FAIR - one pillar towards Convergence

Peter Wittenburg

Max Planck Society, Max Planck Computing & Data Facility

RDA

Dynamics will Continue

volumes
50 Mio smart devices
producing continuous data
streams

dynamics
enormous acceleration
of dynamics and
heterogeneity

Recently at the IoT Week a colleague from WUR gave a talk.
Thus, WUR is aware of these challenges.

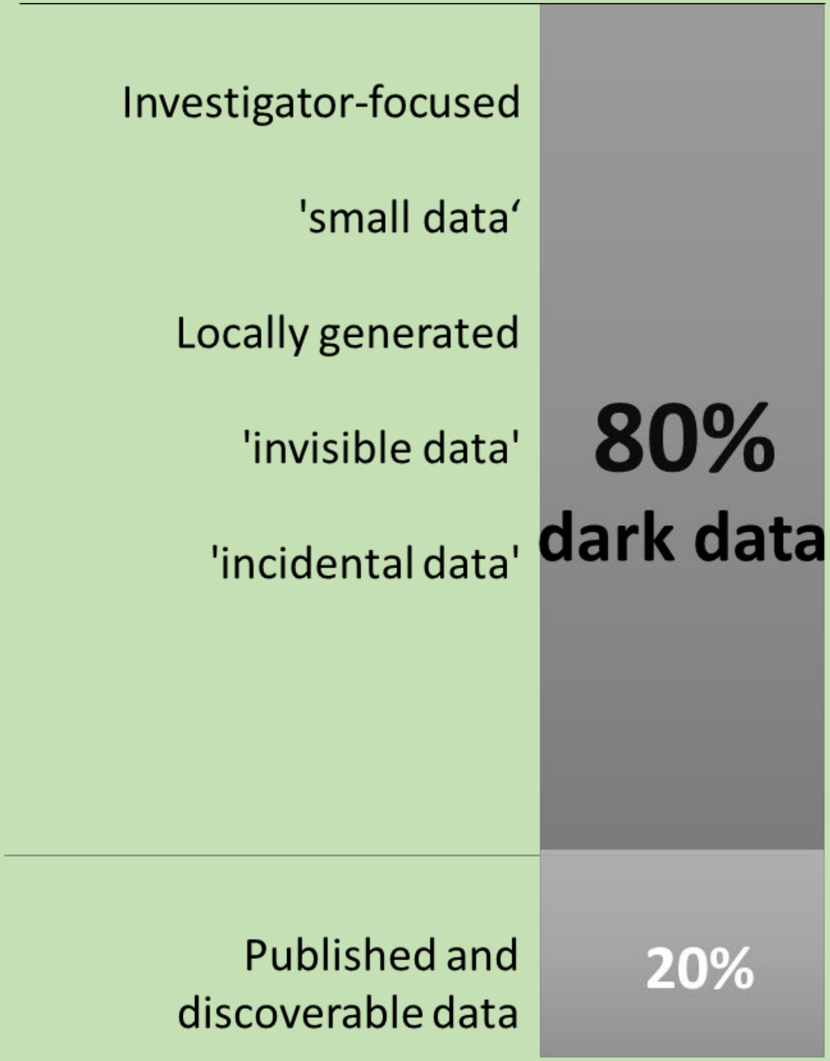
Data Practices are too costly

- **Inefficiencies prevent many data intensive projects and broad participation in science & industry**
  - RDA EU 2013 Survey: 75% of time of data scientists is wasted on data wrangling
  - M. Brodie MIT Survey: 80% of time of data scientists is wasted on data wrangling
  - CrowdFlower 2017 Survey: 79% of time of data scientists is wasted on data wrangling
- **biggest cost factors: bad & non-explicit data organisation,bad data quality**
- **about 60% of data intensive projects fail**

Data science suffers from heterogeneity, proliferations (tools, standards),

lack of interoperability at all layers.

Dark Data Issue



Investigator-focused

'small data‘

Locally generated

'invisible data'

'incidental data'

80% **dark data**

Published and discoverable data

20%

Dark data lost within 20 years

Despite significant investment, data is not being managed effectively

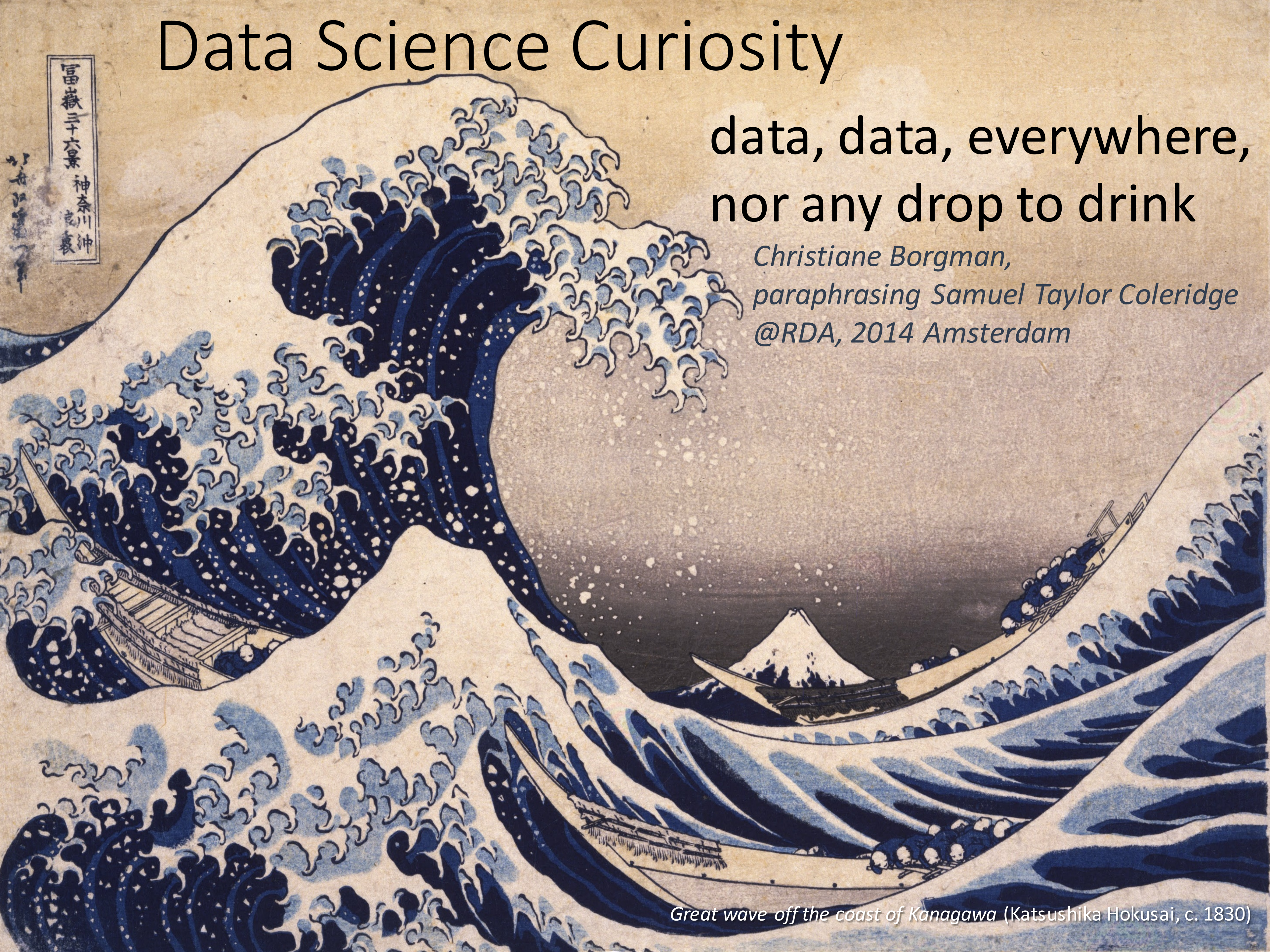
$1.5 TRILLION is the current estimated total global spend on R&D, which could be at risk[3]

