

Group of European Data Experts

Bringing People Together to Advance Data Science!

Large Research Infrastructure Building using FAIR Digital Objects

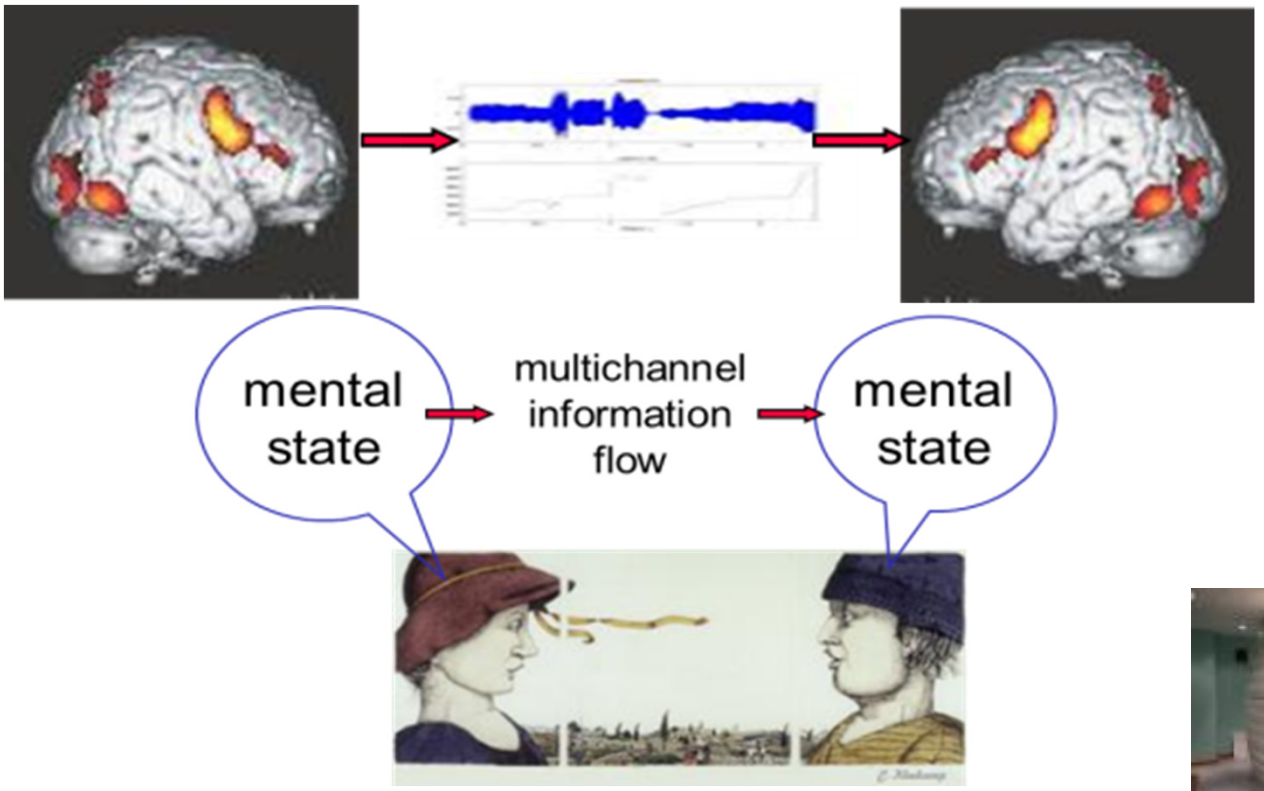
Münster Meeting

Peter Wittenburg

Max Planck Computing & Data Facility

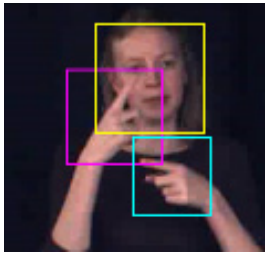


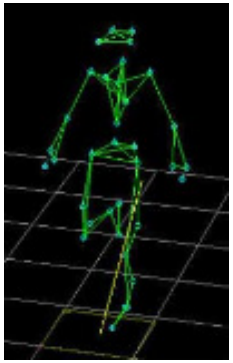
Psycholinguistics – Understanding Human Brain

















from regular to complex data
200 TB of data
80 TB in online archive

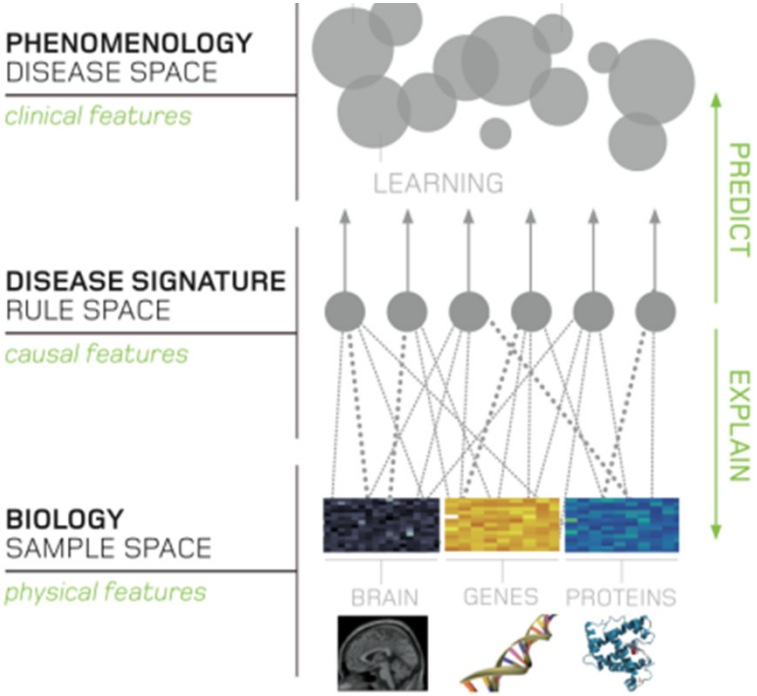
GEDE
Group of European Data Experts

MPCDF: Reoccuring Patterns in Data Science

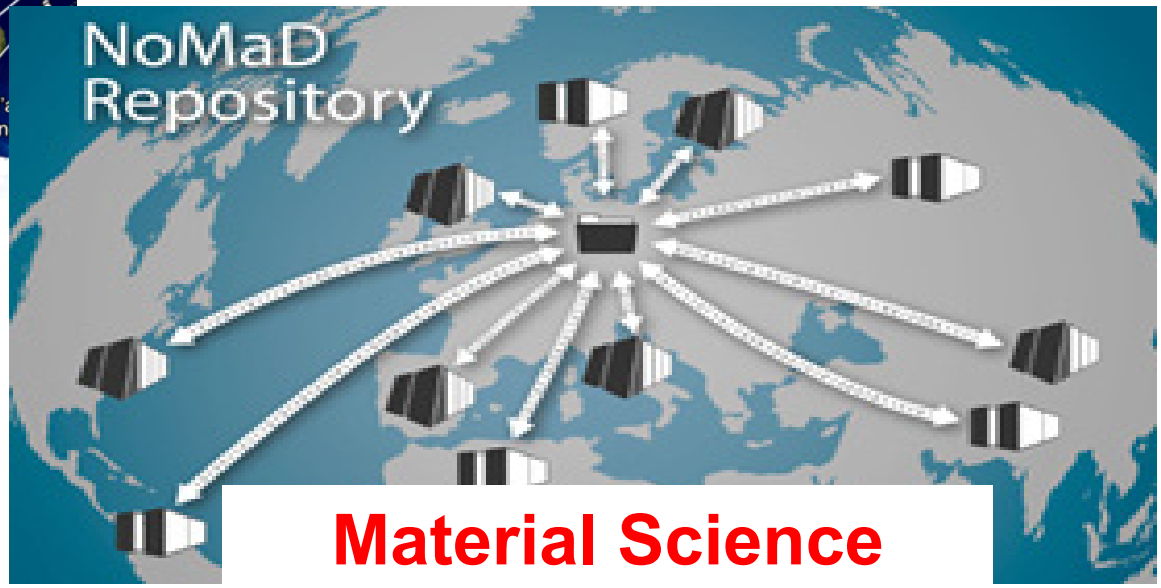
Humanities



Neuro Science



breaking with traditional paradigm:
- experiment -> publication
- forget the data



Material Science

- aggregating large „data lakes“
- fitting parameters of stochastic machines
- extract knowledge from data
- but ...

GEDE
Group of European Data Experts

Data Practices

- 80 % of work in data intensive science is lost with wrangling (science & industry) -> huge inefficiencies and costs
- 80 % of data is „dark data“ – disappearing after 20 y (Heidorn)
- 30 % of costs in healthcare due to non-FAIR data (NL: 30 bil €/y)
- 60 % of data projects in industry simply fail
- many researchers are excluded
- many projects simply not done

- fragmentation is huge even at data organisation layer
  - people don‘t know what they did months ago
  - when you have data you miss metadata and vice versa
  - etc.

