FAIR Digital Objects

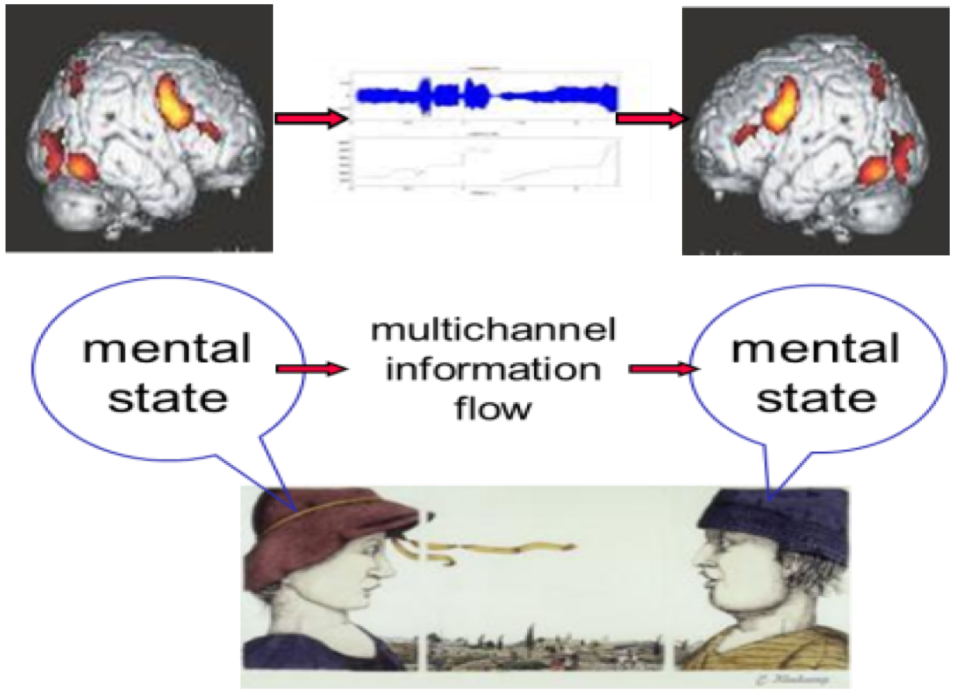
Self-standing Entities for Organising the Global Digital Data Space

Peter Wittenburg

FDO Forum

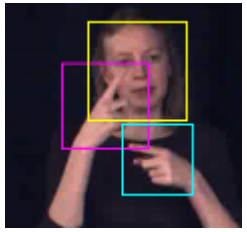
Who am I?

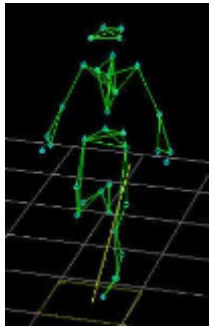
Max Planck Institute for Psycholinguistics
What happens in the brain during language Processing?



Leading Roles: DOBES, CLARIN, EUDAT
Co-Founder of RDA, GEDE, FDO Forum
Co-Author: FAIR Principles
Co-Author: Turning FAIR into Practices

From its beginning data driven research using all kinds of measurements and observations





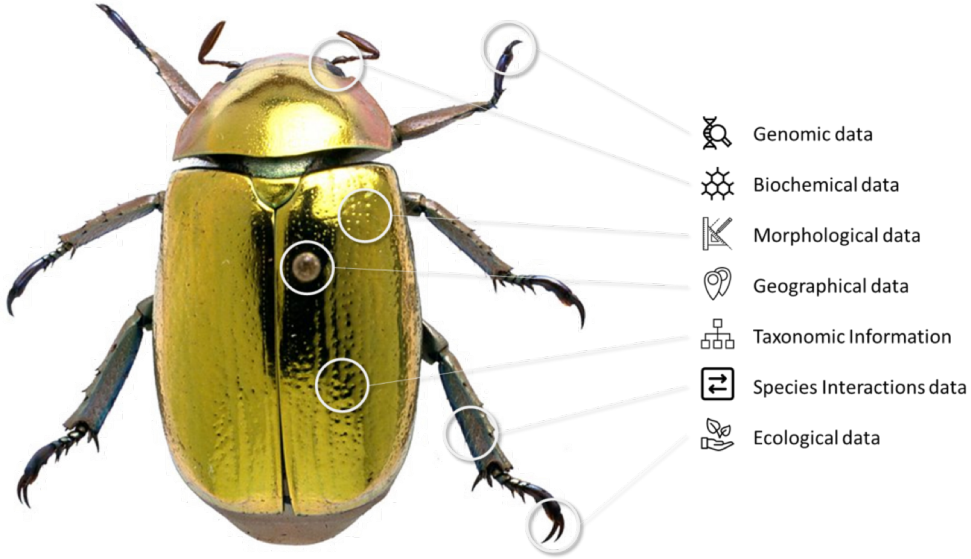


Why do we care about Data?