80% lost

In one study, the odds of sourcing datasets declined by 17% each year, with 80% of datasets over 20 years old not available[4]

[1] Heidorn PB. *Library Trends* **57**:280-299

Data Science Curiosity



data, data, everywhere,
nor any drop to drink

Christiane Borgman,
paraphrasing Samuel Taylor Coleridge
@RDA, 2014 Amsterdam

Great wave off the coast of Kanagawa (Katsushika Hokusai, c. 1830)

Awareness from about 2005

2007: OECD's Principles and Guidelines for Access to Research Data from Public Funding

2007: Jim Gray: A Transformed Scientific Method (4th paradigm)

2010: EC's High Level Expert Group Report *Riding the Wave*

demanding urgent funding actions that would help changing data practices

2012 DAITF workshop at ICRI Conference in Copenhagen (start of RDA)

L. Lannom's four DAIR layers: "Discovery, Accessing, Interpreting and Reusing"

2013 Research Data Alliance start inspired by DAIR

2013 G8 Science Ministers Report

Open Scientific research data should be easily discoverable, accessible, intelligible, useable, and wherever possible interoperable to specific quality standards.

2014 RDA Data Foundation & Terminology Group (and more in RDA)

Core Data Model with Digital Objects as core based on many use cases

2014 Workshop at the Lorentz Centre Leiden (-> FORCE11, Nature)

FAIR principles are now a globally accepted minimal set of behaviours enabling Findability, Accessibility, Interoperability and Reusability by humans and in particular by machines

FAIR Principles (known!?)

F1 (meta)data are assigned a globally unique and persistent identifier;
F2 data are described with rich metadata;
F3 metadata clearly and explicitly include the identifier of the data it describes;
F4 (meta)data are registered or indexed in a searchable resource;
A1 (meta)data are retrievable by their identifier using a standardized communications protocol;
A1.1 the protocol is open, free, and universally implementable;
A1.2 the protocol allows for an authentication and authorization procedure, where necessary;
A2 metadata are accessible, even when the data are no longer available;
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;
R1 meta(data) are richly described with a plurality of accurate and relevant attributes;
R1.1 (meta)data are released with a clear and accessible data usage license;
R1.2 (meta)data are associated with detailed provenance;
R1.3 (meta)data meet domain-relevant community standards;

FAIR Principles (known!?)

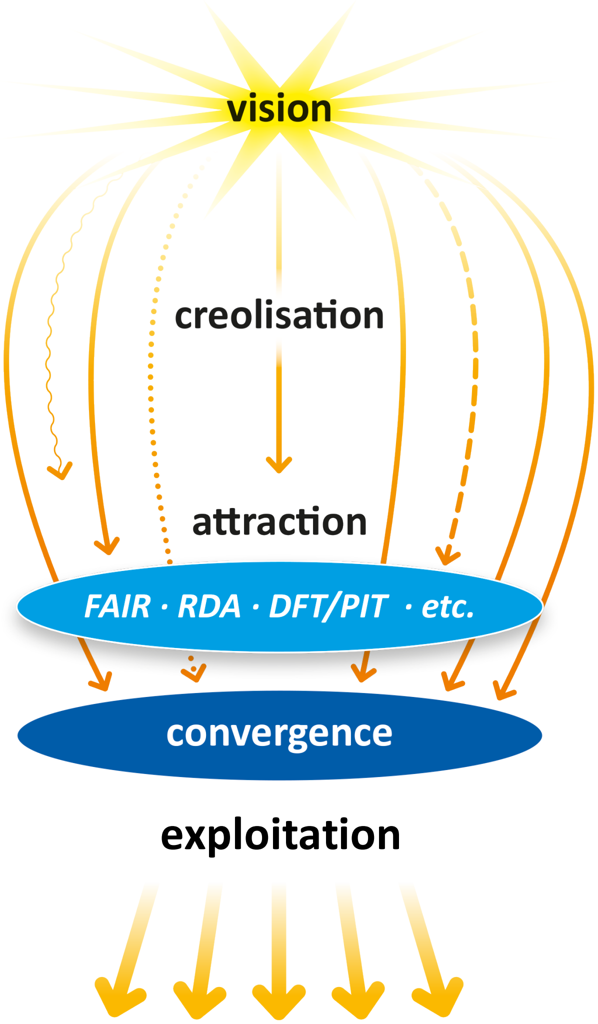
F1 (meta)data are assigned a globally unique and persistent identifier;
F2 data are described with rich metadata;
F3 metadata clearly and explicitly include the identifier of the data it describes;
F4 (meta)data are registered or indexed in a searchable resource;
A1 (meta)data are retrievable by their identifi

This is excellent – much convergence at the level of principles. However, FAIR principles are not a blueprint for building infrastructures.