GEDE
Group of European Data Experts

Data Practices

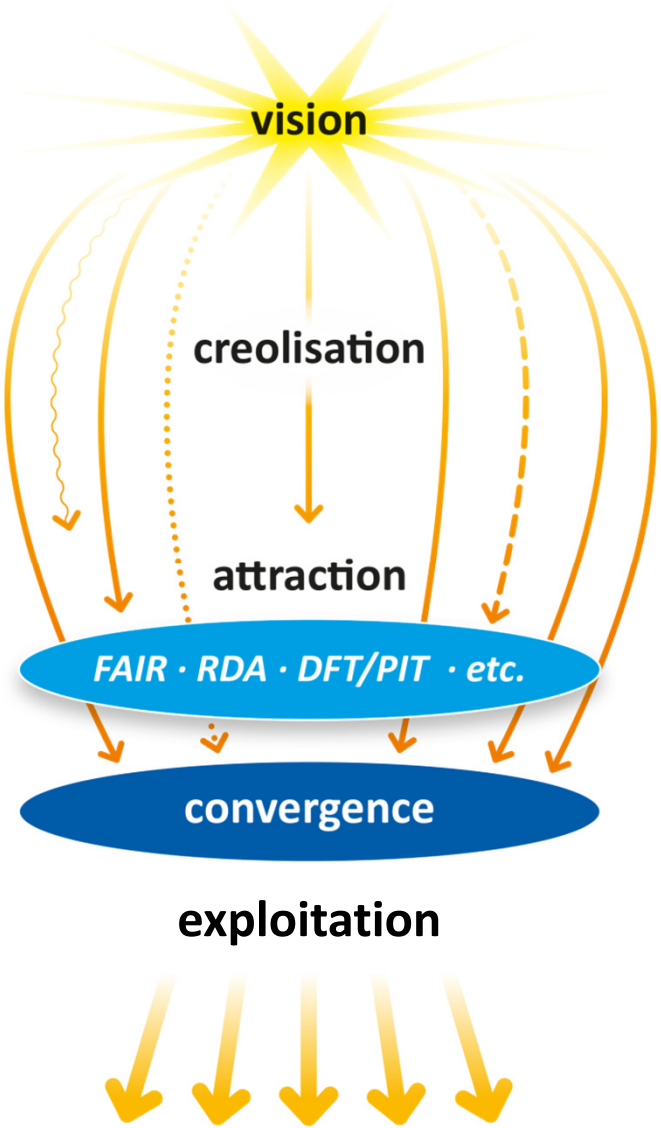
- 80 % of work in data intensive science is lost with wrangling (science & industry) -> huge inefficiencies and costs
- 80 % of data is „dark data“ – disappearing after 20 y (Heidorn)
- 30 % of costs in healthcare due to non-FAIR data (NL: 50 bil €/y)
- 60 % of data projects in industry simply fail
- many researchers are not data literate
- many projects simply not done

something must be fundamentally wrong with our data practices

- fragmentation is huge even at data organisation layer
  - people don‘t know what they did months ago
  - when you have data you miss metadata and vice versa
  - etc.

GEDE
Group of European Data Experts

Infrastructure Patterns



Taken from Wittenburg & Strawn
Common Patterns in Revolutionary Infrastructures and Data

Some attractors

Research Data Alliance (2013)

- Core Data Model (DO)
- PID Kernel Types (PID Attributes)
- Data Type Registry (DO types <> operations)
- Practical Policies (microprocedures)
- etc.

FAIR Principles (2014)

- RDA FAIR Maturity Indicator Group

DONA (2014)

- Handle System (DOI, ePIC, 3600 individual)
- DO Interface & PID Resolution Protocols

GEDE DO / C2CAMP Network (2015)

- 150 experts, 50 RI

Linked Data Platform (2012-15)

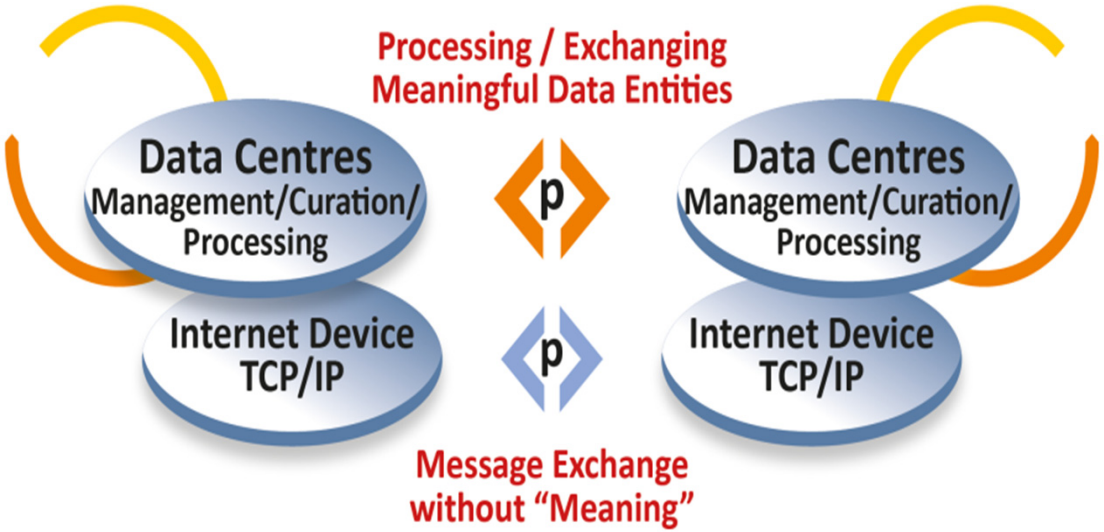
- HTTP, HTML, URIs

DO Model Development I

some applications

FTP
SMTP
GOPHER
etc.