- Observations are the basis for research – not at all new but trends
- J. Gray: Data-Science as 4th paradigm
  - Thousand years ago: science was empirical describing natural phenomena
  - Last few hundred years: theoretical branch using models, generalizations
  - Last few decades: a computational branch simulating complex phenomena
  - Today: data exploration (*without apriori bias*)
- Y. Harari in Homo Deus: become Dataists relying on data analytics
  - is data telling us truth – no, selections and algorithms are determining
  - MPI study: language development with phylogenetic algorithms lead to multiple dependency trees
- L. Floridi in 4th Revolution: Finding small patterns in big (multimodal) data
  - many small patterns – not all are relevant

Changing: massiveness/ubiquity and complexity

Data Types in Biodiversity (Dimitris Koureas, Alex Hardisty)



Genomic
Biochemical
Morphological
Geographical
Taxonomic
Species interaction
Ecological

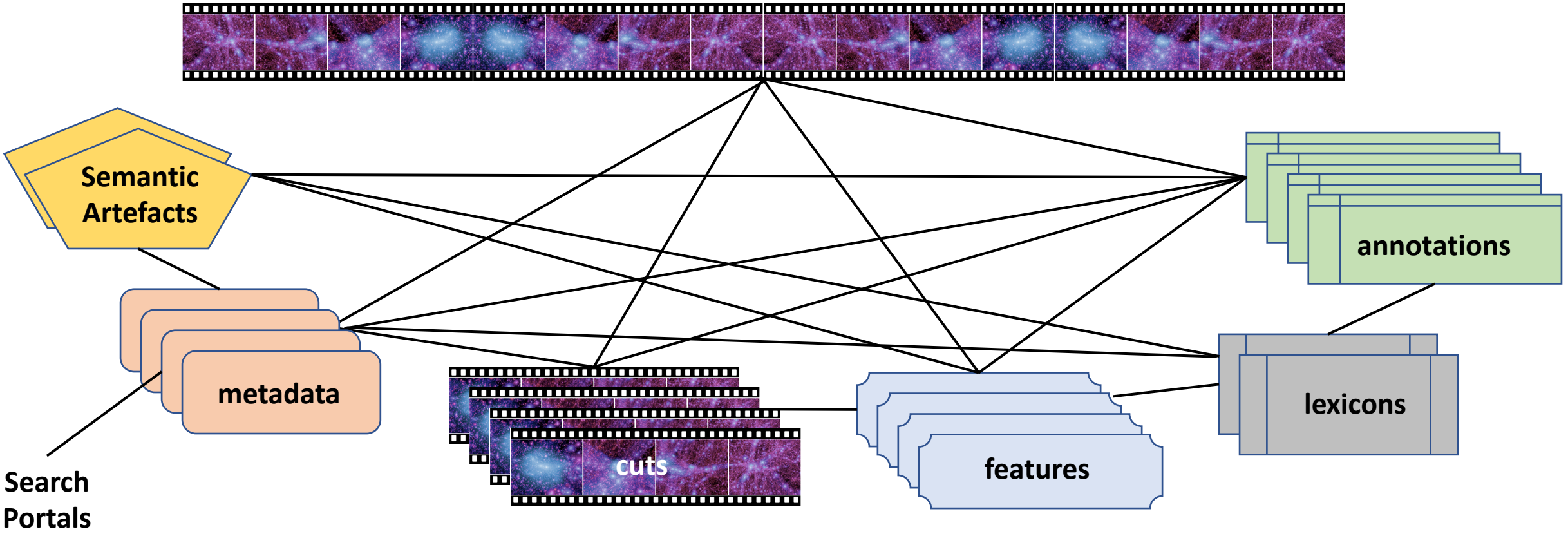
data in different specialised repositories

Natural Science Collections:

1 thousand collections
2 million standards
3 Billion objects
Trillions of relations

How to maintain a stable and reusable domain of data over decades?

Data Types in Language Research (DOBES)



A variety of **related** objects normally in different versions, from different researchers and often using different technologies.

How to maintain a stable and reusable domain of data over decades?

Reality

- Surveys: Researchers spend 75-80 % of their time on data wrangling (or have a whole team of data assistents)
- Individual researchers have so much data, derived data etc. on their notebooks/servers that they soon lose control
- When a PhD left no one knows about his/her data
- Deep analysis of about 75 research infrastructure projects (2020):
  - Most researchers heard about FAIR and promote it on paper.
  - But practices did not really change in last 5 years.
  - FAIRness is shifted to next colleagues in the “production” chain.
  - Researchers prefer Open Science by Publication instead of OS by Design but ...

