RESEARCH DATA ALLIANCE

From FAIR Principles to Blueprints for Data Architectures

Peter Wittenburg
Max Planck Computing and Data Facility

research data sharing without barriers
rd-alliance.org

MPCDF

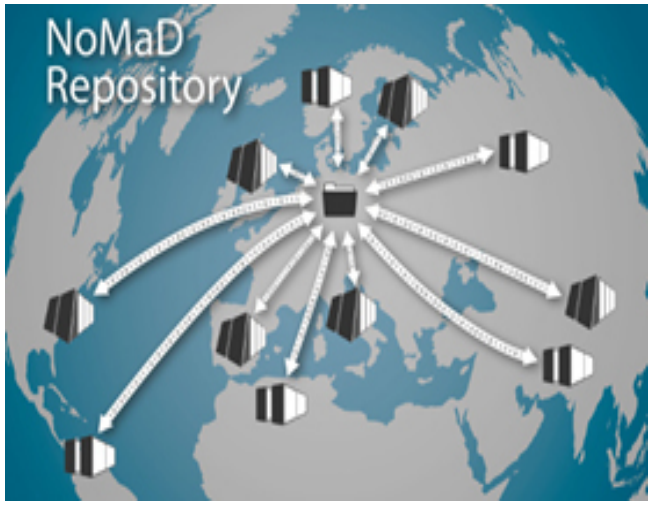
Trend in many disciplines

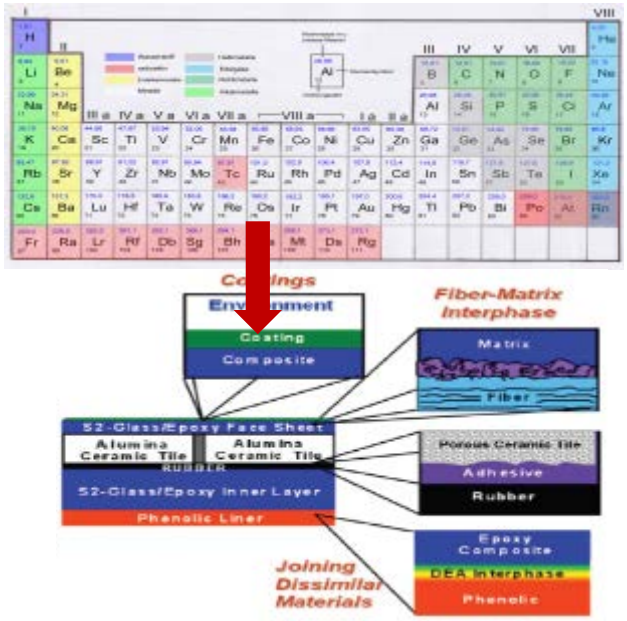
Researchers across disciplines are collecting increasingly large and complex data sets to extract knowledge.

Data Sharing and Re-use is common practice.

just 3 examples ...

NoMaD infrastructure





- Novel Materials Discovery project
  Computational material science
- many Labs create data about materials and compounds (experiments + simulations)
  - **space of Chemical compounds is endless**
  - **how can we categorise space to quickly find useful compound materials**
  - **from Periodical system to multi-dimensional map of compound material**
- categorisation via Machine Learning etc.
- **required is the integration of data from many labs worldwide which is time consuming**