82/85

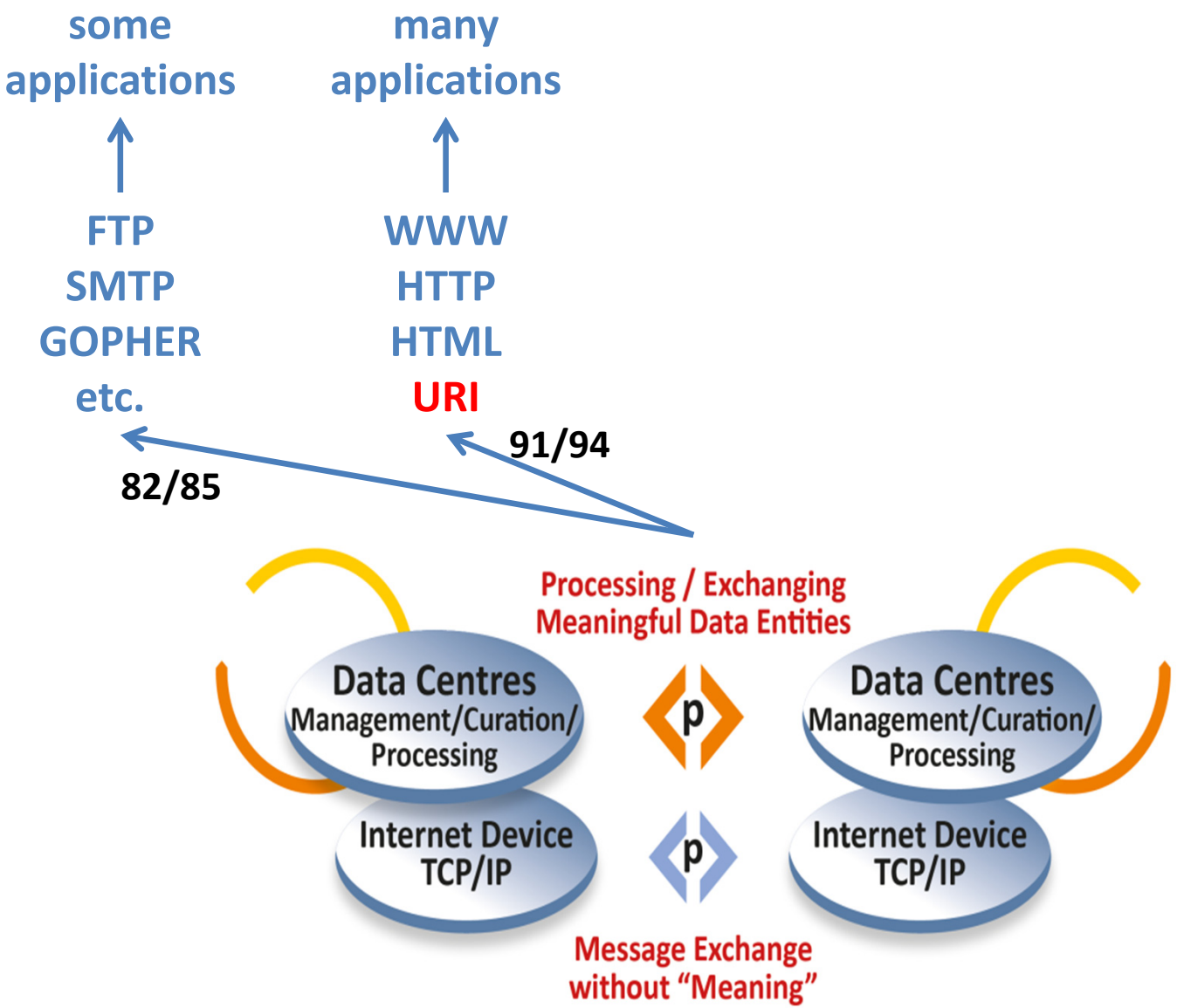


**complex,
many different types,
scientifically driven**

**„simple", few types,
technologically driven**

GEDE
Group of European Data Experts

DO Model Development II

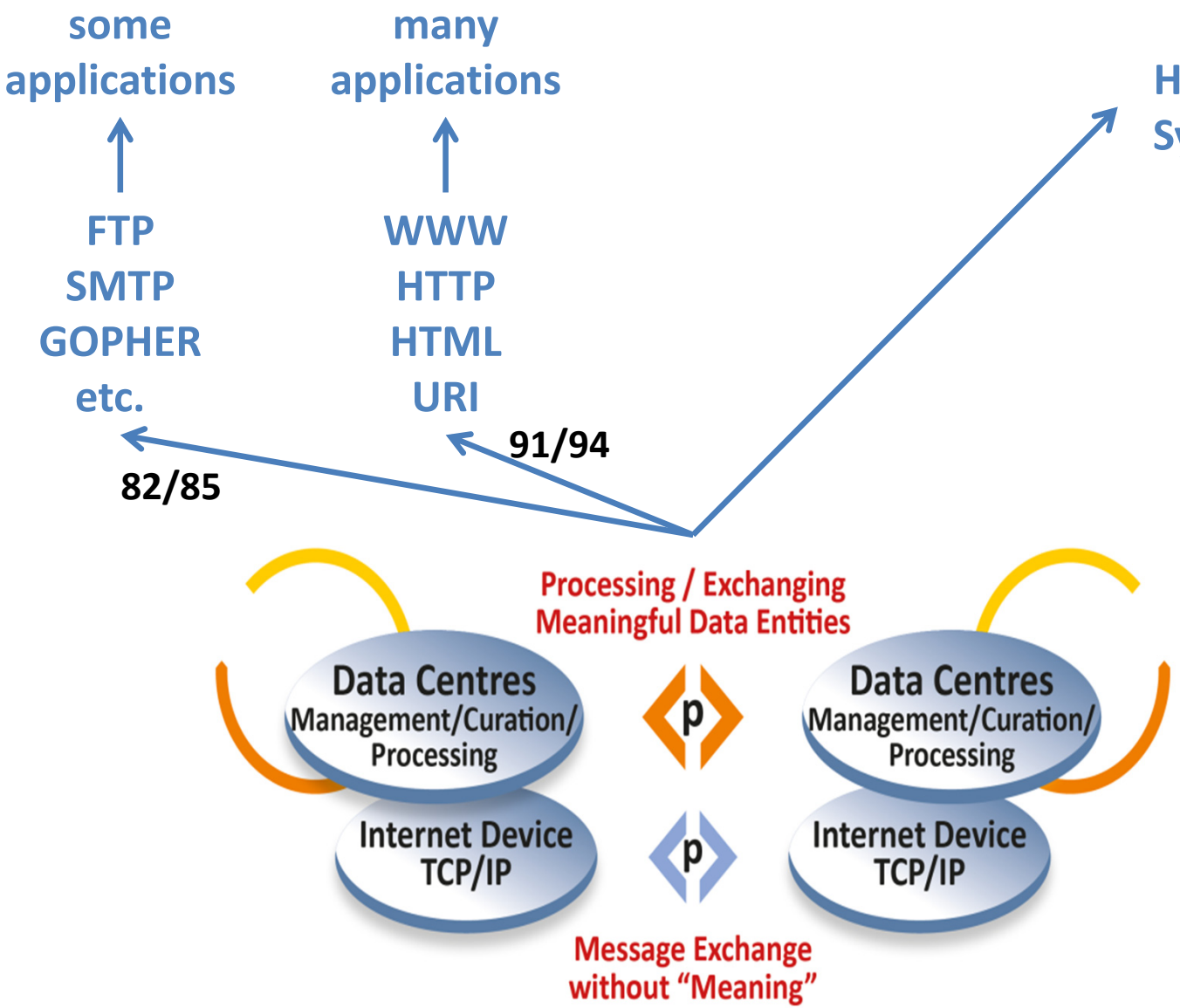


complex,
many different types,
scientifically driven

„simple", few types,
technologically driven

GEDE
Group of European Data Experts

DO Model Development III

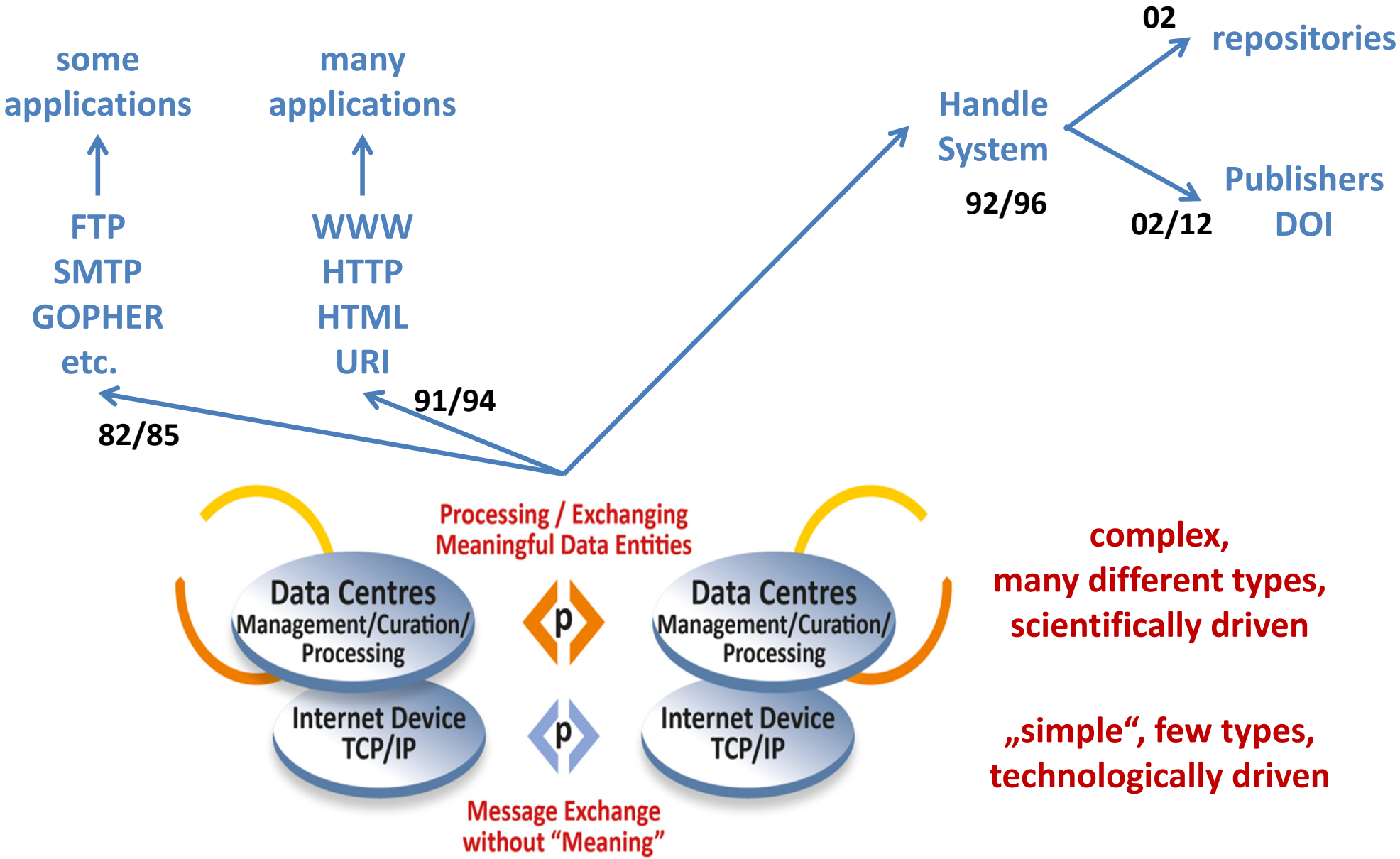


complex,
many different types,
scientifically driven

„simple", few types,
technologically driven

GEDE
Group of European Data Experts

DO Model Development IV

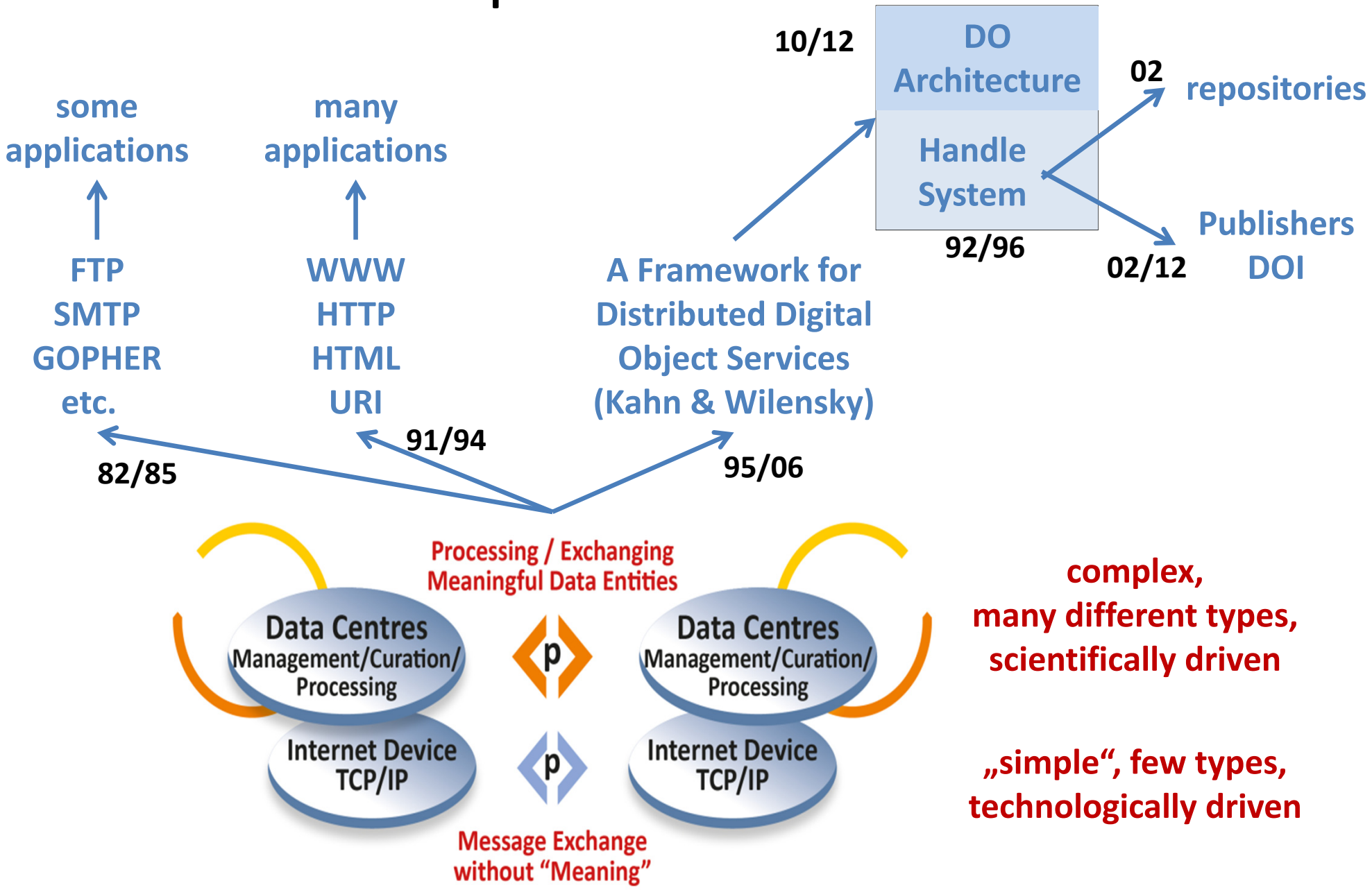


complex,
many different types,
scientifically driven

„simple", few types,
