
heterogenity	100s of networks	100s of standards, 1000s of repositories, 10000s of tools
differences	mode, packaging	data organisation and modelling
basic protocol	TCP/IP	DOIP
rights	no patents, no commerce	no patents, no commerce
achievements	complexity reduction, start of innovation	complexity reduction, start of innovation
key	broad social agreement	broad social agreement (?)

What is the state of FDO work?

- a clear specification called **FDO Framework**
- this is currently turned into technical specs allowing to develop **validators** in 2021 and to enable a "**plug-in" domain**
- a **DO Interface Protocol (DOIP)** specification, a software implementation and a reference repository (server)
- other basic softwre pieces such as **Type Registry** are ready as well

- **FDO Forum** is an independent initiative led by international experts which will be turned into a non-profit organisation (must be similar to Internet Society to prevent take-overs)
- FDO Forum is closely collaborating with RDA, CODATA, WDS, GOFAIR, EOSC

What about community uptake?

quite a number of communities and data centers are testing and piloting with aspects of the FDO concept – some we even don't know

- DISSCO, CLARIN, ICOS, material science, climate modelling, health, social science, biophysics,
- KIT, GWDG, CSC, SDSC, IU, CAS, NIST, GOFAIR, DTL, GESIS,

Trends in Climate Modelling Research (WCDC)

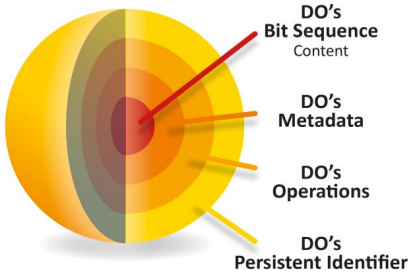
- CMIP 6 to include PIDs (Handles, DOIs) and FDO aspects
- increasingly more automatic workflows for data management & processing (Open Science by Design, reproducibility, etc.)
- from 30 CWFR abstracts for Data Intelligence Journal 8 from environmental sciences

250 experts on our email list, but still a small group of experts

Recommendations

- Need to anticipate requirements for next decades and invent suitable structures.
- Need to capitalise on more than a decade of experience in infrastructure building and dare to abstract from SILO approaches to generic solutions.
- Decisive is not **where** we will store digital objects (clouds) but **how** (FDO) to prevent dark digital age, improve FAIRness, drastically improve efficiency and build-in security, to support tracking, validation etc. and to foster sovereignty & trust
- Need a system of persistent registries and services (PIDs, Types, Mappings, Provenance, etc.) and clarify responsibilities
- **It's time to invest in large FDO testbeds and reference architectures**
- Of course: invest in training young experts (courses, hackathons, etc.)
- Help enthusiastic young people to build smart services (portals, specific mappings, etc.)

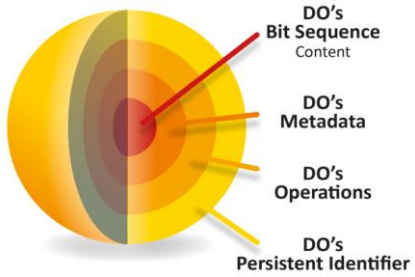
FAIR DIGITAL OBJECTS FORUM
https://fairdo.org/



Thanks so far.

- Wittenburg & Strawn: **Common Patterns in Revolutionising Infrastructures & Data**; http://doi.org/10.23728/b2share.4e8ac36c0dd343da81fd9e83e72805a0
- de Smedt, Koureas & Wittenburg: **Analysis of Scientific Practice towards FAIR Digital Objects**; http://doi.org/10.23728/b2share.e14269d07ce84027a7f79ee06b994ef9
- **FDO Framework**: https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/FDOF
- **Paris Workshop**: https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/Paris-FDO-workshop
- Jeffery, et.al.: **Not Ready for Convergence in Data Infrastructures;** *Data Intelligence* (2021) 3 (1): 116–135, https://doi.org/10.1162/dint_a_00084
- **DOIP V2.0:** https://www.dona.net/specsandsoftware

FAIR DIGITAL OBJECTS FORUM
https://fairdo.org/



Demonstrators

Purpose:

- demonstrate that FDOs are useful and that DOIP is operational and reduces complexity
- show that FDOs are ready for building larger reference implementations

Contributors:

Giridhar Manepalli (CNRI), Christophe Blanchi (DONA), Larry Lannom (CNRI), Marcel Hellkamp (GWDG), Philipp Wieder (GWDG), Chris Aroyo (CSC), Andreas Pfeil (KIT), Antti Pursula (CSC), Rob Quick (U Indiana), Mark van de Sanden (SARA), Rainer Stotzka (KIT), Peter Wittenburg (MPCDF)

Three demonstrators

KIT Karlsruhe:

- new FDO, FAIR Handle generation, validation, dissemination, search support

U Indiana:

- scientific workflow with systematic Handle generation, DOIP usage, tracking

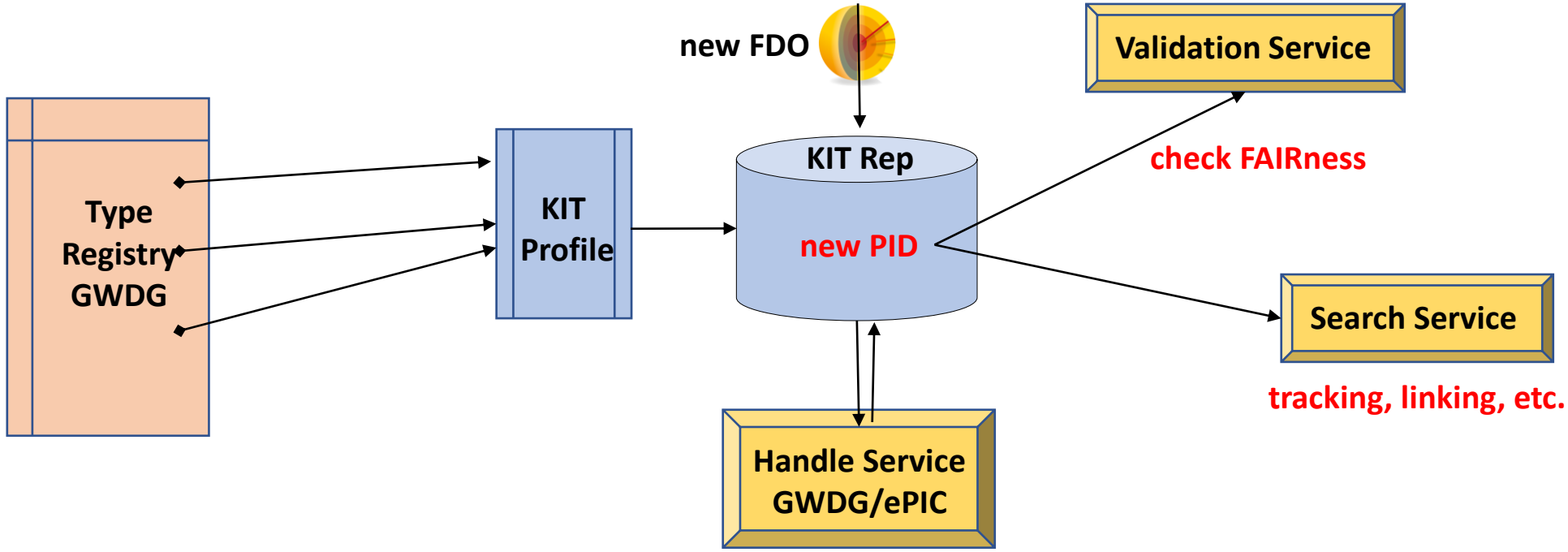
GWDG – CSC – CNRI (US):

- exchange (F)DOs, FAIR Handle generation, DO search

KIT Demo

- A PID is created according to a profile with machine actionable attributes, it is automatically validated, distributed and a search can be executed
- This is part of the KIT work for the Helmholtz Society
- next step to make use of DOIP exchange between different repositories and thus to unify. Validated and machine actionable (FAIR) PIDs are crucial.
- Participants: KIT, GWDG (PID Service, Type Registry Service)