RESEARCH DATA ALLIANCE

DOBES project on endangered languages

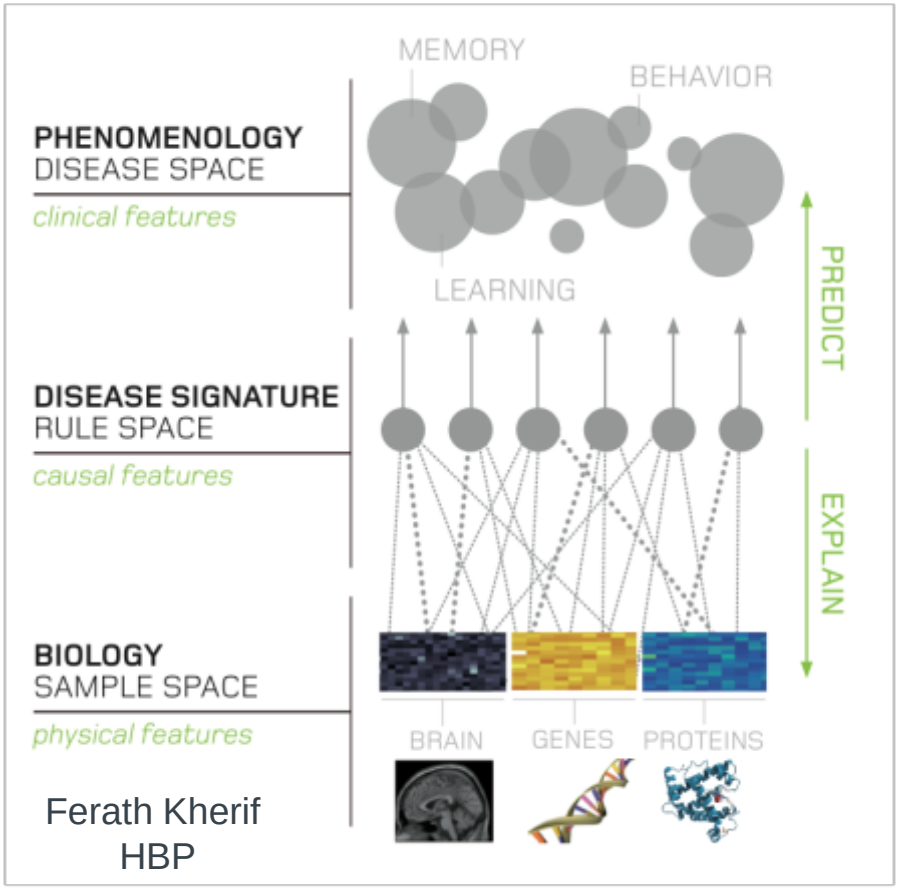
- ~70 global teams
- One central archive
- ~80 TB in online archive
- 4 dynamic external copies
- remote archives

New questions can be addressed:

- how can one use data to validate theories about the evolution of languages (and cultures) over thousands of years
- how to understand which languages are more "economic" than others

- **also here: Integration of much data from many teams worldwide**