technologically driven

GEDE
Group of European Data Experts

DO Model Development V

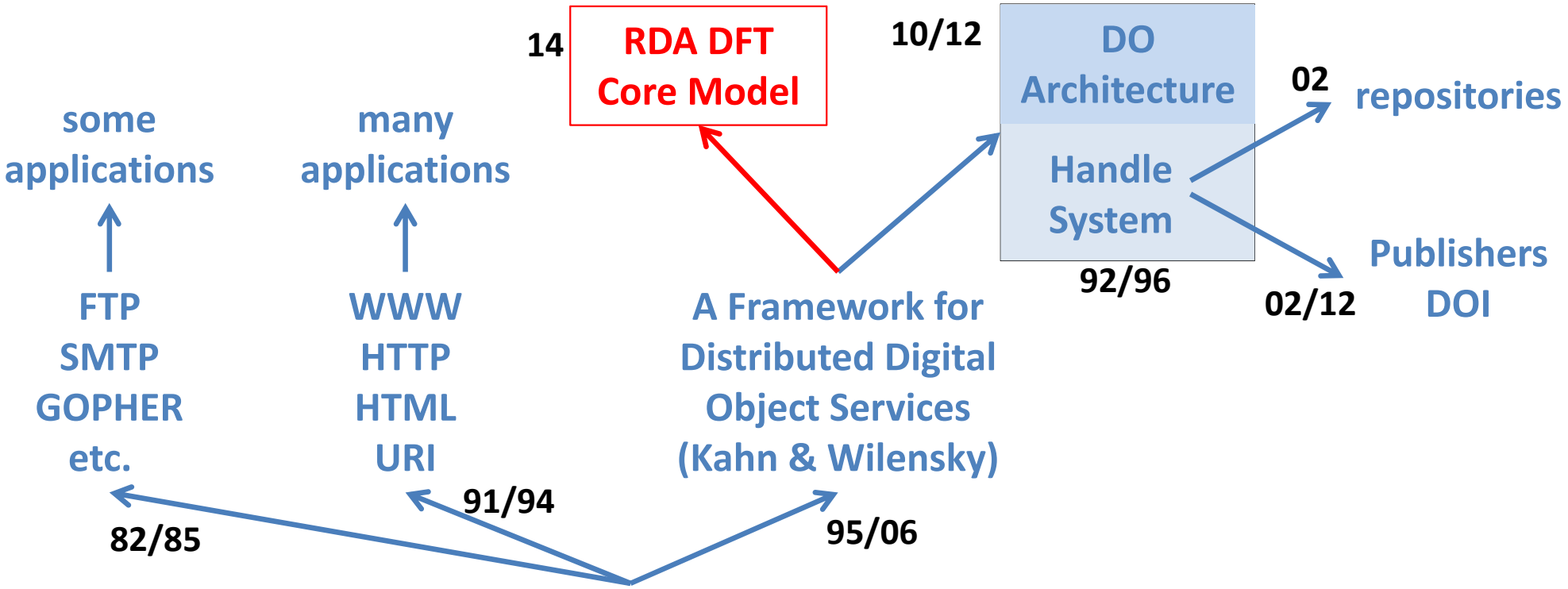


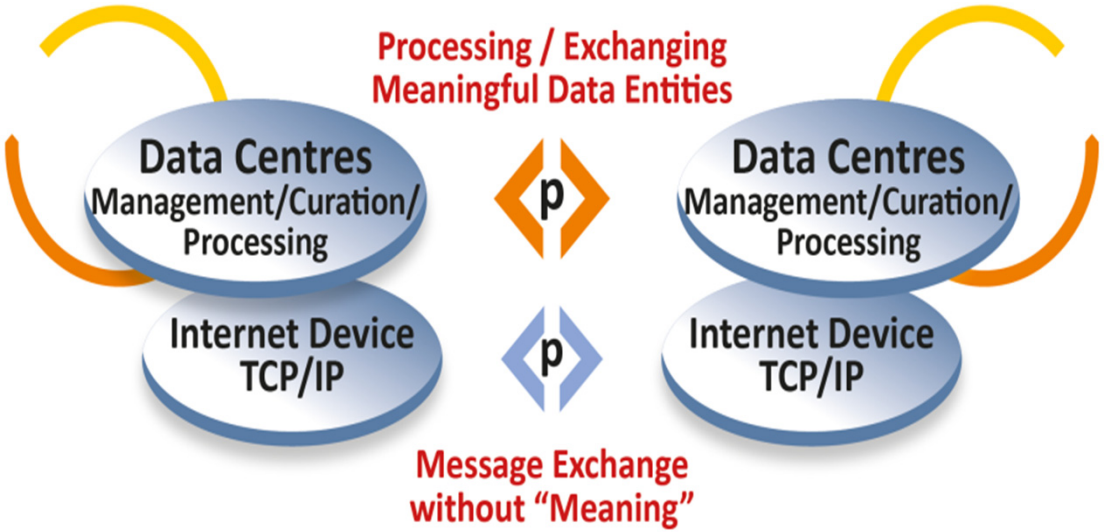
complex,
many different types,
scientifically driven

„simple", few types,
technologically driven

GEDE
Group of European Data Experts

DO - RDA came in





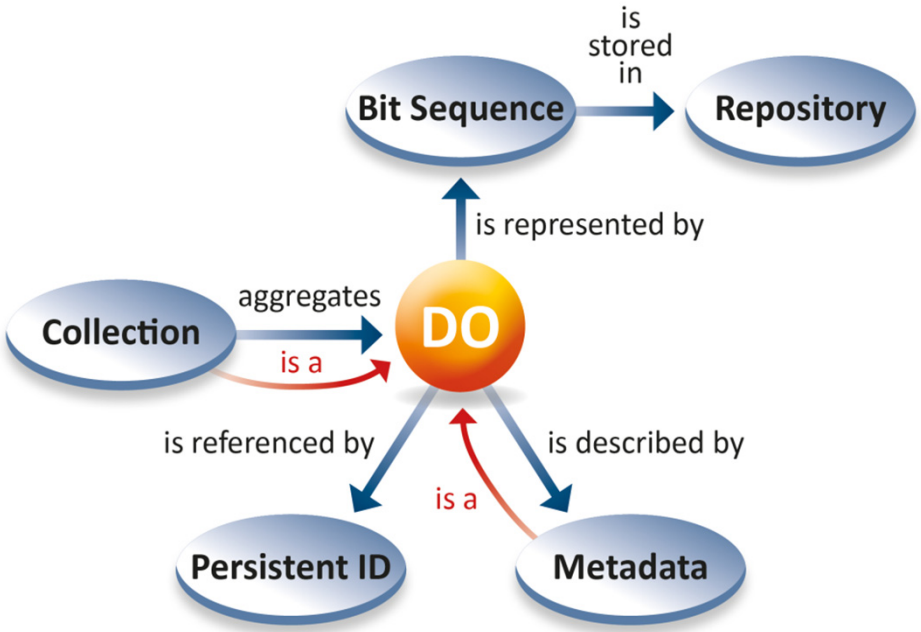
complex,
many different types,
scientifically driven

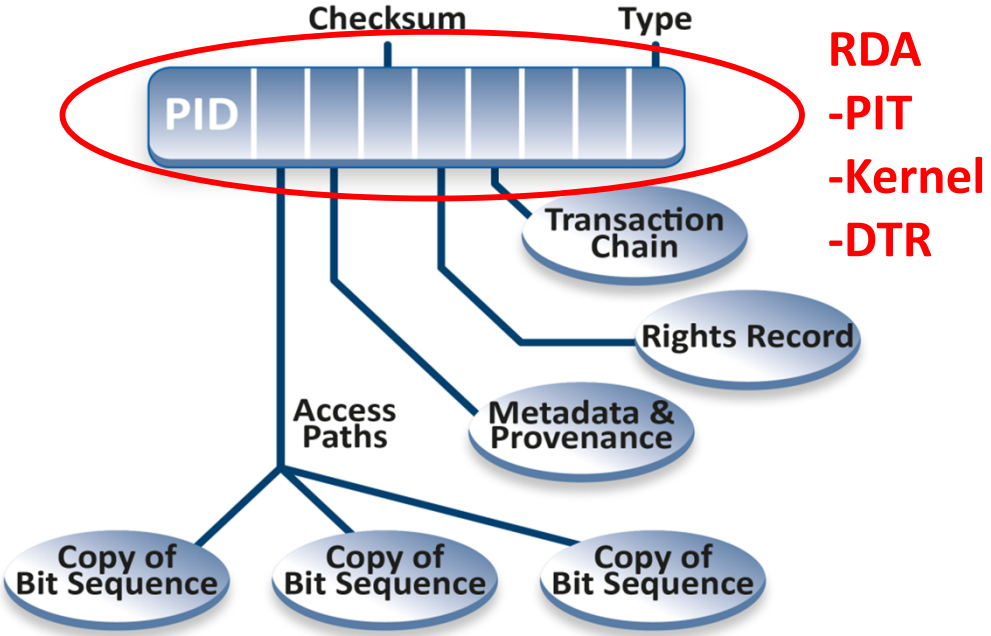
„simple", few types,
technologically driven

GEDE
Group of European Data Experts

DO: RDA Data Foundation & Terminology (2014)

Kahn & Wilensky: DO is an instance of an abstract data type that has two components, data and key-metadata. The data is typed. The key-metadata includes a handle.