We need to do more in particular when we want to realise federations of repositories to integrate data from different sources (see EOSC).

principles;
I3 (meta)data include qualified references to other (meta)data;
R1 meta(data) are richly described with a plurality of accurate and relevant attributes;
R1.1 (meta)data are released with a clear and accessible data usage license;
R1.2 (meta)data are associated with detailed provenance;
R1.3 (meta)data meet domain-relevant community standards;

Creolisation in Data Domain



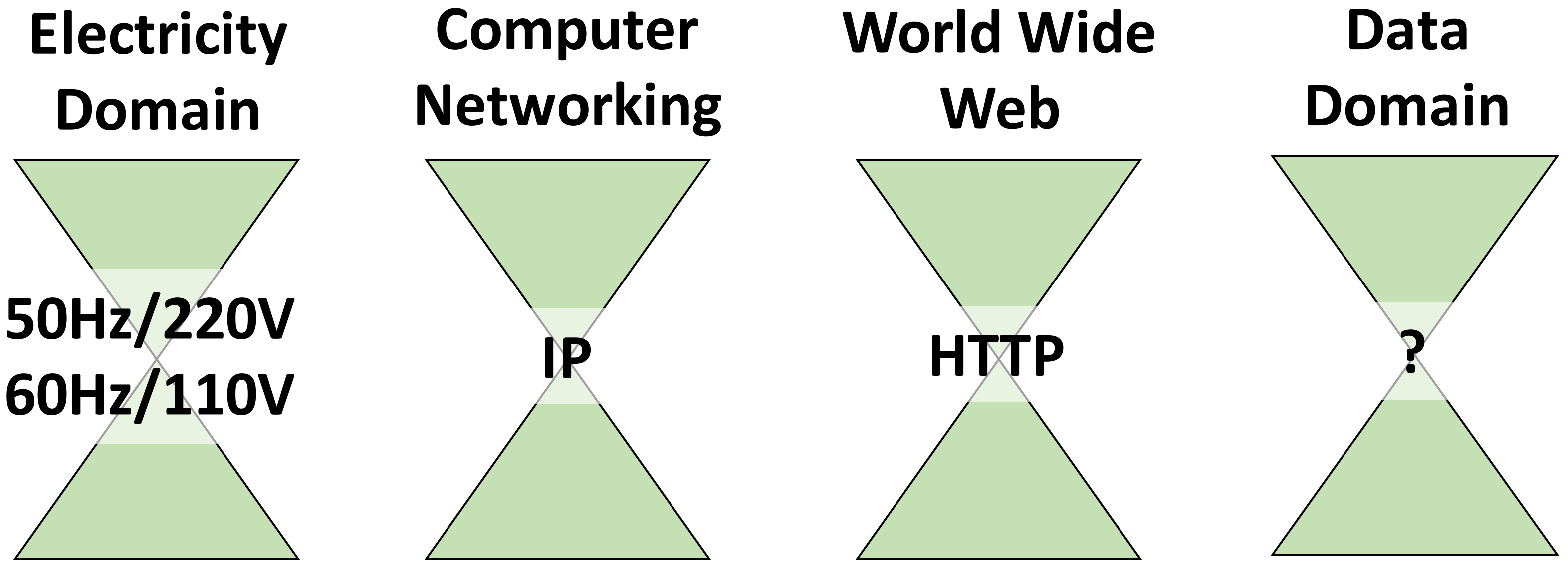
Recent Paper from Wittenburg & Strawn
Common Patterns in Revolutionary Infrastructures and Data

Right in the Creolisation Phase

- so many brilliant minds
- enormous solutions space – 1000 flowers ...
- tested quite some approaches

Convergence is needed! ...
... but at which level?
... but how to organise the cross-border process?

One way to compare



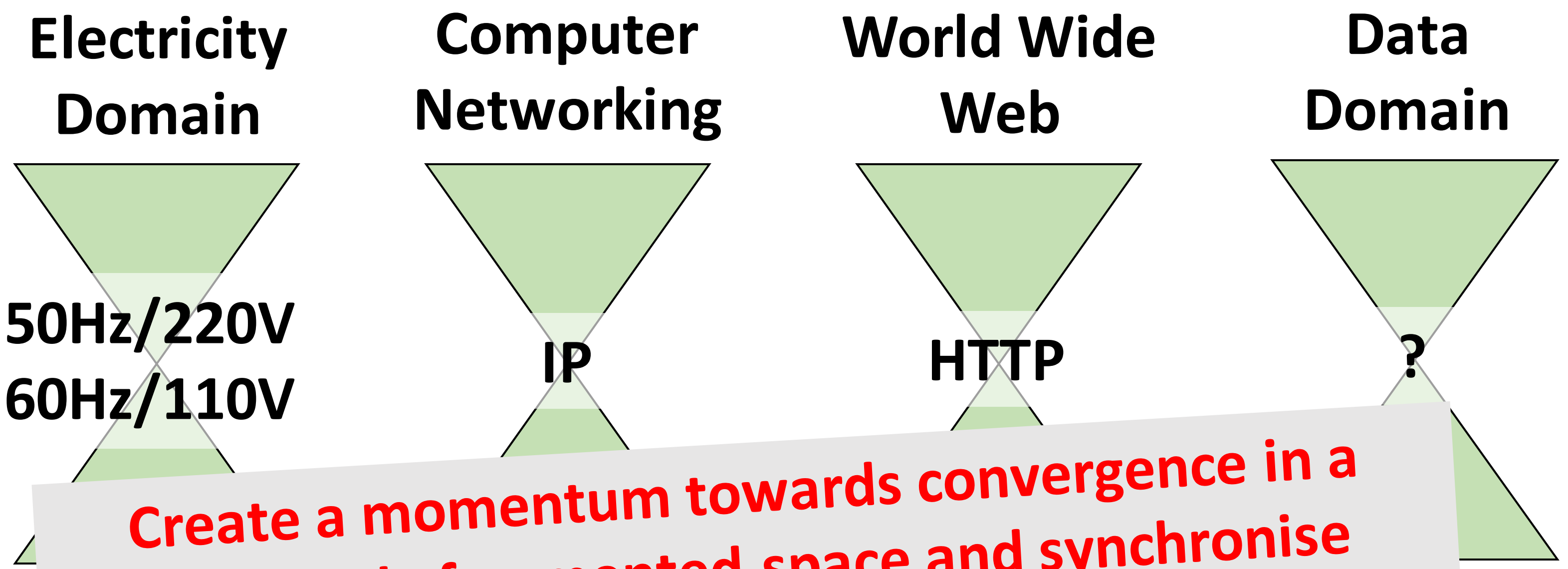
Define a solid basis for future developments and big investments.