Data use is still very much „local“

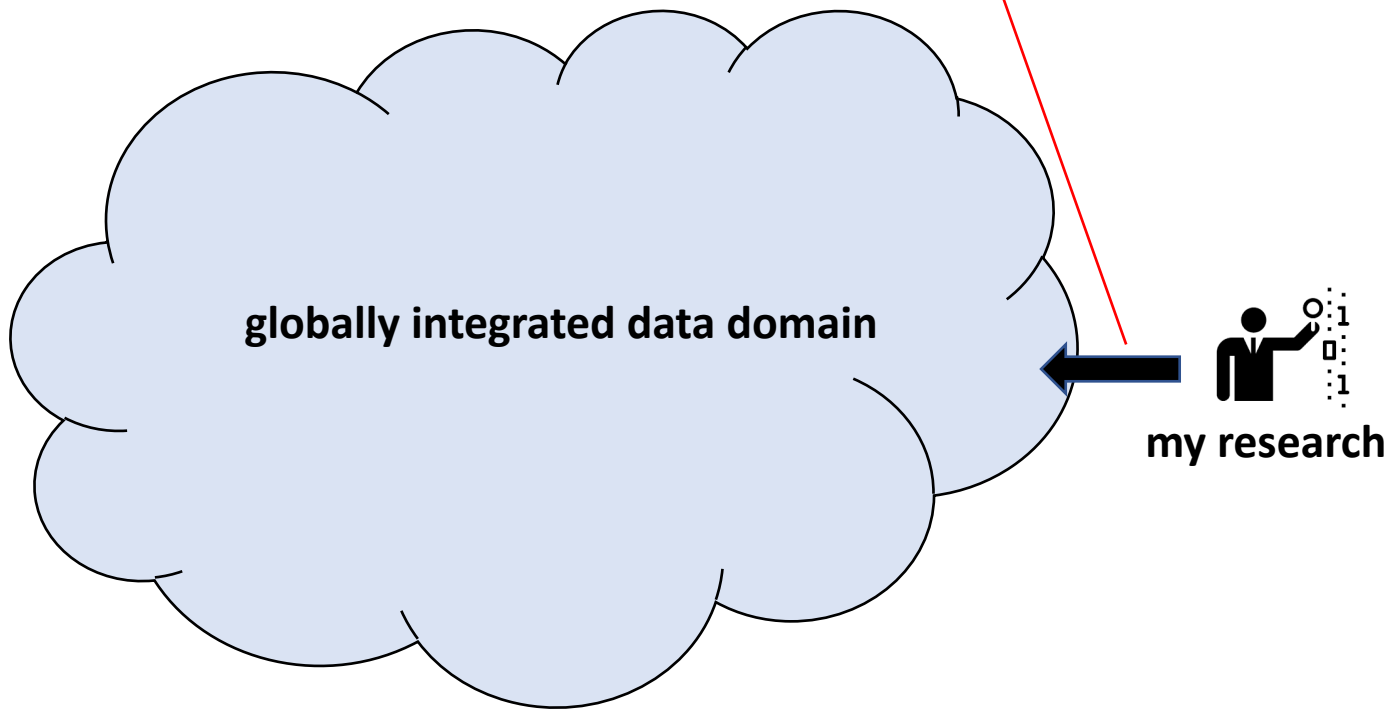
- know the data & context
- know its structure & variables
- have probably some tools
- have trust

- exchange via publications

Will/Should it change?

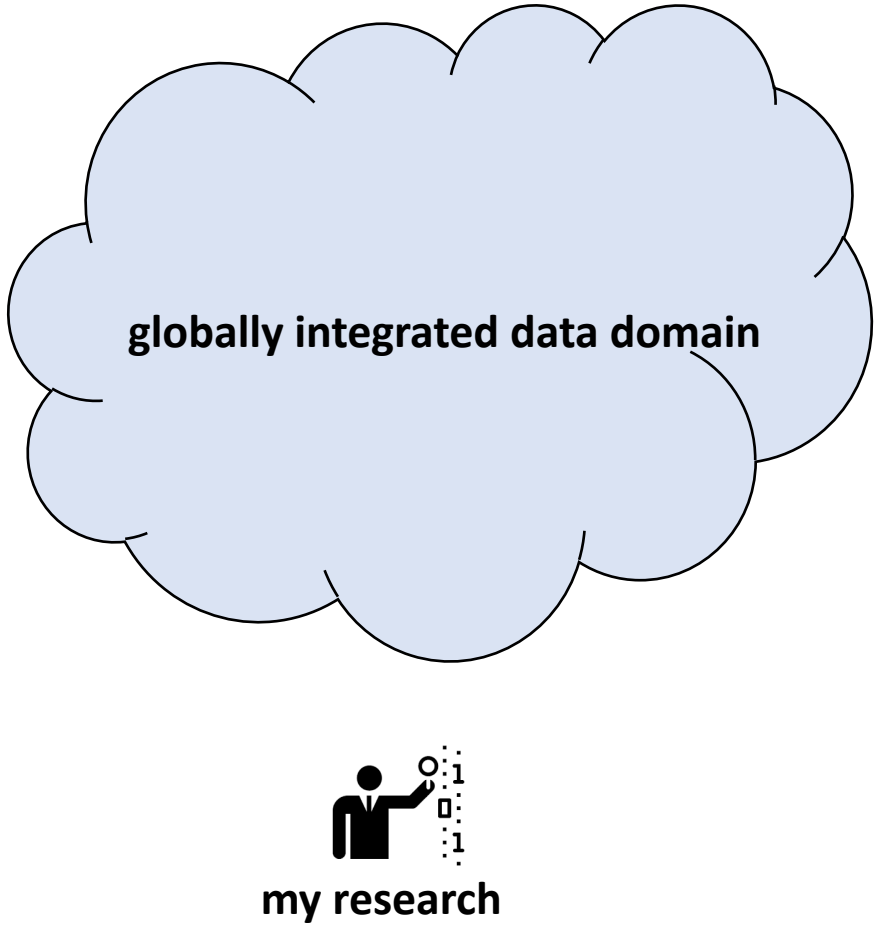
- know the data & context
- know its structure & variables
- have probably some tools
- have trust

decoupling of producers and consumers
(10 years of discussions in DOBES)



- data & context are not obvious
- structure & variables are not clear
- tools to operate on data lake not evident
- can one trust data (quality, authenticity, etc.)
- data security & protection is an issue