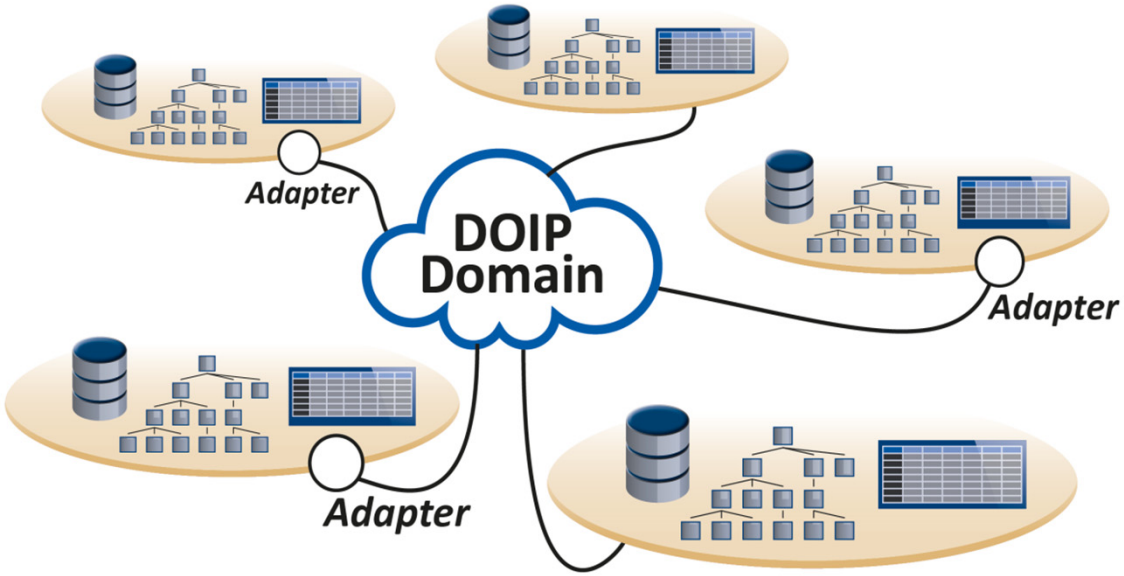


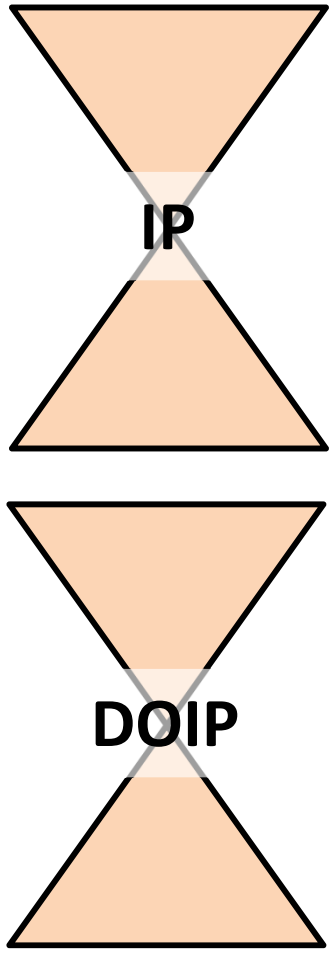
RDA DFT: a DO has a structured bit sequence stored in some repositories, is assigned a PID and is described by metadata. DOs can be aggregated to collections which are also DO. Metadata descriptions are DOs.
The DO‘s PID Record is resolved to machine-actionable attributes enabling human/machine actions.

GEDE
Group of European Data Experts

DOIP V2.0 from DONA

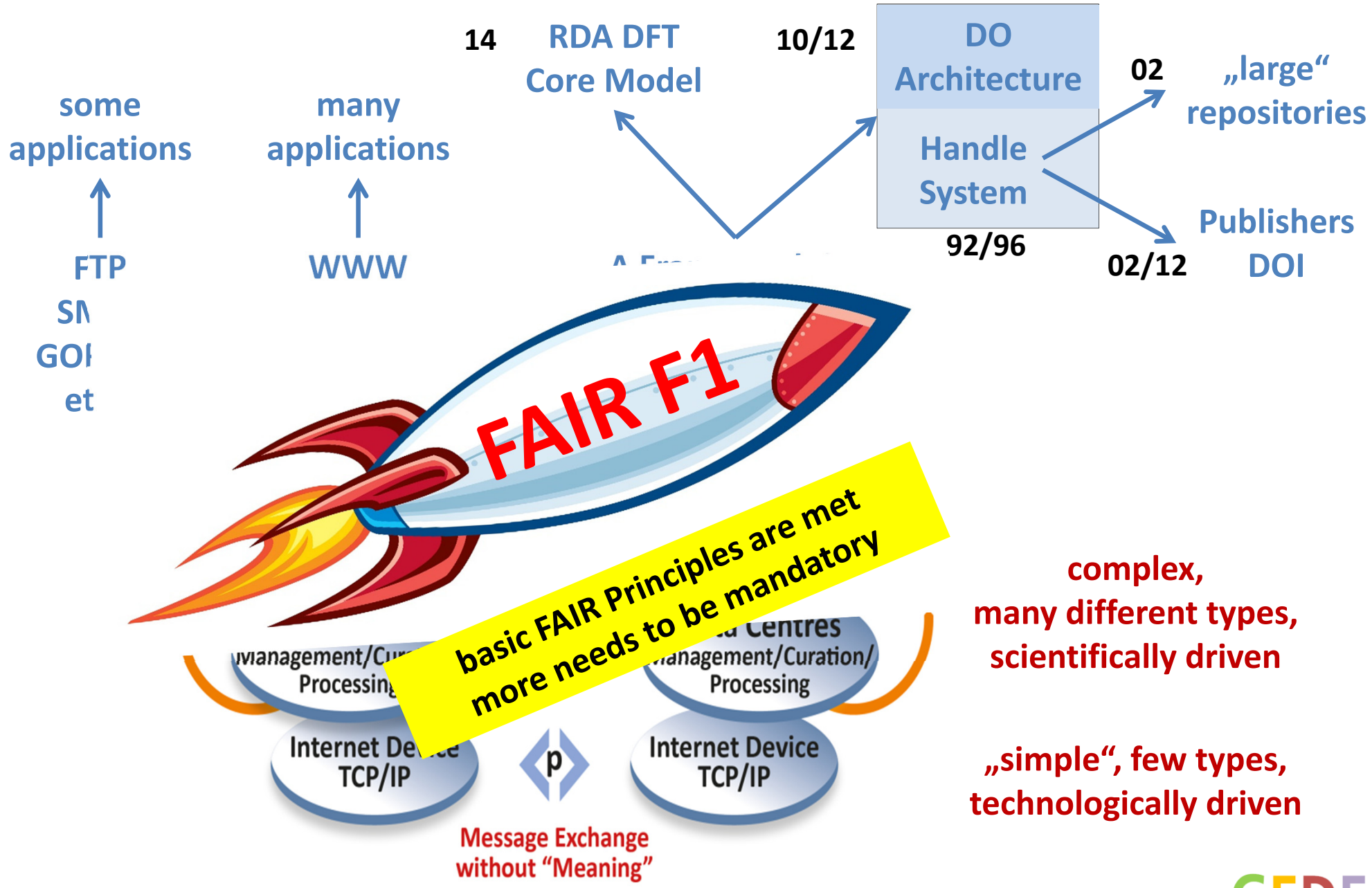
- improved specification and implementation of DO Architecture
- DOIP V2.0 specifying unified client – DO Server interaction
  - CORDRA reference implementation ready
  - DOIPV2.0 SDK almost ready
  - all based on PIDs





GEDE
Group of European Data Experts

Do DOs support FAIR?