Where do we talk about: 5 y – 20 y – 100 y?

J. Hendler (W3C): need a new commodity layer again

G. Strawn (BRDI): seem to be at a point comparable
to the moment where Internet was born

One way to compare



Create a momentum towards convergence in a hopelessly fragmented space and synchronise minds who all have so brilliant ideas.

1. **connect islands**
2. **changing internal structures**

De **ents.**

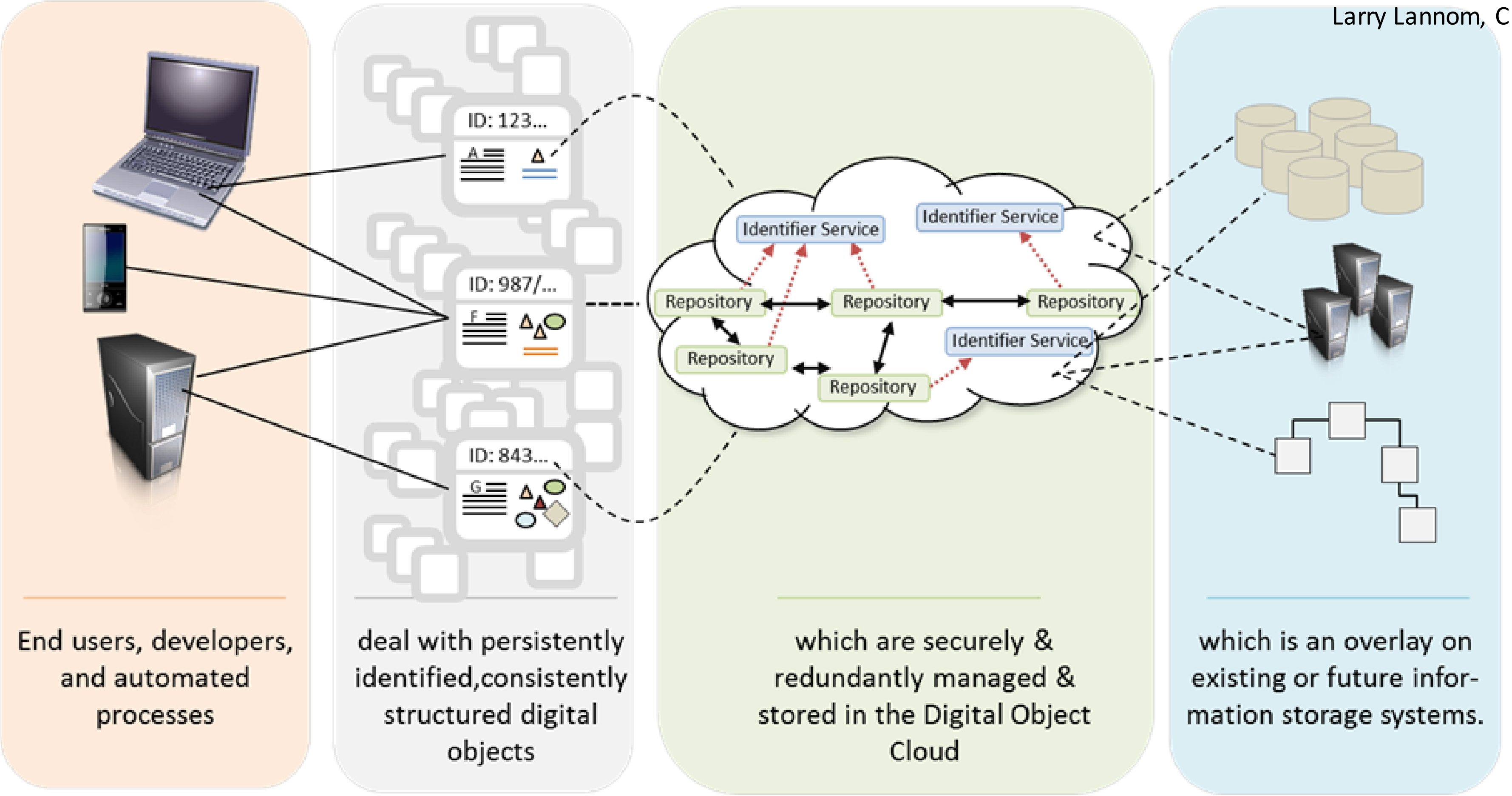
Wh

J. Hendler (W3C): need a new commodity layer again

G. Strawn (BRDI): seem to be at a point comparable to the moment where Internet was born

Need for Abstraction

diagram from
Larry Lannom, CNRI



End users, developers, and automated processes

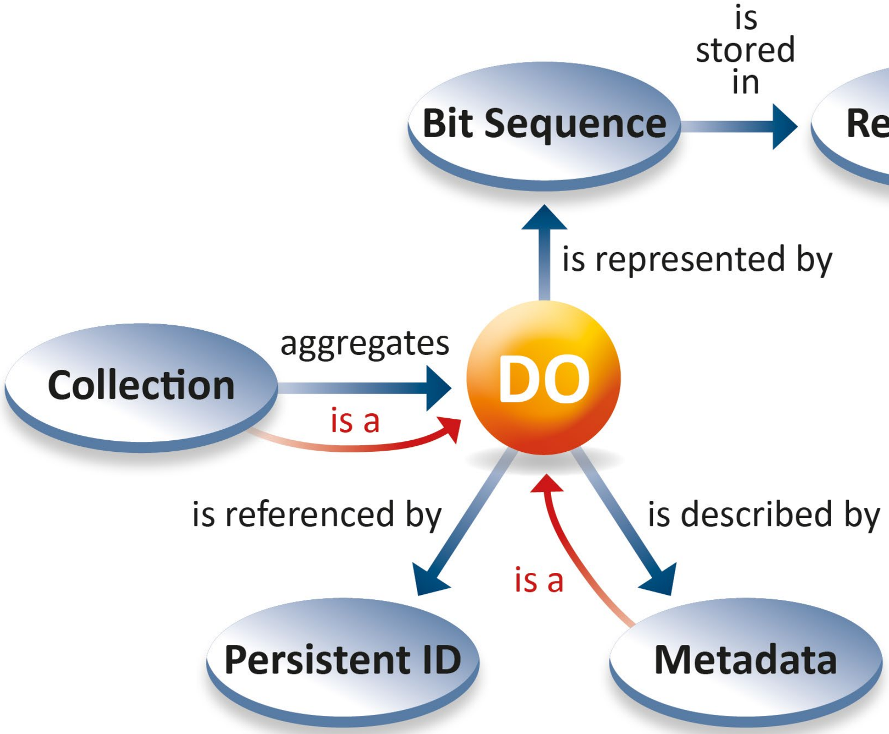
deal with persistently identified,consistently structured digital objects