Understanding the human brain



- increase of brain diseases
- how can we detect their causal basis, how to detect them early, how to medicate them?
- machine learning allows to correlate patterns in data (brain images, genes, proteins, reactions, etc.) with phenomena

- but: much sensitive data from various specialized labs and hospitals is required

Can we simply continue?

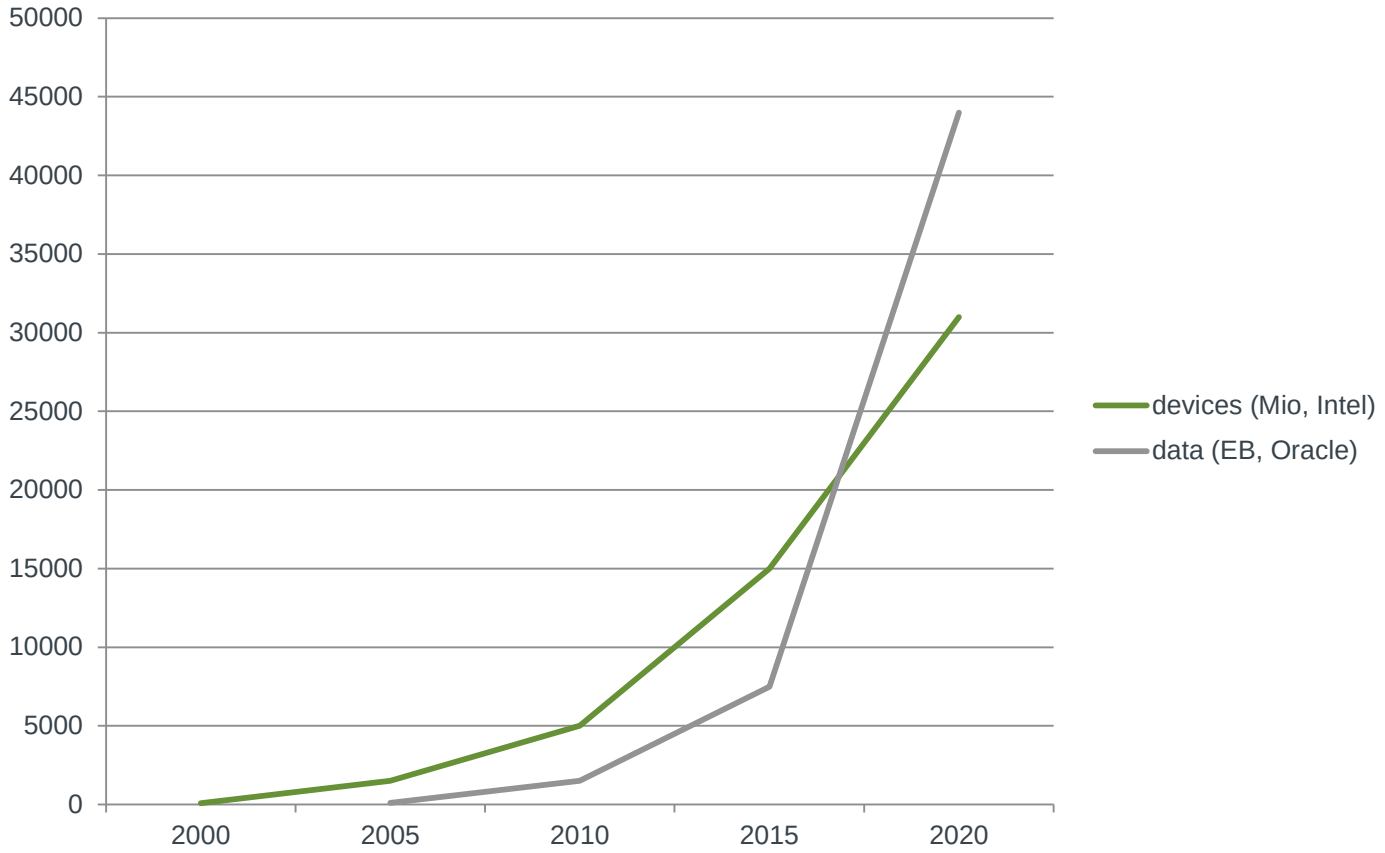
Noooooo, because ...