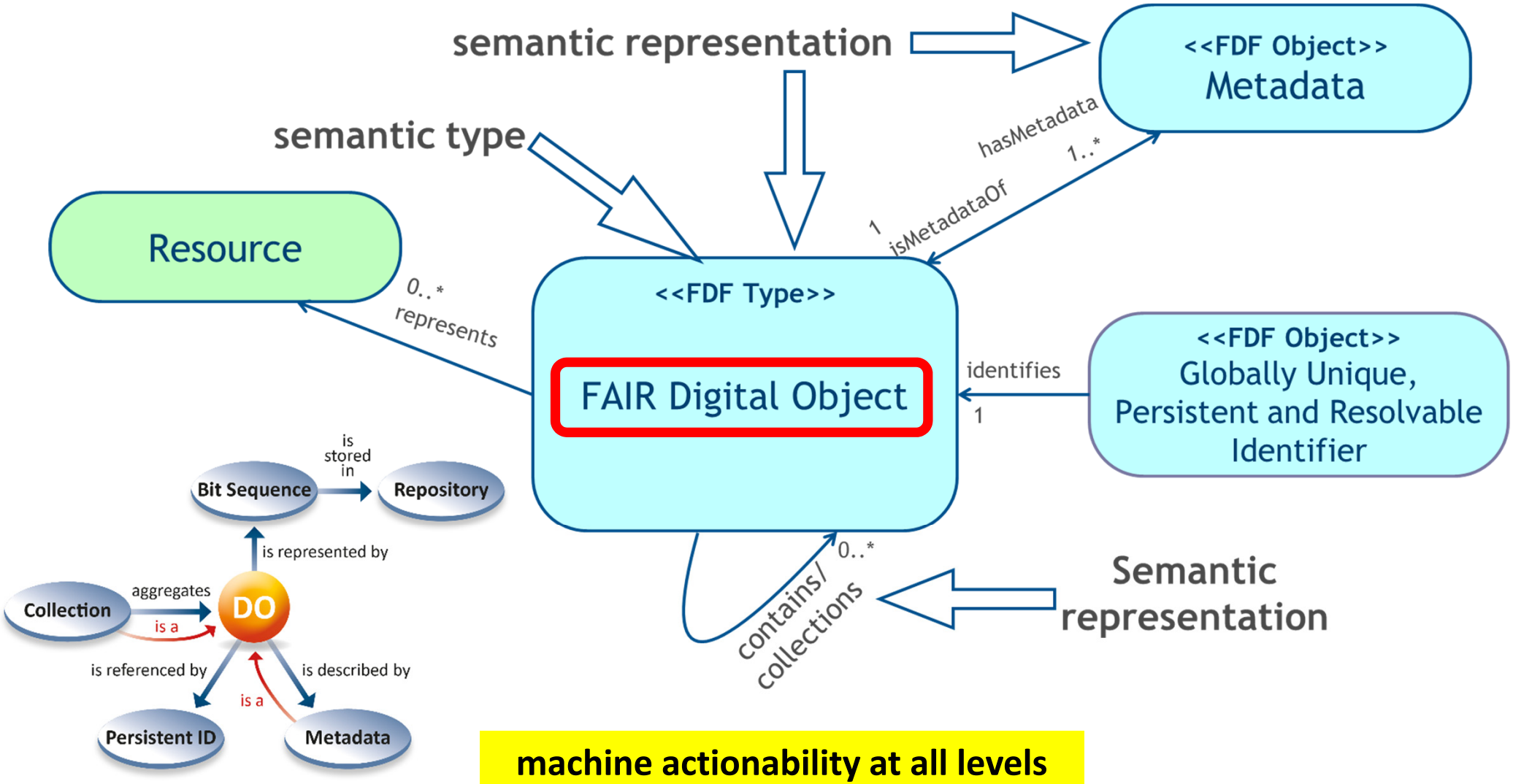
complex,
many different types,
scientifically driven

„simple“, few types,
technologically driven

GEDE
Group of European Data Experts

FAIR requires Semantic Explicitness

(in close collaboration with Luiz Bonino, applying mechanisms from LD)



machine actionability at all levels
what about metadata ???

GEDE
Group of European Data Experts

FAIR DO Framework (Version 1.01)

could be several implementations (DOA, LDP, DBMS, etc.)

General Guidelines

G1: Show a path for infrastructure investments for **many decades.**

G2: Demonstrate **trustworthiness** to researchers and developers to become engaged.

G3: Offer compliance with the **FAIR principles** being turned into indicators of FAIRness by an RDA Working Group (https://www.rd-alliance.org/groups/fair-data-maturity-model-wg).

G4: Support **machine actionability** which includes referential integrity, which states that all references need to be valid without temporal limitation, and explicitness of semantic relationships.

G5: Support the **abstraction principle**, i.e. abstract away from details that are not needed at a specific layer. At the management layer there is no difference to be made between data, metadata, software, semantic assertions, etc.

G6: Support **stable binding** between all informational entities that are required for machines to act.

G7: Support **encapsulation** which means that operations can be associated with types of FDOs.

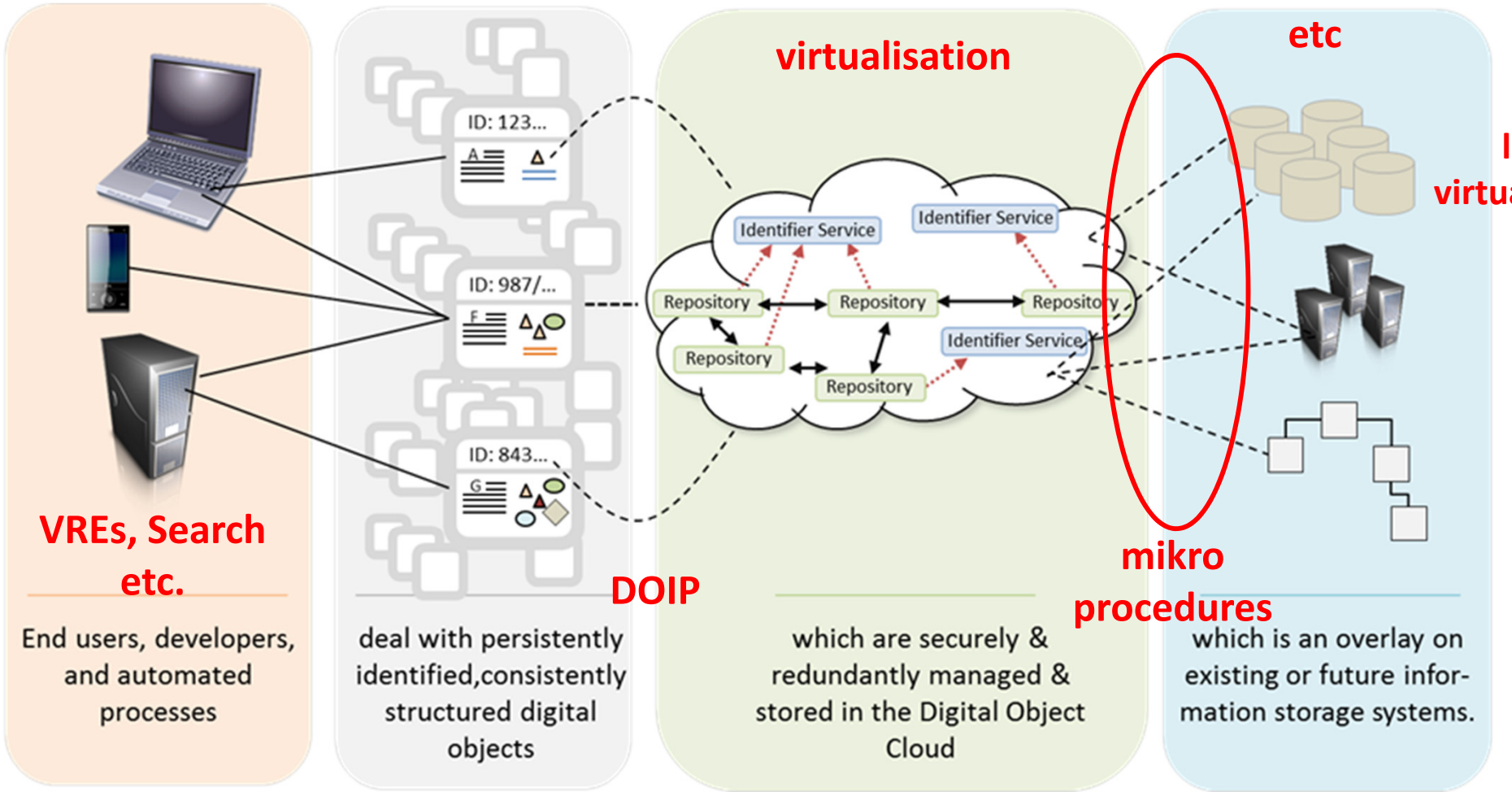
G8: Support **technology independence** allowing implementations using different technologies

FAIR DO Framework

FDOF1: A PID, standing for a globally unique, persistent and resolvable identifier, is assumed to be the basis of the Internet of FAIR Data and Services.
FDOF2: A PID is resolved to a structured record with attributes which are semantically defined within a type ontology which can have different forms.
FDOF3: The structured record includes at least a reference to the locations of the bit-sequences, a PID pointing to the metadata of FDO(s) and the DO's type.
FDOF4: The structured record can include other typed attributes that are important to characterize specific types of FDO or that are required by applications..
FDOF5: Each FDO identified by a PID can be accessed or operated on using an interface protocol by specifying the PID of a registered operation and the PID of the access point.
FDOF6: This protocol offers the typical CRUD operations on FDOs and a possibility to use extended operations.
FDOF7: The relations between FDO Types and operations are maintained in a type ontology.
FDOF8: Metadata descriptions being FDOs and describing the properties of the FDO are made available as semantic assertions enabling machines to act.
FDOF9: Metadata assertions can be of different types such as descriptive, deep scientific, provenance, system, access permissions, transactions, etc.
FDOF10: Metadata schemas are maintained by communities of practice. FDOF requires that such metadata are FAIR.
FDOF11: A collection of FDOs is an FDO and semantic assertions are to be used to describe their construction, i.e. the relationships of their constituents.