which are securely & redundantly managed & stored in the Digital Object Cloud

which is an overlay on existing or future information storage systems.

ideally: users only deal with Metadata and PIDs independent of types and implementation details

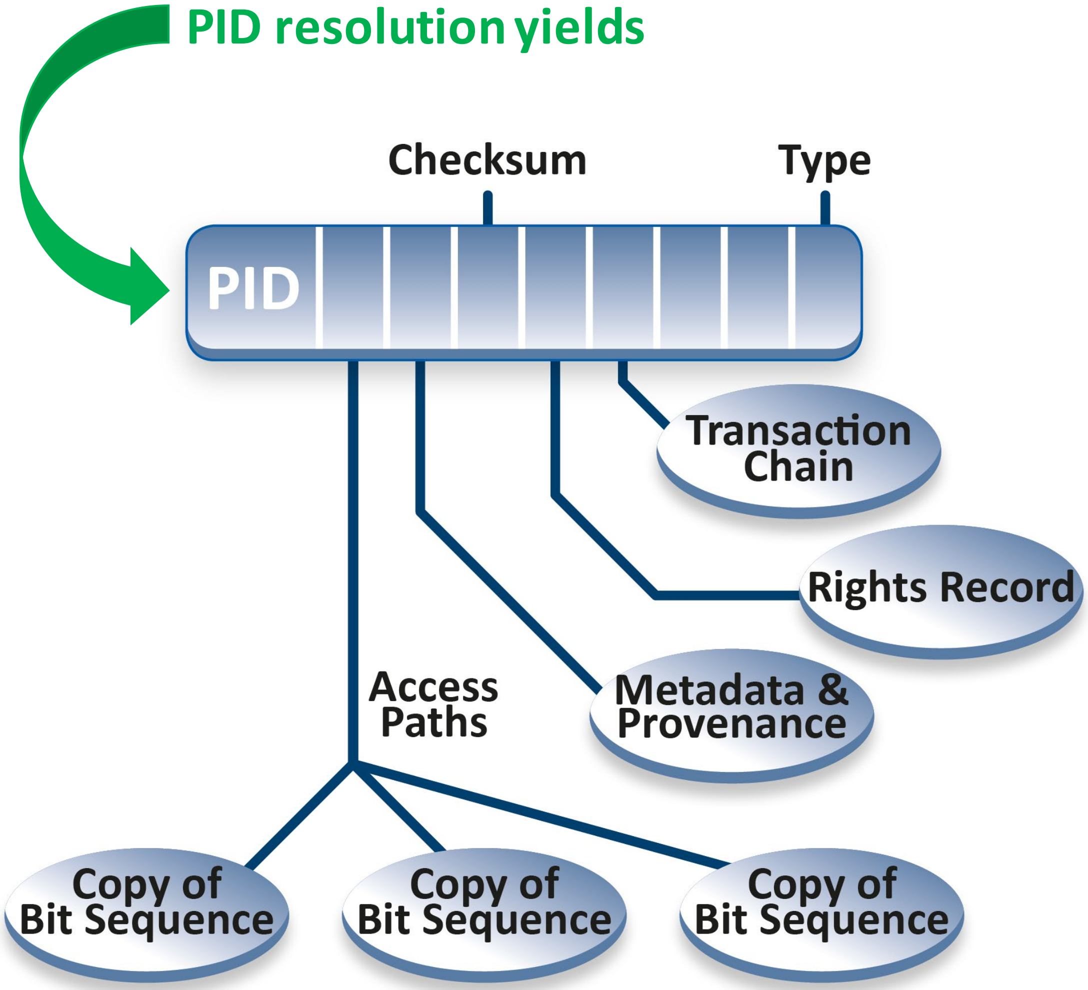
Digital Objects offer Abstraction

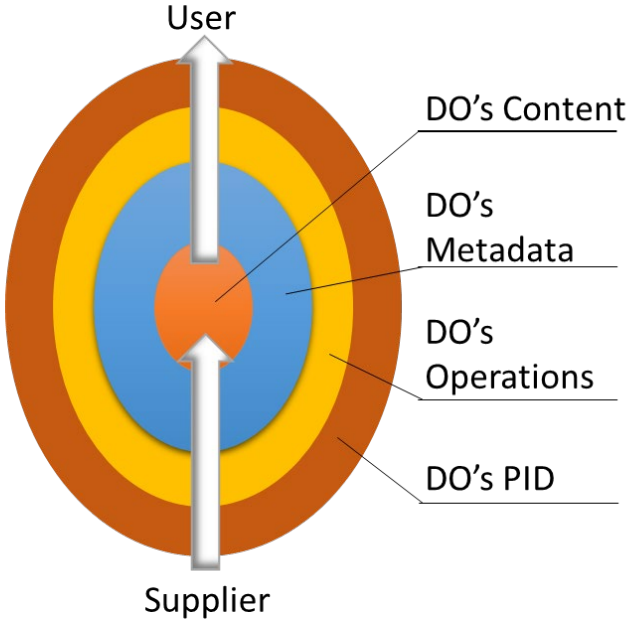


can be many repositories

- **RDA DFT Core Model based on many use cases across disciplines**
- **very simple – but expressive in terms of specifying interoperability at data organisation**