Yes it will change but some challenges



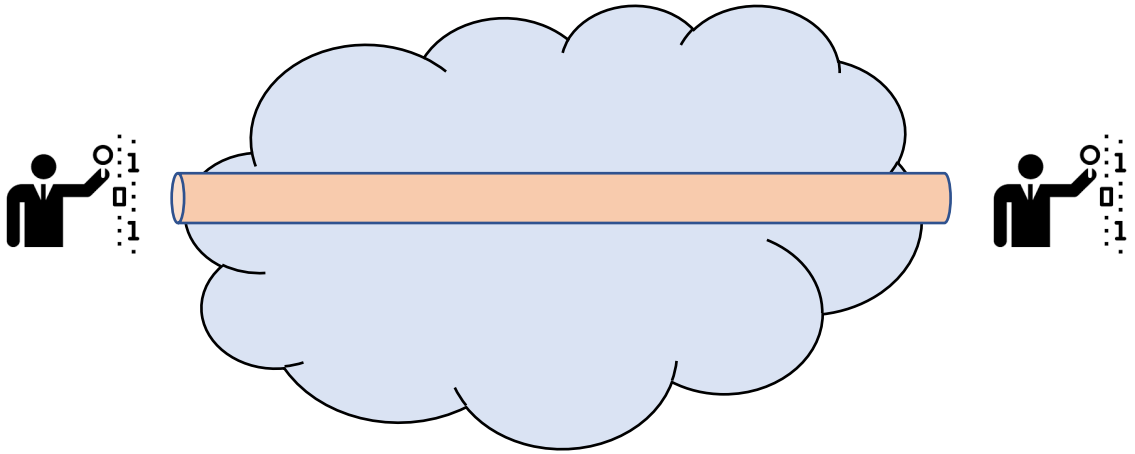
need to

- specify and register structure
- define and register variables
- describe context, provenance & relations
- establish mechanisms of trust
- establish mechanisms for protection
- manage data (stability, sustainability)
- manage relationships

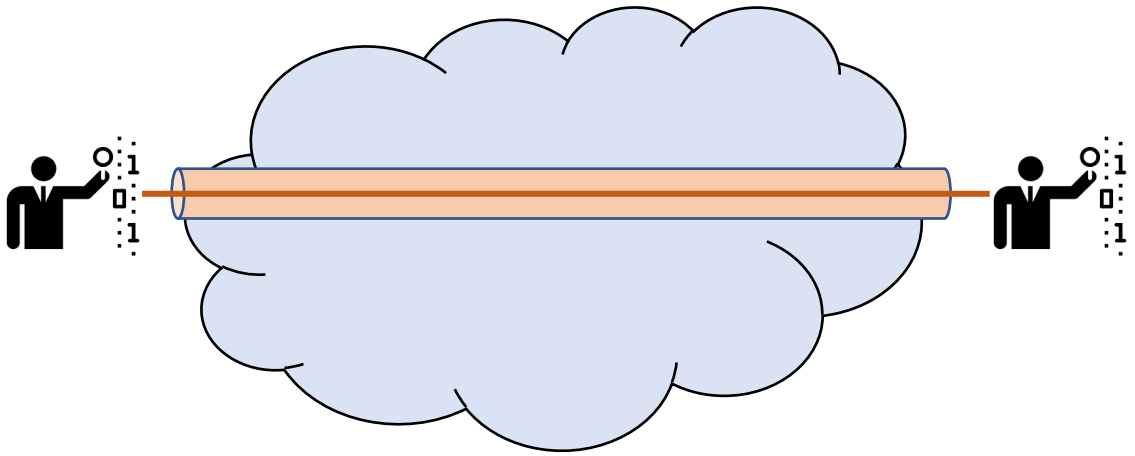
metadata

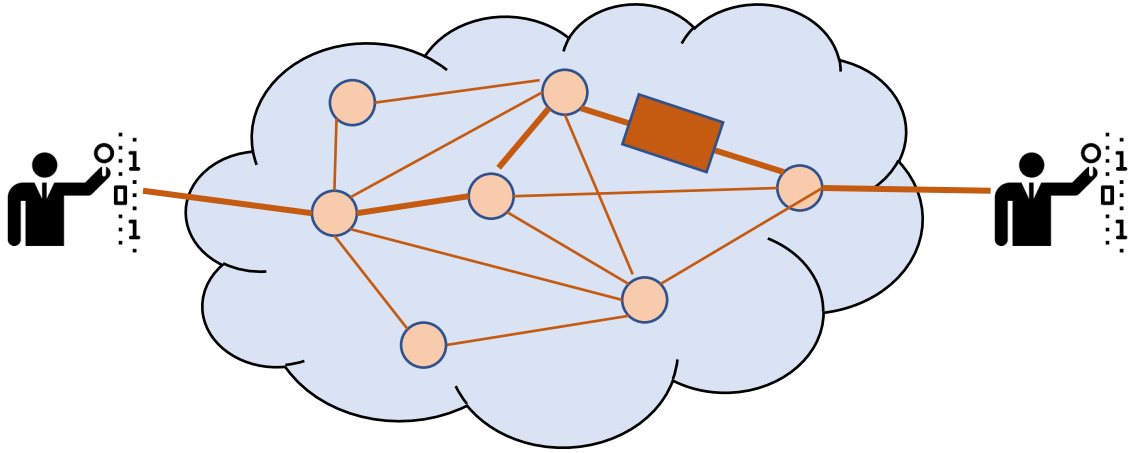
- what about tools?
- how much extra load on researchers?