KIT Demonstrator

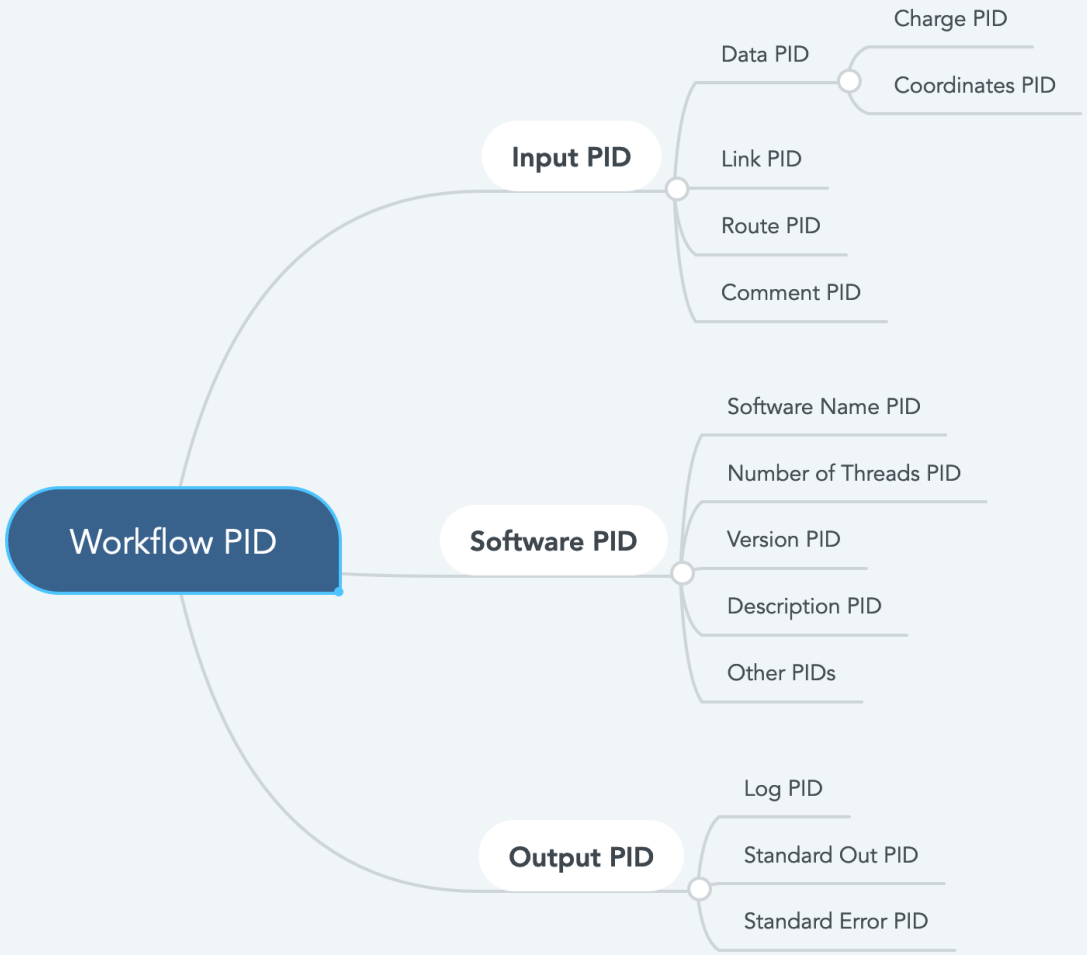


Demo Video Kit

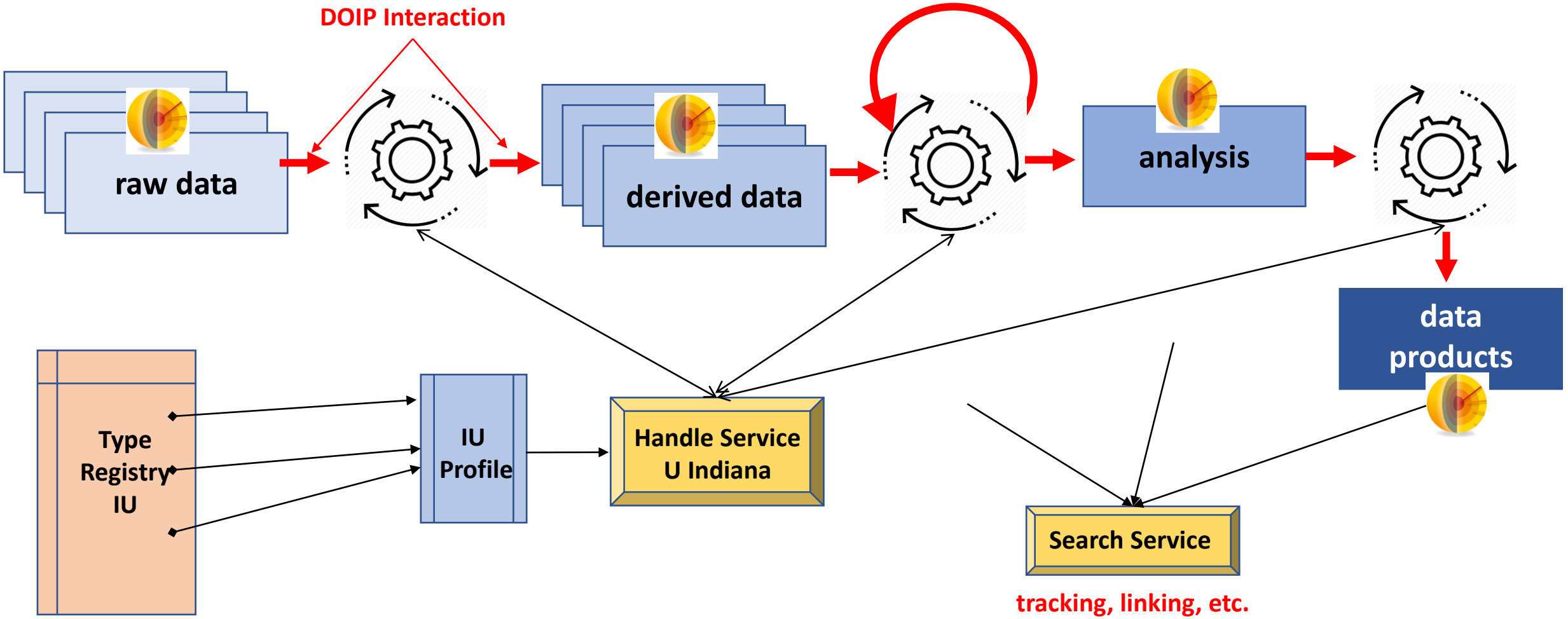
U Indiana Demo

- U Indiana is working on a variety of testbeds
- Workflow testbed based on a VRE supporting Handles and DOIP
  - Raw data –> data preparation SW -> Intermediate data -> data analysis -> data products
  - PIDs for workflow, data & software
  - at all stages PIDs are being assigned for tracing and provenance tracking
  - all stages well-documented