GEDE
Group of European Data Experts

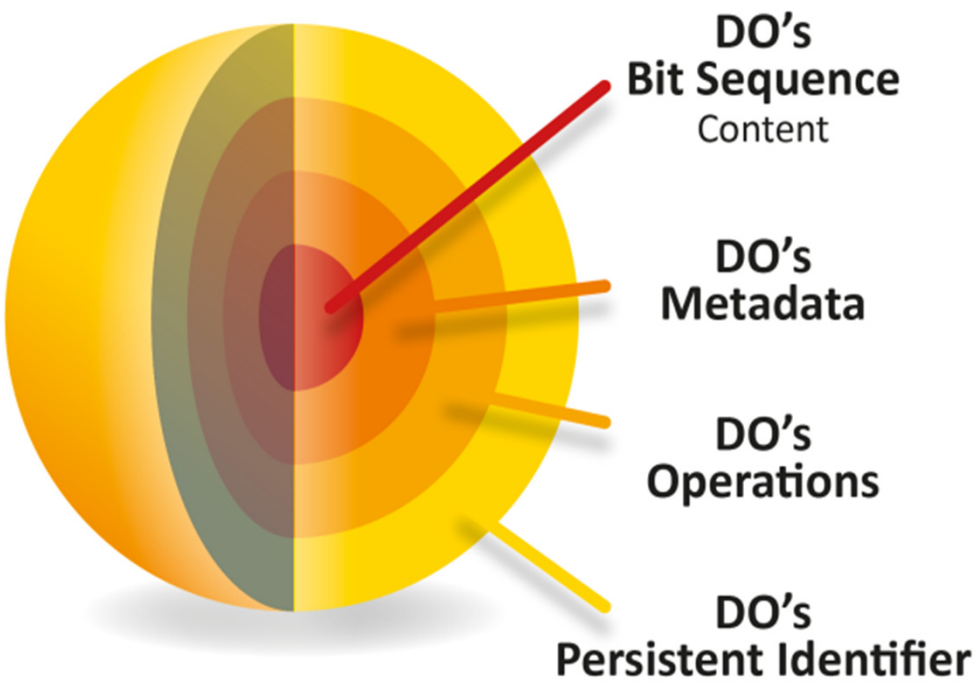
What does it solve?



Building now large infrastructures (EOSC, NFDI, etc.).
FDOs are an integrative technology for federative infrastructures!

GEDE Group of European Data Experts

FDO fit well



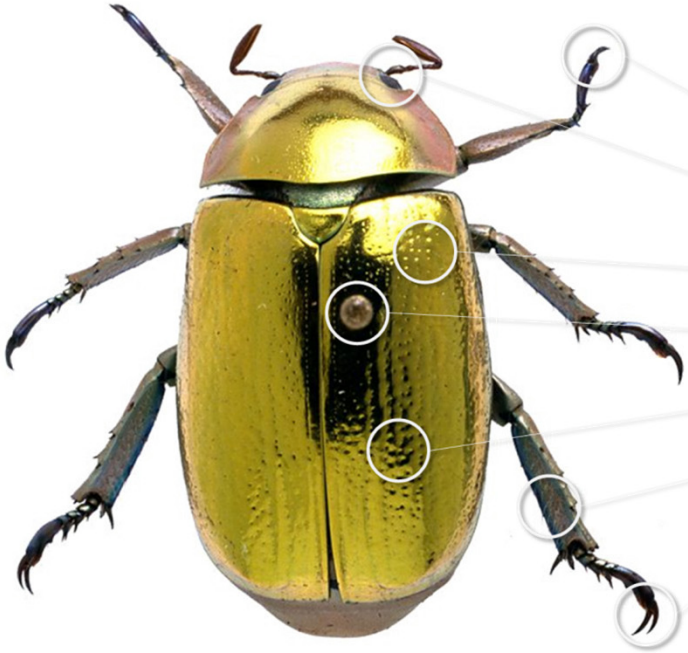
- very simple model
- all based on PIDs
- supporting
  - abstraction
  - stable binding
  - encapsulation

But ...

- do not address difficult semantic aspects
  (metadata semantics, scientific annotation, mapping, etc.)
- do not address operations on content
  (content transformation, knowledge extraction, etc.
- free of commercial influence (like TCP/IP)!!

GEDE
Group of European Data Experts

Biodiversity Use Case (Dimitris Koureas, Alex Hardisty)



- Genomic data
- Biochemical data
- Morphological data
- Geographical data
- Taxonomic Information
- Species Interactions data
- Ecological data

Natural Science Collections:

2 million standards

3 Billion objects

Trillions of relations

Digital Surrogate

FAIR Digital Object

=

an actionable

knowledge unit

Trust

Persistent link to the specimen supports validation of observations

Reproducibility

Persistent link to the specimen supports validation of observations



New insights

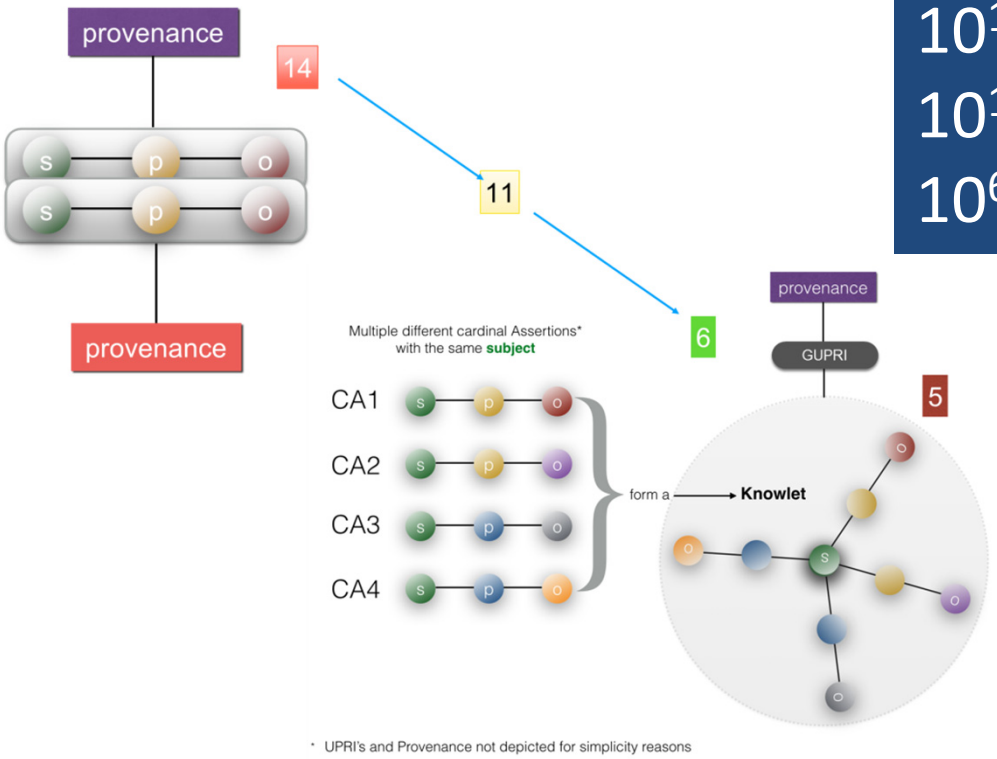
Robust linking of different data types

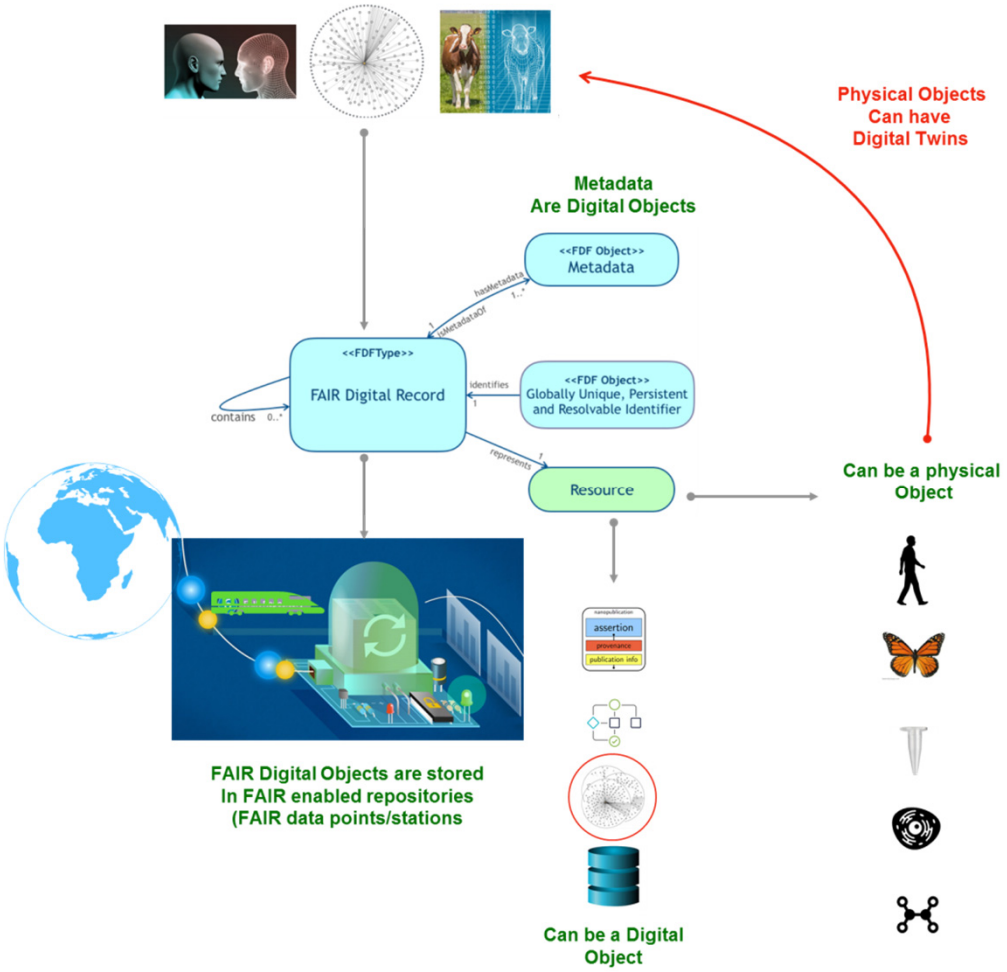
Expert Community

Community Annotations - Move the curation space from the analogue to the digital

Biomed Use Case (Barend Mons, Marco Roos)