About self-standing entities in large infrastructures



circuit switching: first channel – then talk





Internet

- Datagrams floating throught Internet
- Datagrams are self-standing, nodes know what to do with them

Data Entities in global space to be self-standing, since used by many

FAIR Principles give directions

- Findable (PIDs, rich metadata, indexed & searchable)
- Accessible (retrievable by PID with standard protocol, long term)
- Interoperable (standard language for knowledge representation, FAIR vocabularies, qualified references)
- Reusable (accurate & relevant attributes, license, provenance, community standard)
- **for Humans AND Machines -> machine actionable, i.e., when value is found machine needs to know how to interpret and what to do**
- **for Data & Metadata: in the sense of any bit-sequence**

- no direct statements on open access, long-term sustainability, care takers, etc.
- many different interpretations – no validators

FAIR Principles give directions

- Findable (PIDs, rich metadata, indexed & searchable)
- Accessible (retrievable by PID with standard protocol, long term)
- Interoperable (standard language for kn… …es, qualified refe…
- …rd)
- …hine
- f… …equence

FAIR Principles are guidelines for proper data practices.
Global awareness, but changing practices is difficult because …
People have their „repositories“, tools, spreadsheets and do not want to change ☺
FAIR is not a template for building an infrastructure!

- no direct statements on open access, long-term sustainability, care takers, etc.
- many different interpretations – no validators

Plug-In Concept

Dream in 1970ies

a user simply plugs-in a computer
and is part of the Internet
(TCP/IP was the game changer,
but it took 3 decades to take up)

Dream in 2020ies

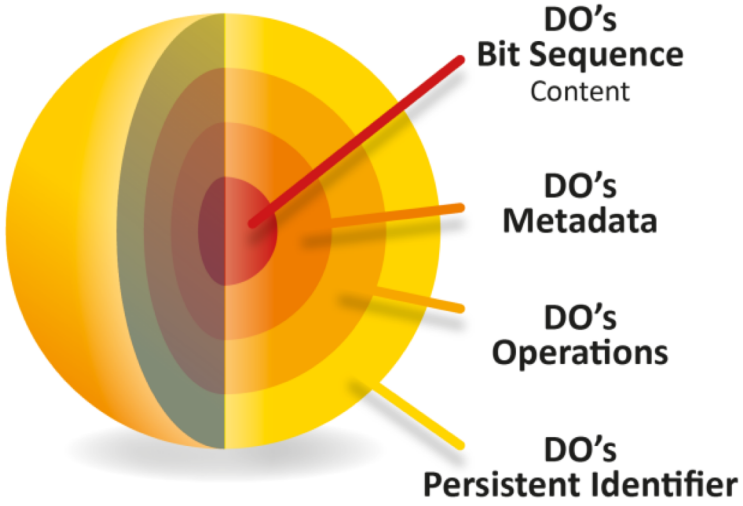
a community simply plugs-in a repository or FDO
and is part of the GIDs
(it will come, but what is the game changer
and how much time will it take)

What does „plugging-into GIDS“ mean?

- need to have a clearly identifiable, self-standing and traceable entity (like Internet‘s datagrams)
- need to have an entity that binds all relevant information persistently (type, metadata, rights, licenses, etc. for reuse)
- a repository is added to the GIDS: it offers its holding (data, all metadata) to all interested crawlers by a DO Interface Protocol to update collections, registries & portals automatically
- 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

And now the FAIR Digital Objects

Are they more than just another bit of noise?



FDO characteristics

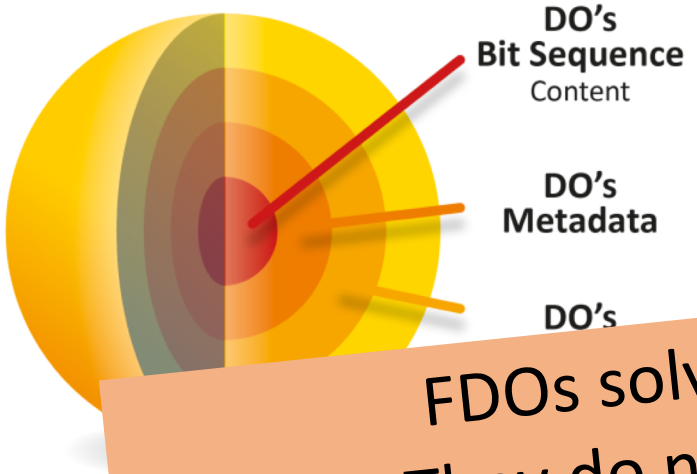
- abstraction (any content)
- persistent binding
- encapsulation

- A FAIR digital object is a unit composed of data and/or metadata regulated by structures or schemas, and with an assigned globally unique and persistent identifier (PID), which is findable, accessible, interoperable and reusable both by humans and computers for the reliable interpretation and processing of the data represented by the object.
- are atomic and self-standing by bundling all relevant information to process digital concent
- Identifier System (Handles/DOIs) is globally administered, distributed, secure, redundant, free of patents and is owned by the Swiss non-profit DONA Foundation

  (DOIs = Handles with prefix 10 & business model)

And now the FAIR Digital Objects

Are they more than just another bit of noise?



FDOs solve the data-organisation aspects of FAIR
They do not address (yet) richness of Metadata!!!!