- Material Science Repository Integration



U Indiana Workflow Demonstrator

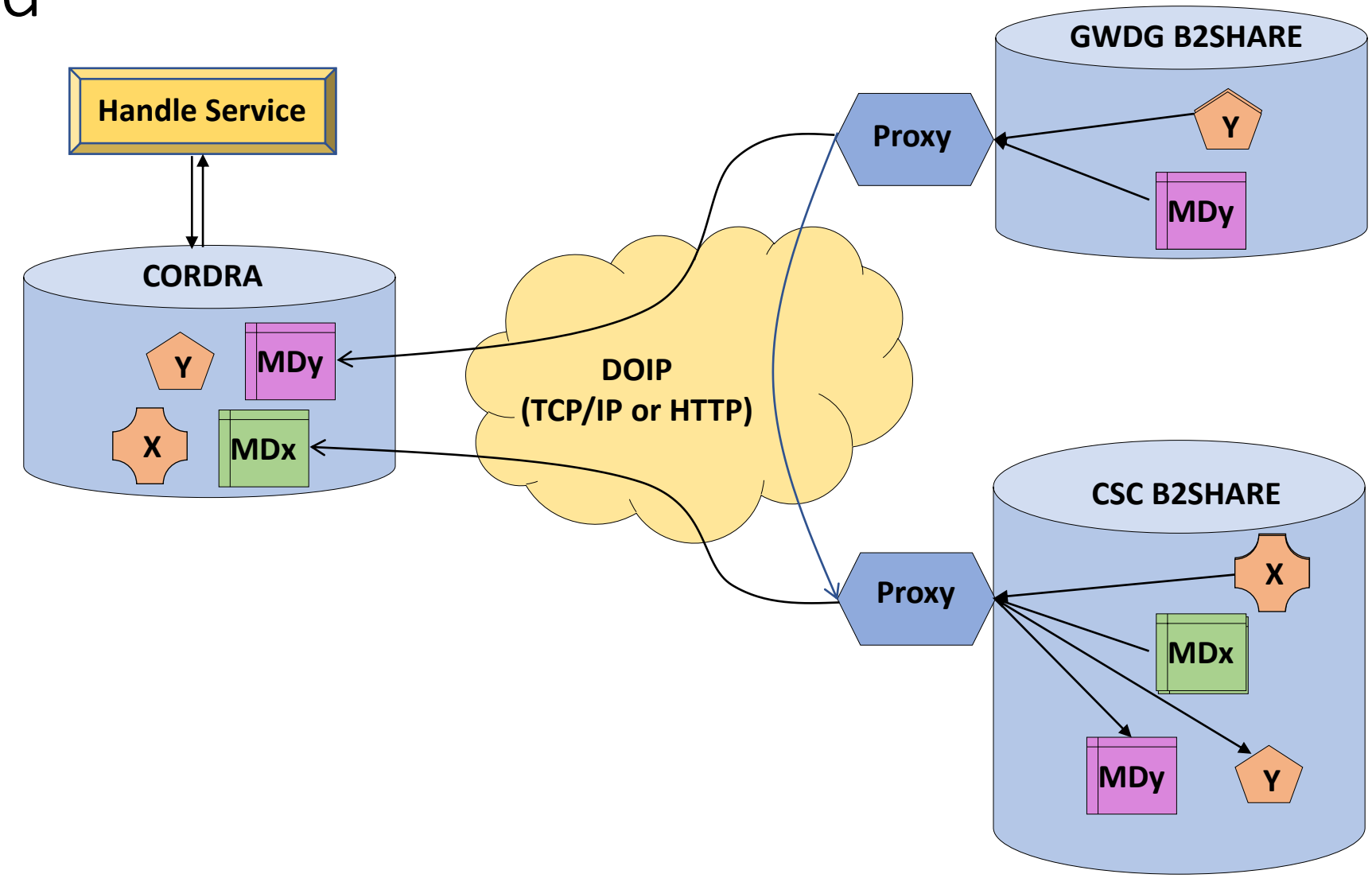


Demo Video U Indiana

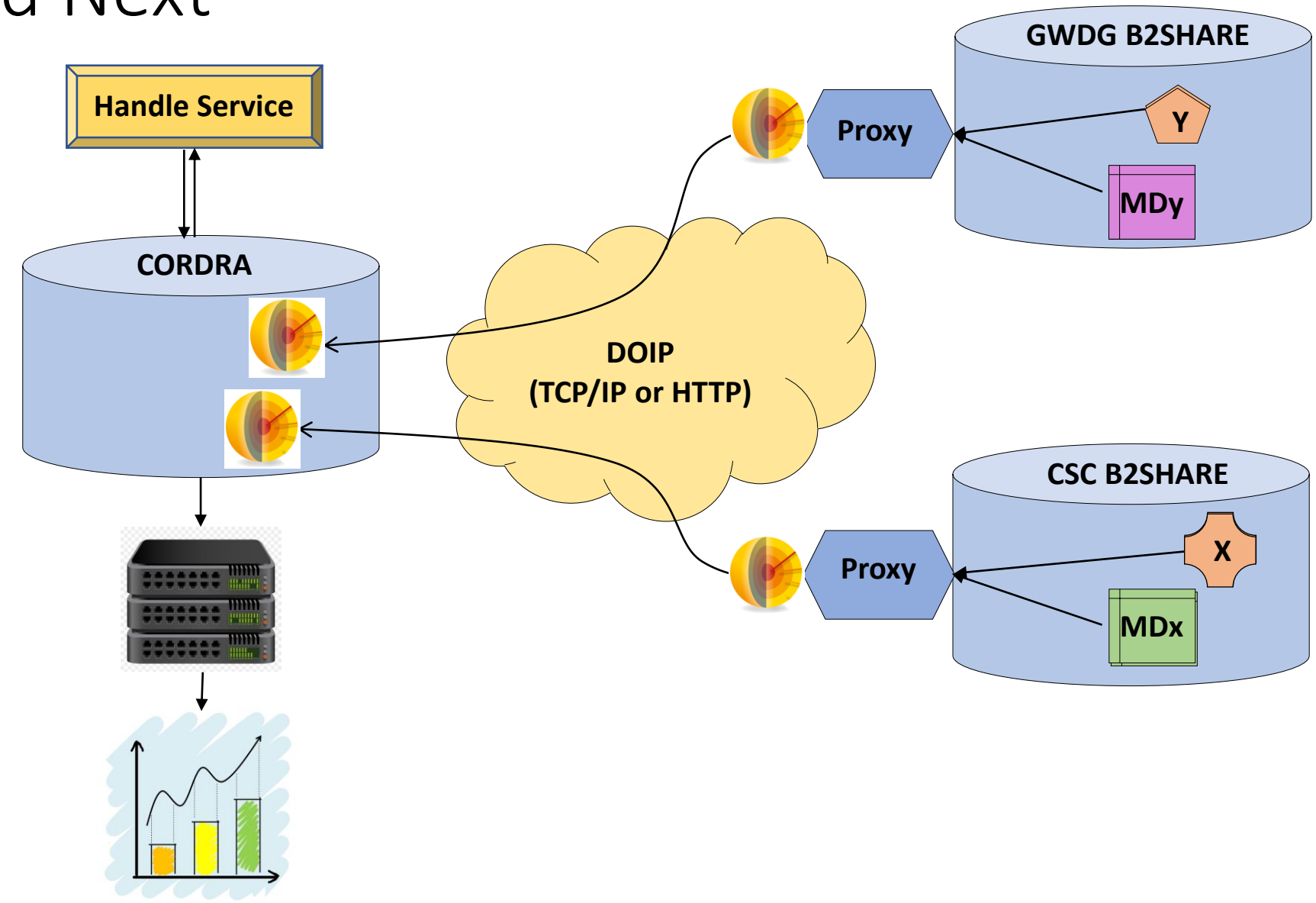
B2SHARE – CNRI Integration Testbed

- Intention is to integrate B2SHARE (EUDAT) with CNRI repository via DOIP
- CNRI developed a proxy (one week analysing and programing)
- Proxy allows to connect all B2SHARE instances (GWDG, CSC, FZ Jülich, U Helsinki, etc.)
  - usual operations: „copy“ FDO, copy bit-sequence, move FDO, delete bit-sequence
- next steps
  - nice Virtual Environment (GWDG) to be finished
  - transformation to full FDOs
  - connection of San Diego SC repository, automatic execution of docker image
  - integration of other repositories
- participants: CNRI, GWDG, CSC, SDSC