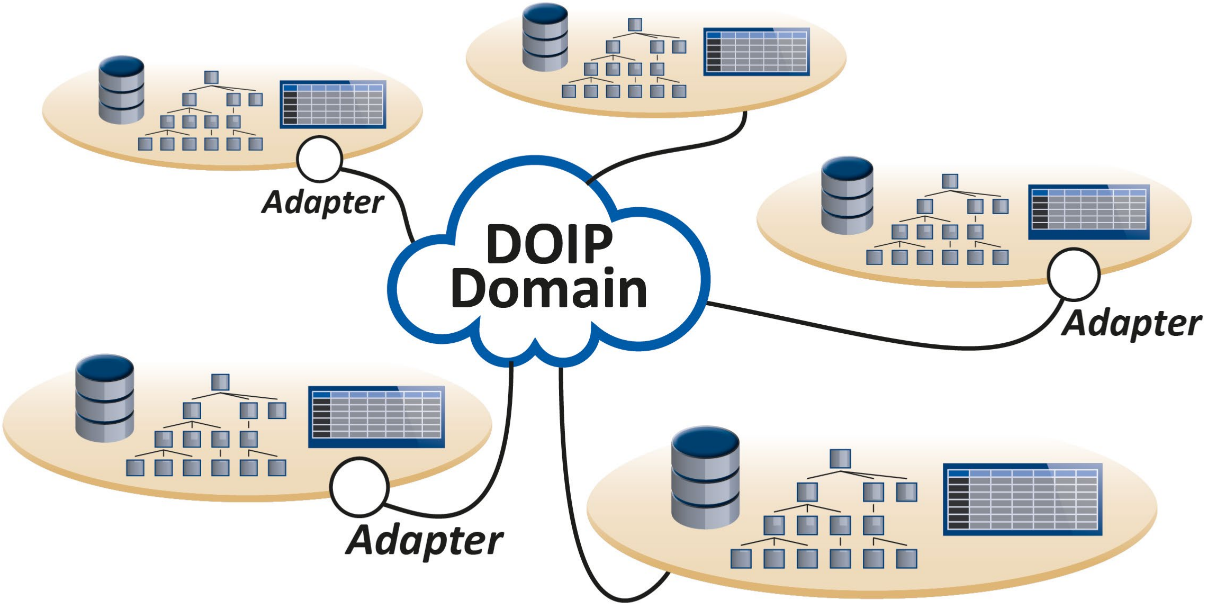
DOs also do the Binding





- persistent PIDs can be used to store relevant attributes in record
- specify „passport“ attributes such as „checksum, type“ to facilitate machine interpretation
- currently standardisation of core attributes in RDA
- but flexibility required

Connecting Repositories with DOs



Adapter

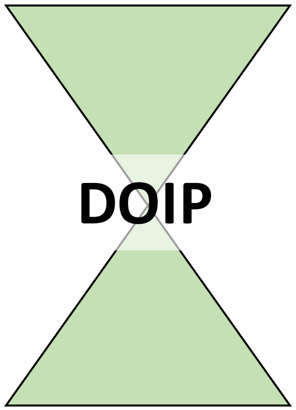
DOIP Domain

Adapter

Adapter

Adapter

Data Domain

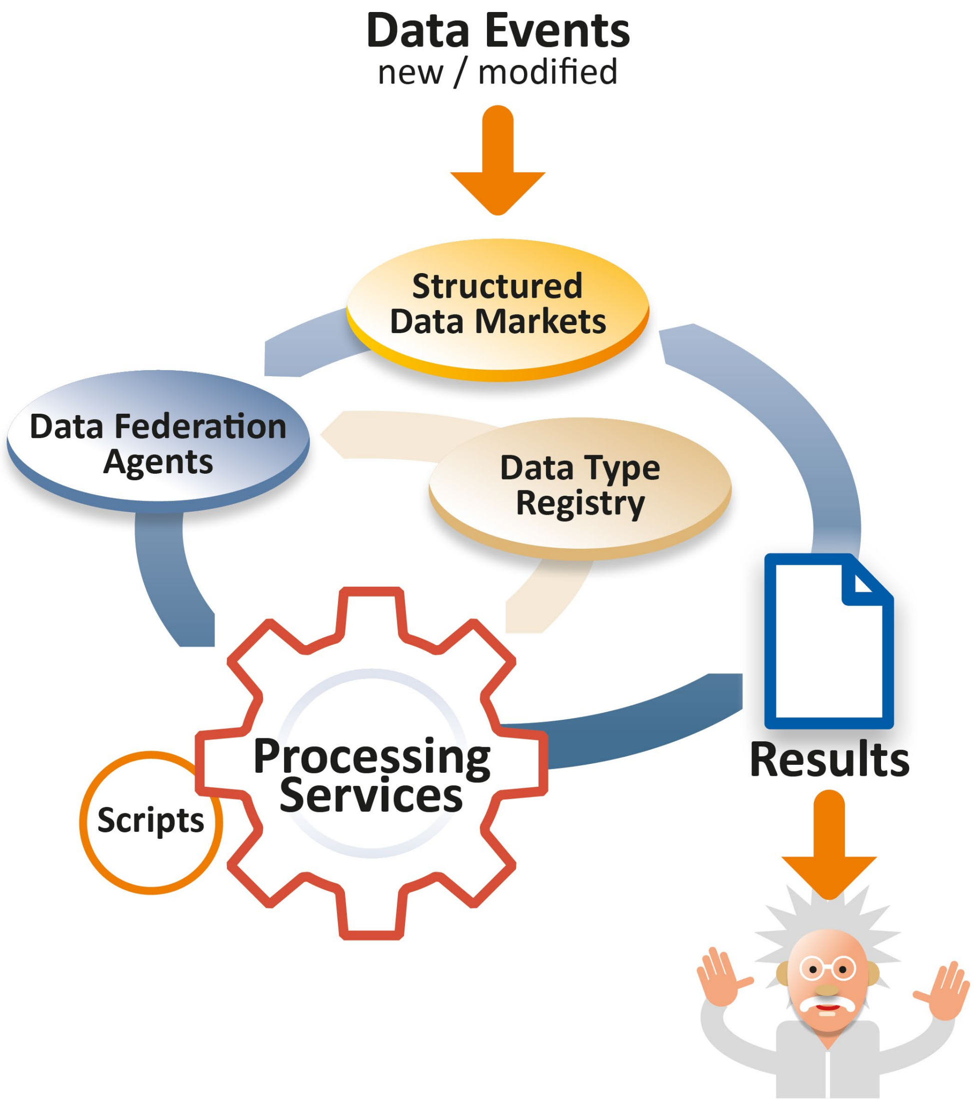


DOIP

DOIP V2.0 published all open & free

- **at least interoperability between repositories whatever data model and organisation they use**
- **some have compatible native organisation, others need to write more complex adapters**
- **not addressing Semantic interoperability at scientific encoding level, but facilitating**