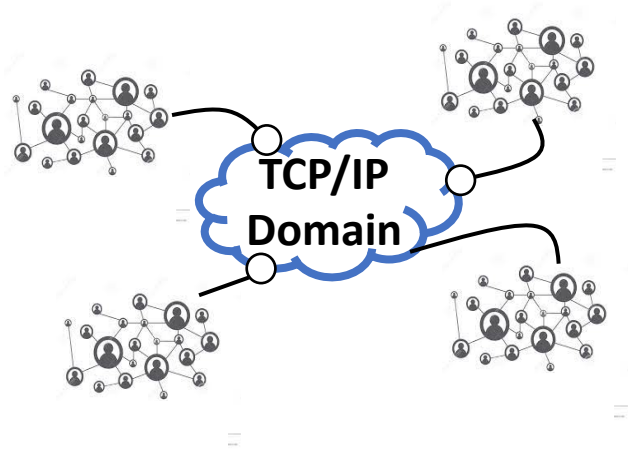
- A FAIR digital object is a unit composed of data and/or metadata regulated by structures or schemas, and with an assigned globally unique and persistent identifier (PID), which is findable, accessible, interoperable and reusable both by humans and
- ed by
- are atomic and self-standing by bundling all relevant information to process digital concent
- Identifier System (Handles/DOIs) is globally administered, distributed, secure, redundant, free of patents and is owned by the Swiss non-profit DONA Foundation

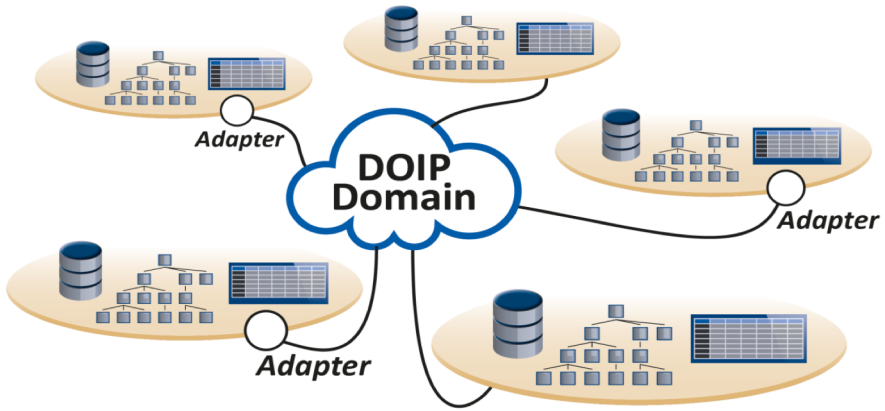
  (DOIs = Handles with prefix 10 & business model)

FDO characteristics

- abstraction (any content)
- persistent binding
- encapsulation

A Domain of FAIR Digital Objects





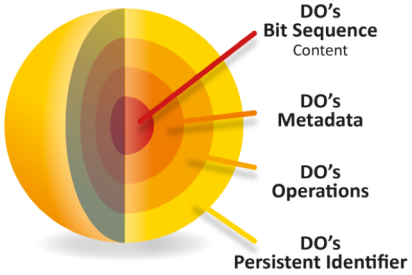
	Internet	Integrated Data Space
challenge	creating an integrated computer network	creating an integrated global data space
heterogeneity	100s of networks	100s of standards, 1000s of repositories, 10000s of tools
heterogeneity	mode, packaging	data organisation and modelling
basic protocol	TCP/IP	IRP, DOIP
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** (being extended)
- this is currently turned into technical specs allowing to develop **validators** and to enable a "**plug-in" domain**
- a **DO Interface Protocol (DOIP)** specification, a software implementation and a reference repository (server), **Data Type Registry**, **Proxies**, etc.
- a variety of initiatives (ESFRIs, US, etc.) working on **implementations** and **demonstrators**
- Missing is a large & integrated demonstrator

- **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, DONA, IDF

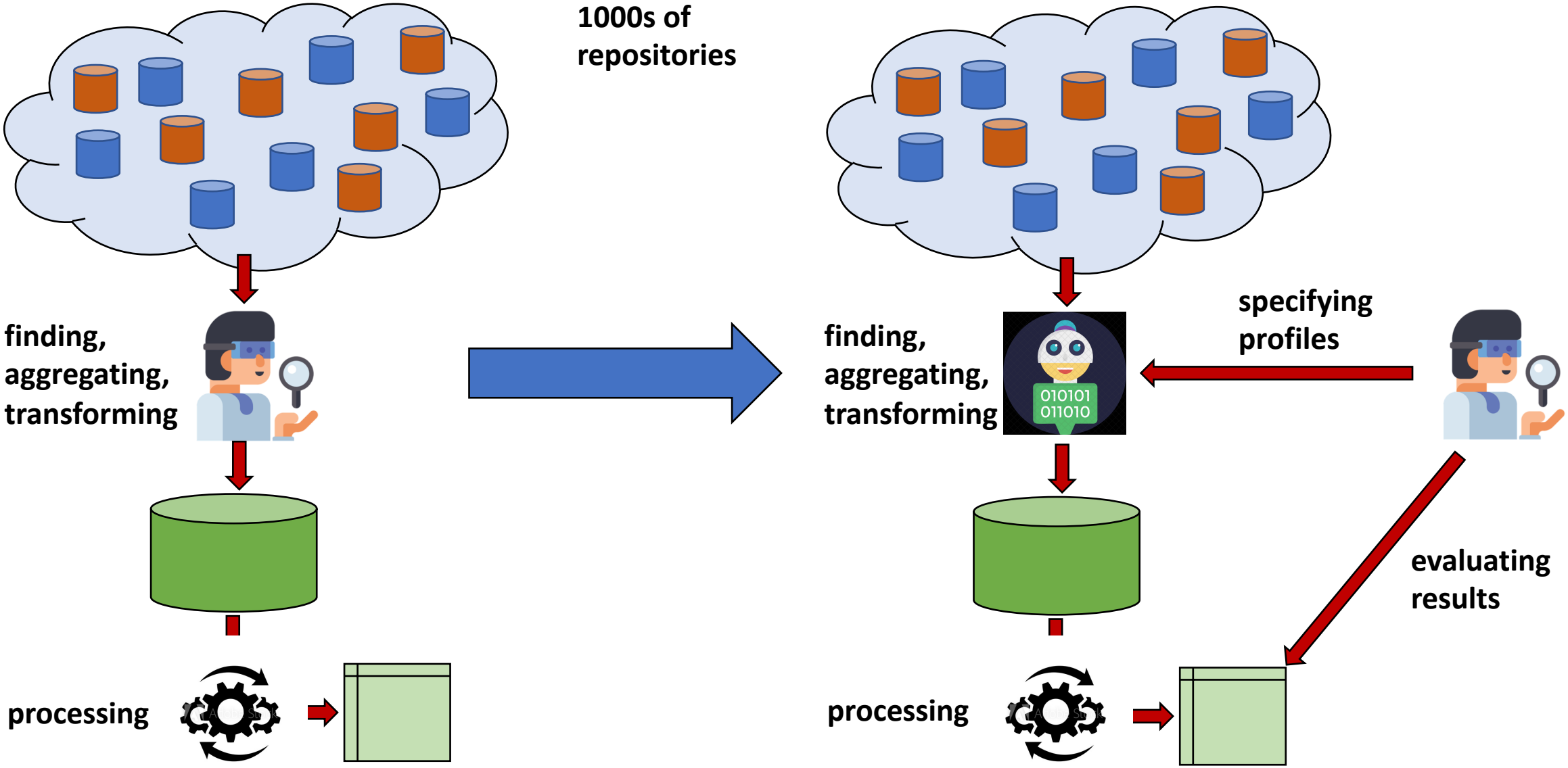
FAIR **DIGITAL OBJECTS** FORUM
https://fairdo.org/