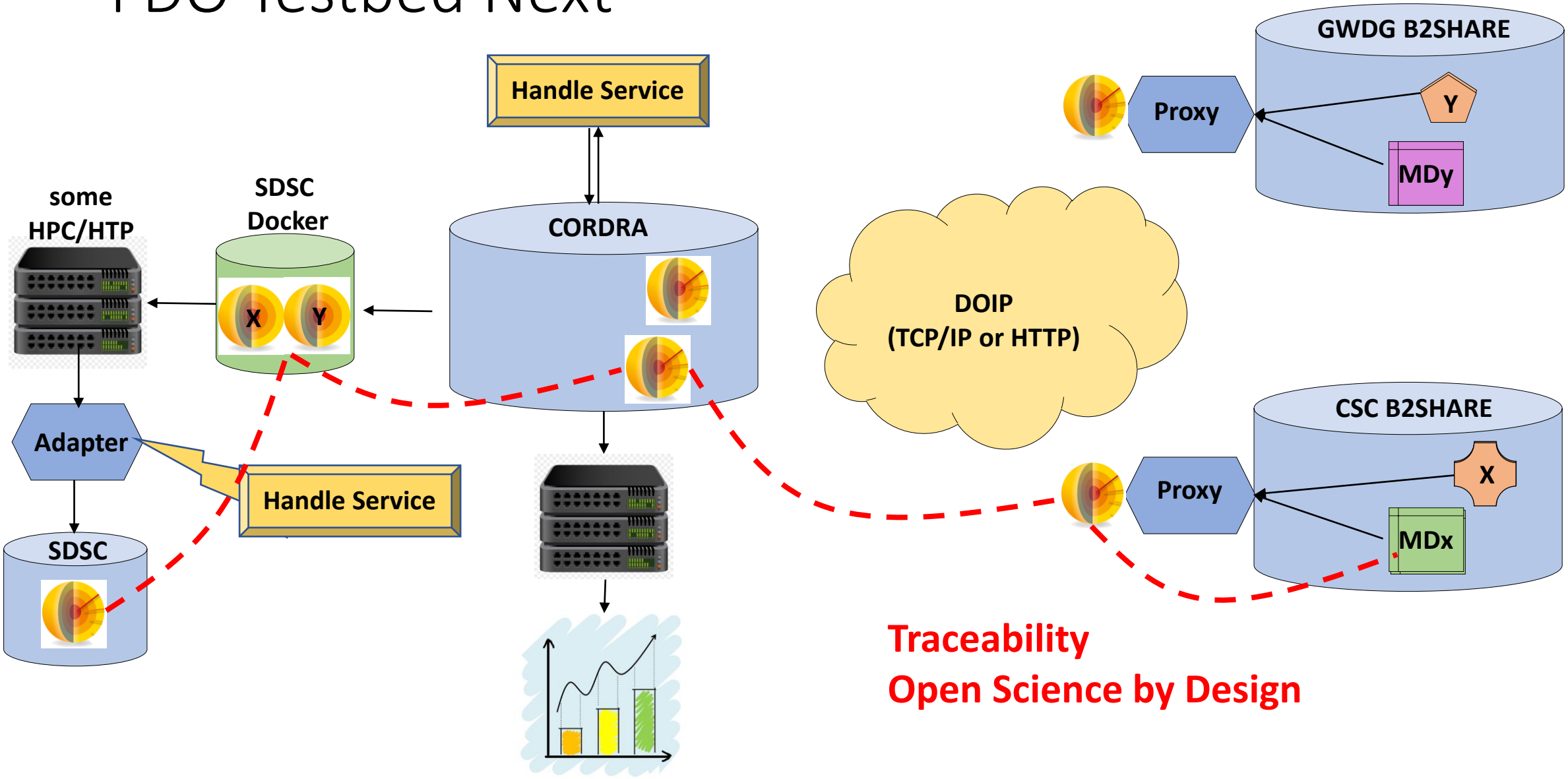
FDO Testbed


Handle Service
CORDRA
Y
MDy
X
MDx
DOIP
(TCP/IP or HTTP)
Proxy
GWDG B2SHARE
Y
MDy
Proxy
CSC B2SHARE
X
MDx
MDy
Y