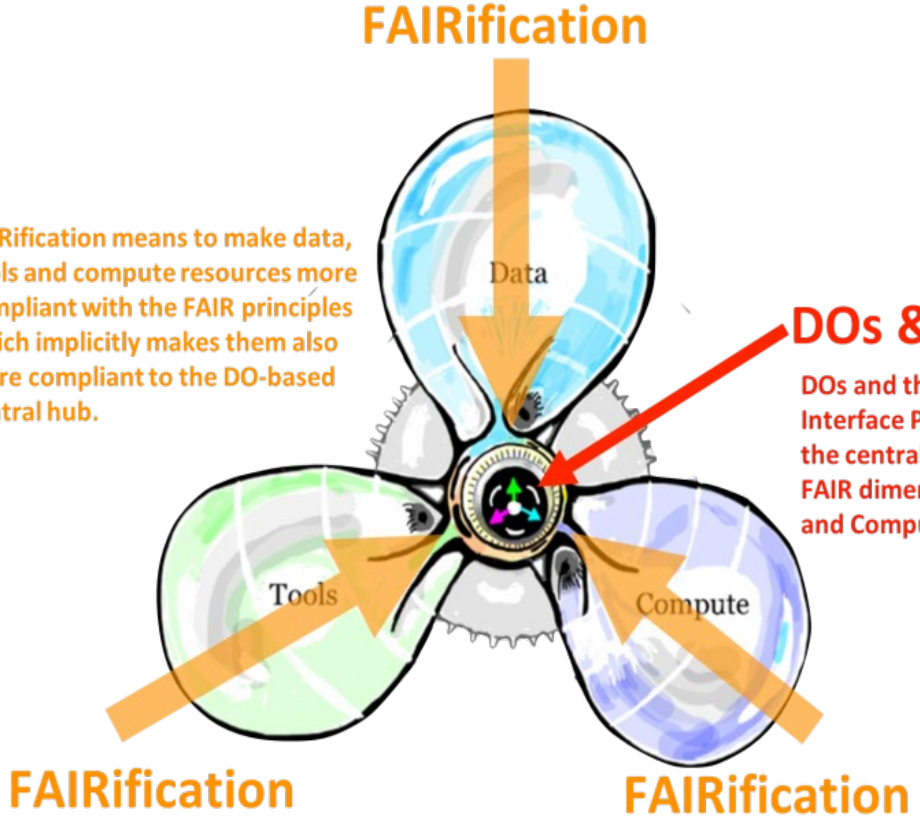
DOs facilitate Automatic Processing



- **Massiveness of data streams and wish to re-combine data requires radical shifts**
- **Agents should react on incoming data which are suitable for the specific business case**
- **Digital objects "find themselves"**

Basis are Digital Objects (Data, Software, Configurations, etc.) and their Types

DOs and FAIRification are Twins



FAIRification means to make data, tools and compute resources more compliant with the FAIR principles which implicitly makes them also more compliant to the DO-based central hub.

DOs and the unified DO Interface Protocol can act as the central hub relating the FAIR dimensions Data, Tools and Compute Resources.

FAIRification ideas in GOFAIR aligned with the DO concept.

DOs Relevance for Open Science

Scientists develop and thrive within their respective small communities of practice, but results need to be global.

Trust lost when datasets disconnect from:

context in which they were created, or **communities** who created them.

- DOs can be self-contained and can convey the context in which datasets were generated
- it gives each digital entity an identity allowing to prove identity and authenticity even after years
- types of metadata are available even for machine processing (descriptive, system, rights, provenance, etc.)
- transactions can be verified
- respect the domain-specific specificities

Seem to agree & many interactions

- RDA a global initiative supporting PID and DO work
- C2CAMP a bunch of global actors implementing DO
- CODATA a global initiative at policy level
- GEDE Initiative bringing scientific communities together
- FAIR Principles guiding science (PID, Metadata, etc.)
- GO FAIR implementation network
- EC EG on FAIR Implementation (FAIR DO)
- EOSC plans to build European Infrastructure (federating ESFRIs)
- RDA - IoT Forum collaboration (industry)
- RDA – BDVA collaboration (industry)
- RDA – ITU collaboration (industry)

still so many open questions wrt rights, licensing and ethics

Seem to agree & many interactions

- RDA a global initiative supporting PID and DO work
- C2CAMP a bunch of global actors implementing DO
- CODATA a global initiative at policy level
-
-
-
-
-)
- (industry)
- RDA – BDVA collaboration (industry)
- RDA – ITU collaboration (industry)

still so many open questions wrt rights, licensing and ethics

It‘s now time to go beyond the FAIR guidelines and „build“ the complex eco-system (bottom-up, top-down).
A DO based data domain can be one essential pillar.
It will solve basic interoperability issues and kick off redesign.
Basic components are ready to go.

What to do in a RO?

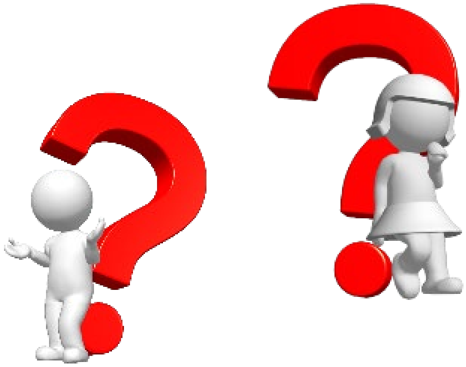
extract knowledge from data – integrate knowledge

have the infrastructure that enables this

- organise yourself (if not already done)
  - create FAIR compliant policy guidelines and ask for DMPs
  - have a specialist group to help researchers, to give advice (metadata, PIDs, FAIRification, CoreTrustSeal, etc.) and to be active (RDA, GOFAIR, GEDE, etc.)
  - provide a trustworthy repository (CoreTrustSeal, FAIRmetrics)
  - train data science methods (ML, semantics, etc.)
  - invest in workflows training & snippets (Jupyter, ...),
- try to get involved – one has to get hands dirty
  - engage in data projects and test adaptations (EOSC, AI)
  - run cross-disciplinary projects (eScience center, etc.)