Thanks for the attention.

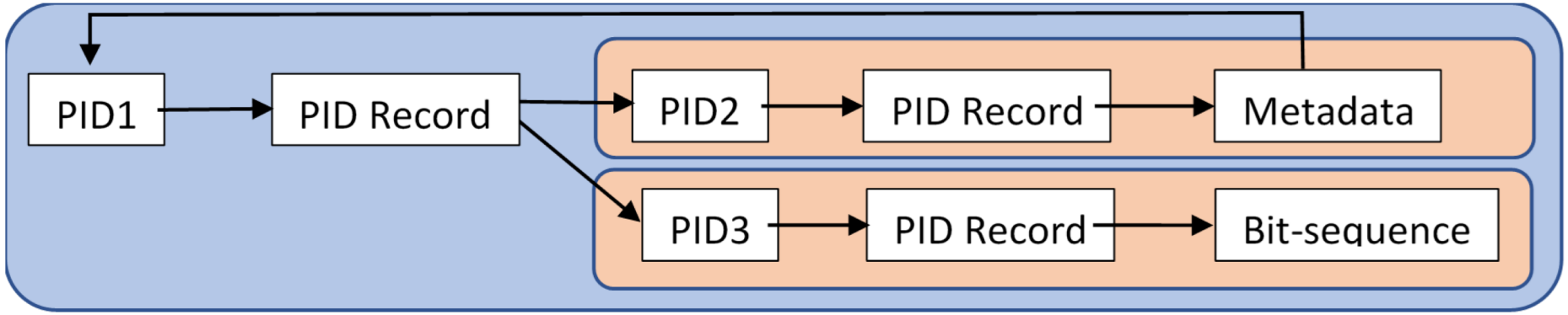
- Wittenburg & Strawn: **Common Patterns in Revolutionising Infrastructures & Data**; http://doi.org/10.23728/b2share.4e8ac36c0dd343da81fd9e83e72805a0
- de Smedt, Koureas & Wittenburg: **Analysis of Scientific Practice towards FAIR Digital Objects**; http://doi.org/10.23728/b2share.e14269d07ce84027a7f79ee06b994ef9
- **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
- EOSC: **Turning FAIR into Reality**: https://ec.europa.eu/info/sites/default/files/turning_fair_into_reality_1.pdf

Possible Impact: changed researcher role



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.