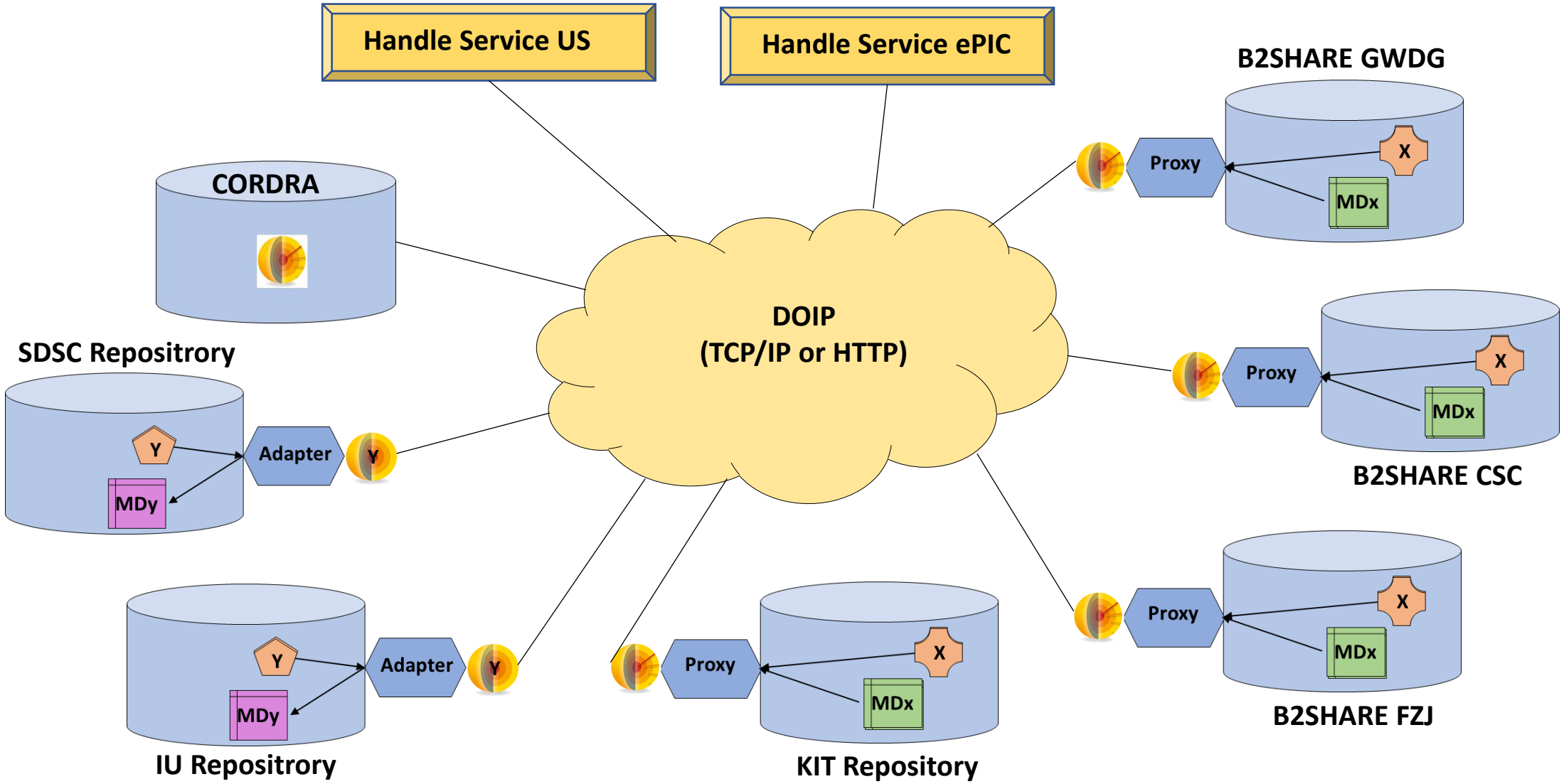
FDO Testbed Next



FDO Testbed Next



FDO Testbed Next



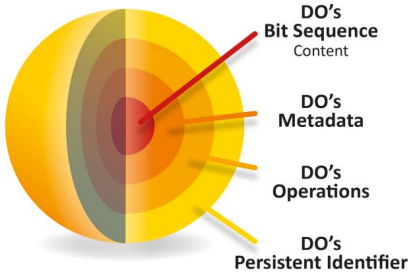
Demo Video FDO Testbed

Resumee

- a variety of piloting projects in addition to massive Handle/DOI usage (-> Maggie)
- with relatively small effort (1 week) one can create an FDO adapter/proxy (if developer is professional)
- all kinds of repositories (different data organisations & models) can be connected to DOIP network reducing complexity to N*1
- FDOs are FAIR compliant if community metadata is made FAIR compliant
- „encapsulation“ by specifying the FDO Type – Operation relationship

- **biggest problem for our demonstrator: access to developers**
- **urgent: connect these partial demonstrators together (incrementally growing testbed)**

FAIR DIGITAL OBJECTS FORUM
https://fairdo.org/



Thanks for the attention.