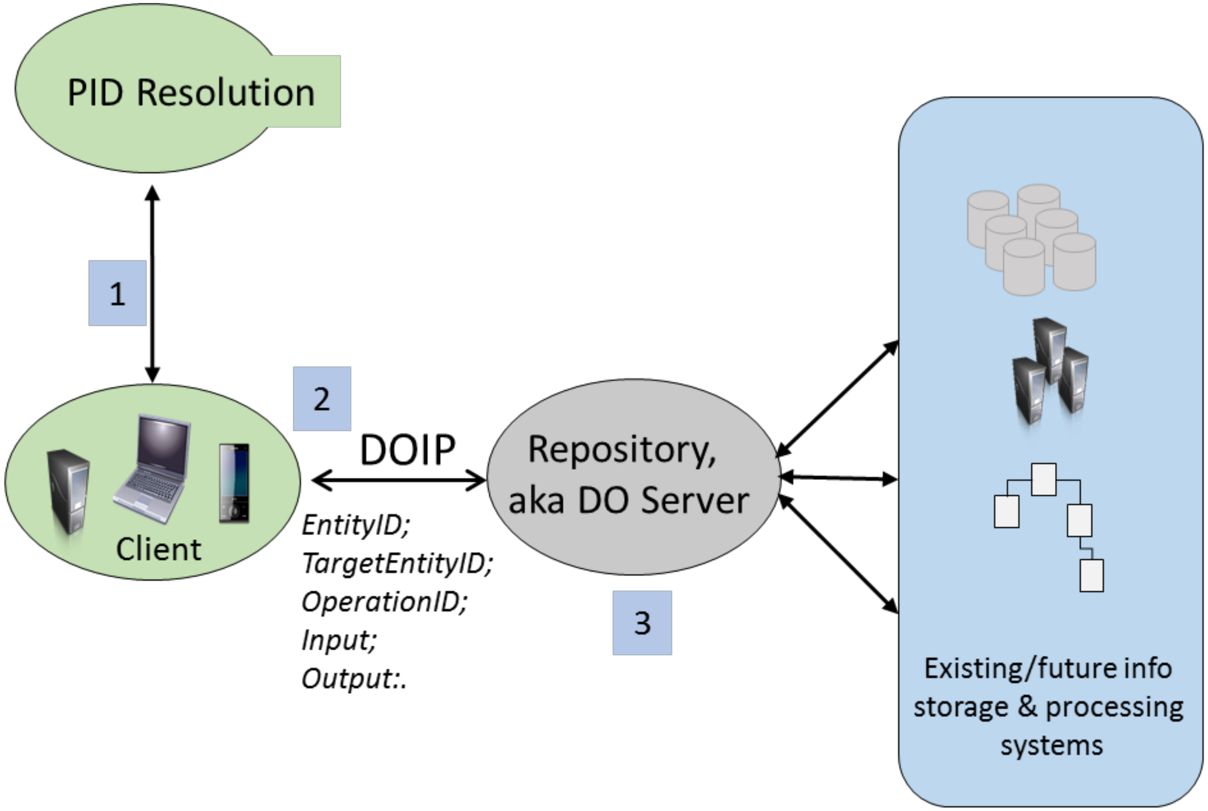
References

- Wittenburg& Strawn: Common Patterns in Revolutionary Infrastructures and Data: http://doi.org/10.23728/b2share.4e8ac36c0dd343da81fd9e83e72805a0
- GEDE Workshop on DOs: http://doi.org/10.23728/b2share.0347cfc5bddb4124a4abadbcf180bef5
- FAIR Implementation Report: https://doi.org/10.2777/1524
- DOIP Specification: https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf



Thanks for your attention.

DO Interface Protocol V2.0 published

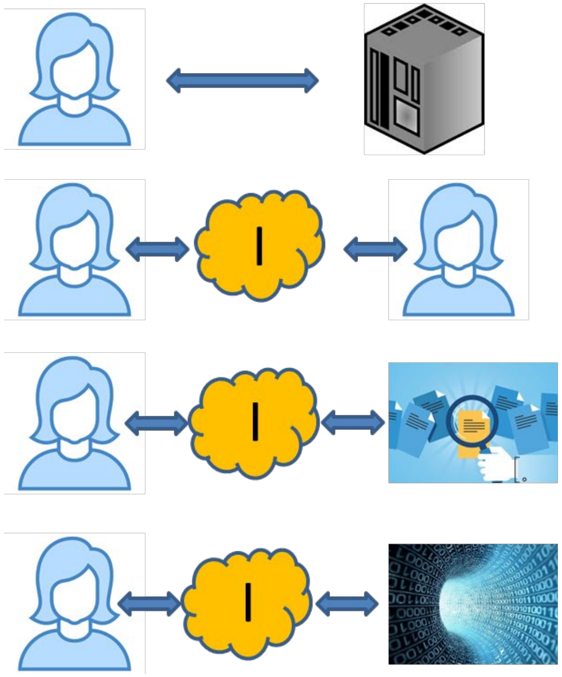


1 Client resolves PID to get current state data, minimally incl. network location.

2 Client sends DOIP request to relevant repository.

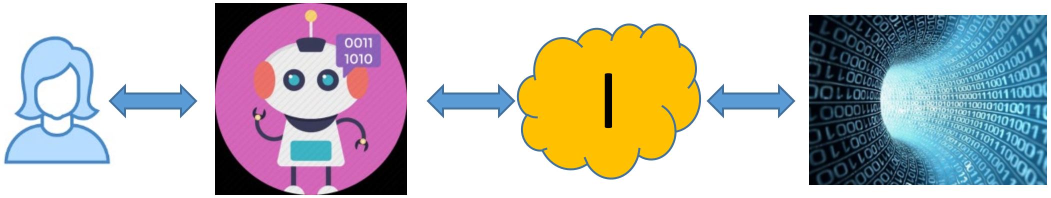
3 Repository finds or computes data to respond to client request.

Towards Automatic Processing



People to time-sharing computers	TTY
People to Files People to People	FTP + TCP/IP SMTP + TCP/IP
People to Info	HTML/HTTP + TCP/IP
People to Data	TCP/IP + ??

1. **We do not have a succifient approach wrt to phase 4.**
2. **What comes next given the data trends?**



Agent to Data
Data to Data