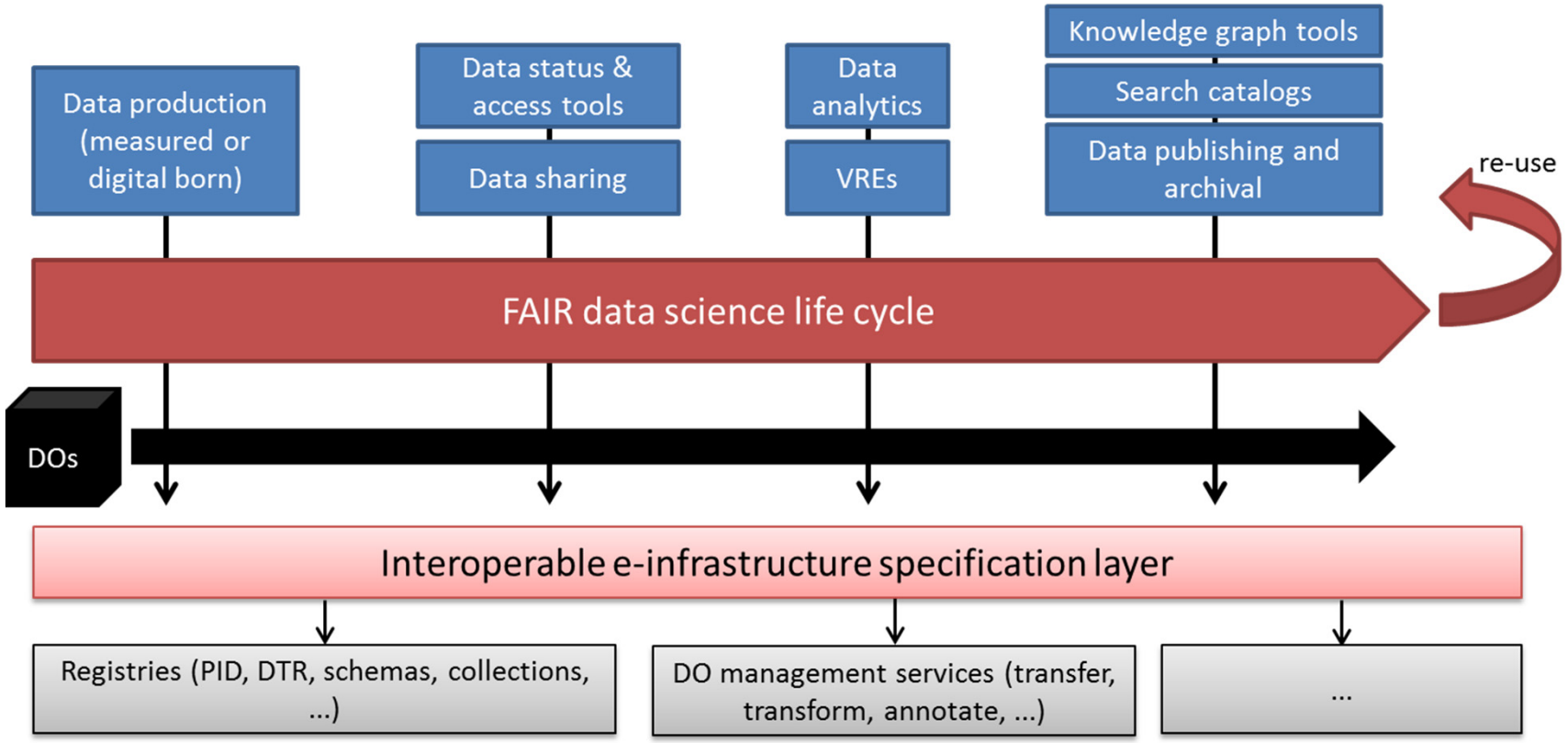
10^{14} nano-pubs (augmented RDF)
10^{11} cardinal assertions
10^{6} knowlets around key concepts





Knowlets & Digital Twins
=
FAIR Digital Object
Types

Climate Modeling (Tobias Weigel, et. al.)

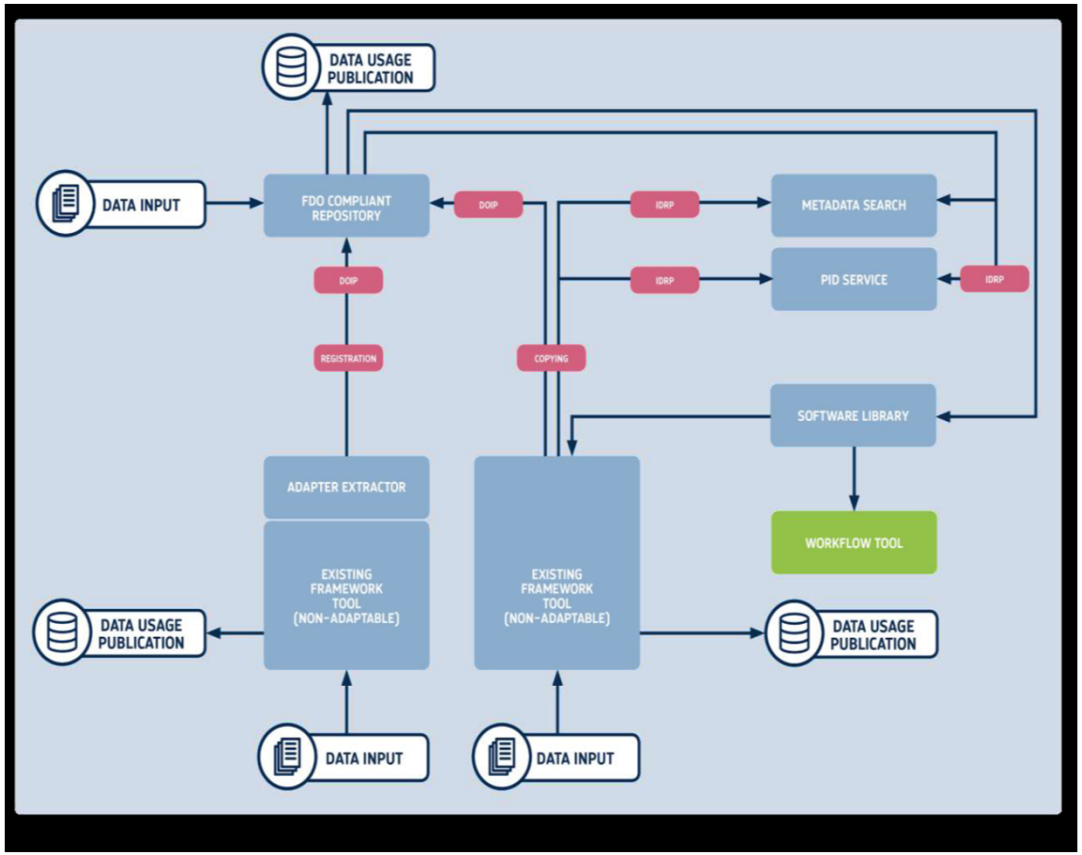


International Climate Modeling Commmunity

- only automatic management will work
- from the beginning FDOs through all life cycle states
- FDO to be supported by HPC processes

GEDE
Group of European Data Experts

Experimental Disciplines (Humans) (J. Weimann, et. al.)



Integrated Infrastructure for cross-disciplinary reuse

- 20 sub-disciplines plus use of OSF services
- 20+ repositories all different organisations, formats, metadata, etc.
- 20++ tools with some special functions
- how to mke this feasible – FDOs is a way to reduce complexity

GEDE
Group of European Data Experts

State

- EOSC is a must, FAIR is a must for EOSC – FAIR DO a choice
- FAIR DO is a federative infrastructure useful for EOSC

- Technically still much is missing
  - European system for PIDs ready to support DIS
  - some essential registries
  - repository adaptors
  - etc.

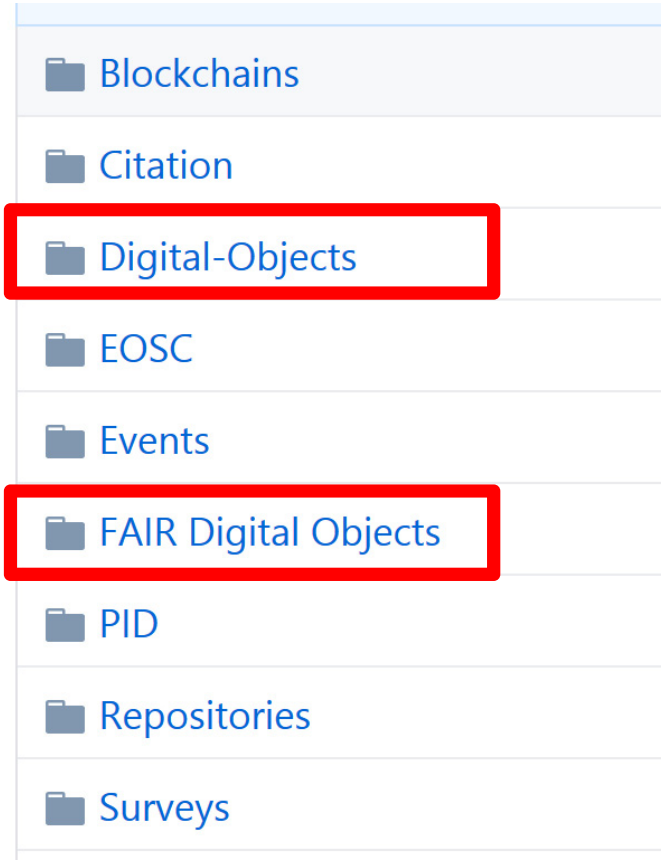
- Community:
  - ~ 400 experts (GEDE-DO/C2CAMP, DONA, GOFAIR IN, RDA DF, US, CAS)
  - some RIs adopt the FDO concept, some projects testing FDOs now
  - many community actions

- Governance
  - Coordination Group working out a governance structure
  - Technical Implementation Group (as open as possible)
  - pushing work through RDA IG/WG

Where?

- Google: GEDE Github → DO and/or FDO
- https://github.com/GEDE-RDA-Europe/GEDE

Thanks for the attention.



GEDE Group of European Data Experts