Typical Domain of Data Types - Neuroscience

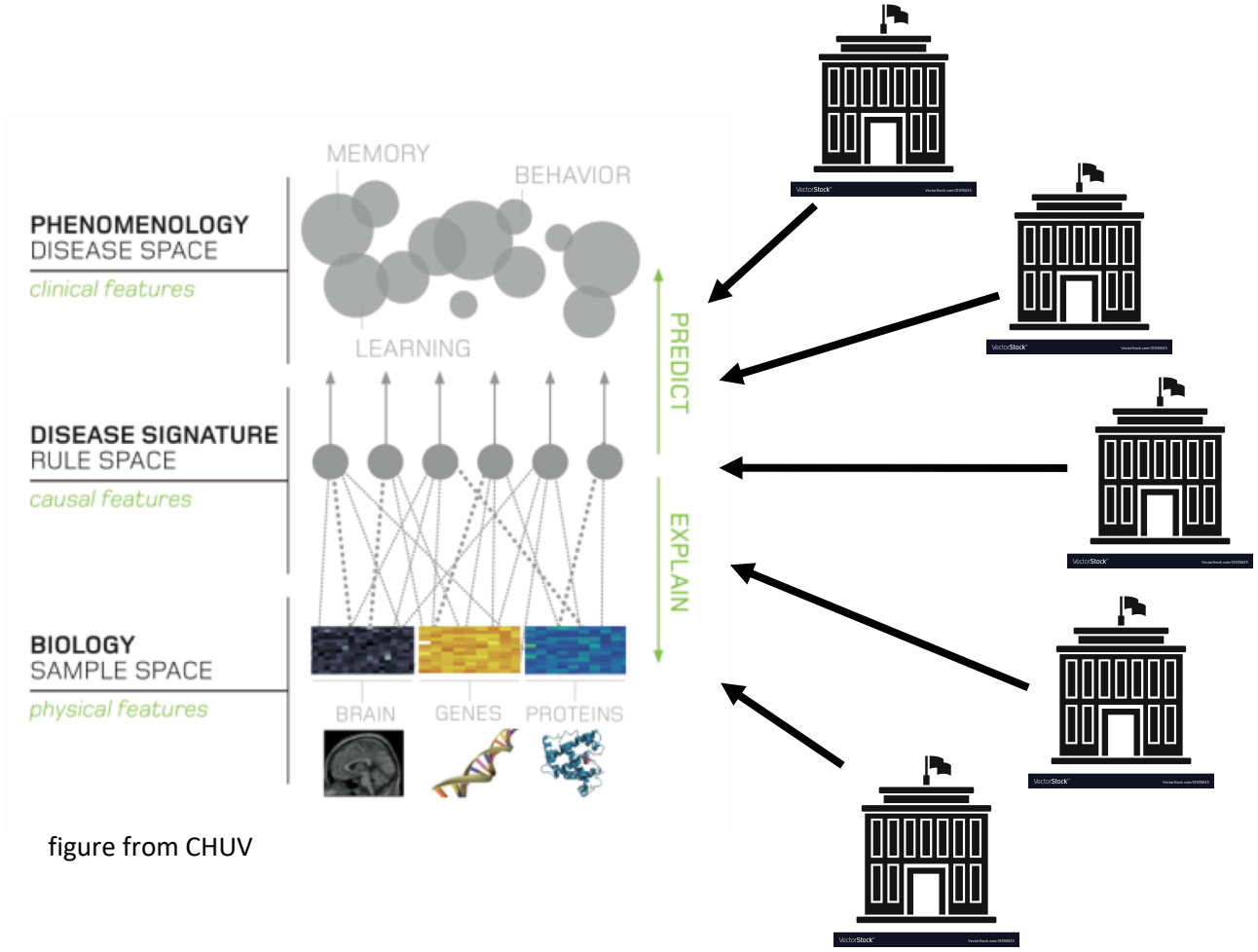


figure from CHUV

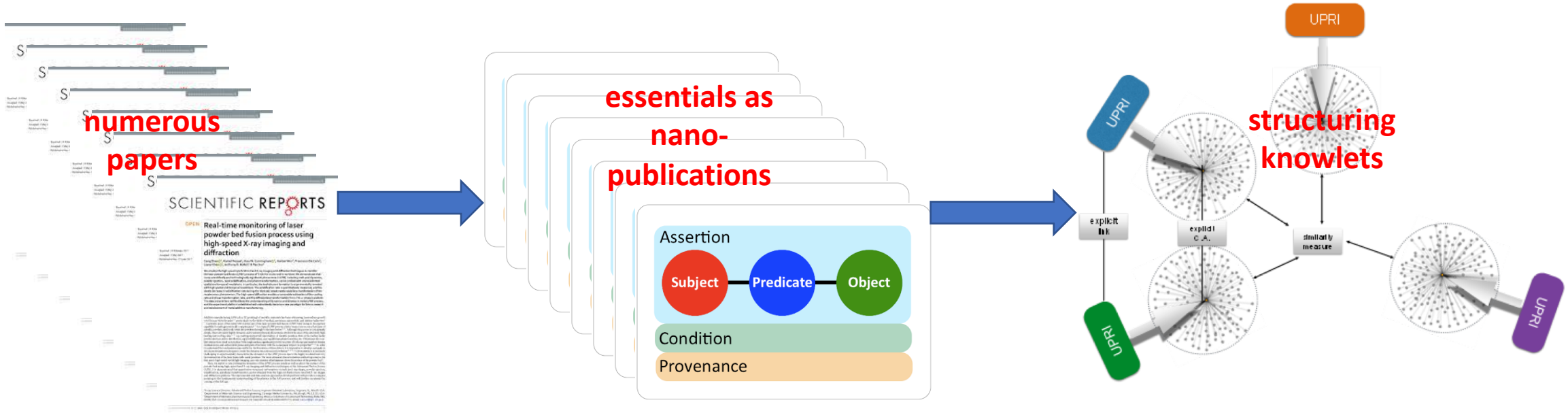
Much data of different types (observations, brain images, gene sequences, etc.) from different institutes to identify all model parameters.

Many computations are done until model is satisfying.

Has everything incl. various transformations been documented to know what was done 6 months ago?

How to maintain a stable and reusable domain of data over decades?

Data Types in „Knowledge“ Work



- essentials of papers extracted to nano-publications (augmented RDF)
- Nano Publications can be analysed (statistics, knowlets, etc.) (Mons)
- knowlets: structuring the domain of knowledge
- needs to be part of our scientific memory

How to maintain a stable and reusable domain of data over decades?