- ☐ our data landscape is fragmented - only little fits together
  (Identification, organization and description of data, storage systems, etc.)
- ☐ in industry 60% of costs are devoted to data integration
- ☐ 80% of all created data no longer accessible after short time periods
- ☐ 80% of the time of expensive data scientists is wasted on typical data management tasks
- ☐ data volumes and complexity will increase extremely due to new developments (in science and industry)

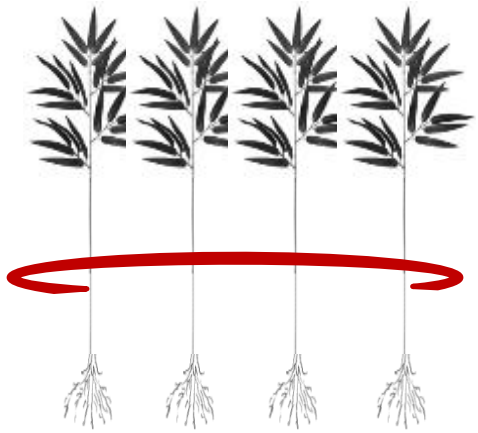
50 billion Smart Devices will create true data monsters.

- ☐ we are not fit for this new phase! (one of the reasons for RDA)

Development of devices and data



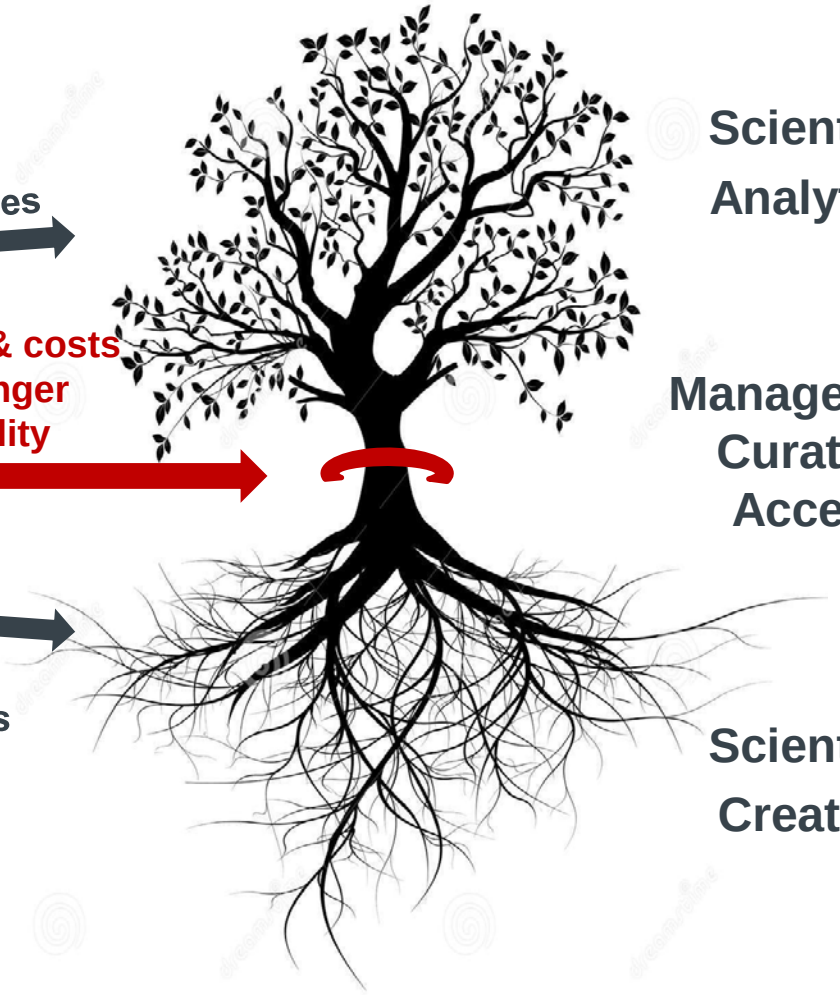
Assumption & Hope



Leave flexibility
Even more opportunities

Reduce heterogeneity & costs
Make solutions stronger
Achieve sustainability

Leave flexibility
Even more opportunities



Scientific Analytics

Management
Curation
Access

Scientific Creation

PID, AAI, MD, WF, Registries, Repositories, meta-semantics, etc.

Coming to agreed principles

Pre-ICRI Meeting Copenhagen March 2012	G8 Data Group June 2013	Data Foundation & Terminology Sept 2013	FAIR Principles Summer July 2014
discovery	discovery	store data in	findable
access	access	trustworthy	accessible
interpretation	re-usable	repositories	interoperable
re-use	manageable	assign PIDs	re-usable
		assign MD	

increasing convergence & explicitness

Layers to work with “Digital Objects”

Scientists, Data Curators, End Users, Applications

Enabling Technologies

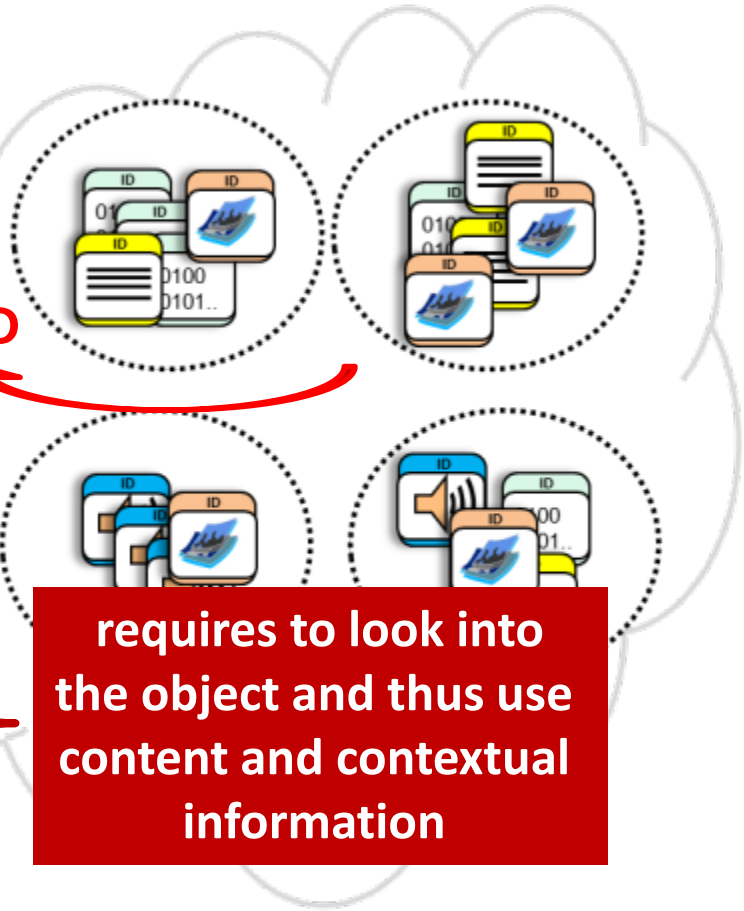
Discovery

Access (ref. resolution, protocols, AAI)

Interpretation

Reuse

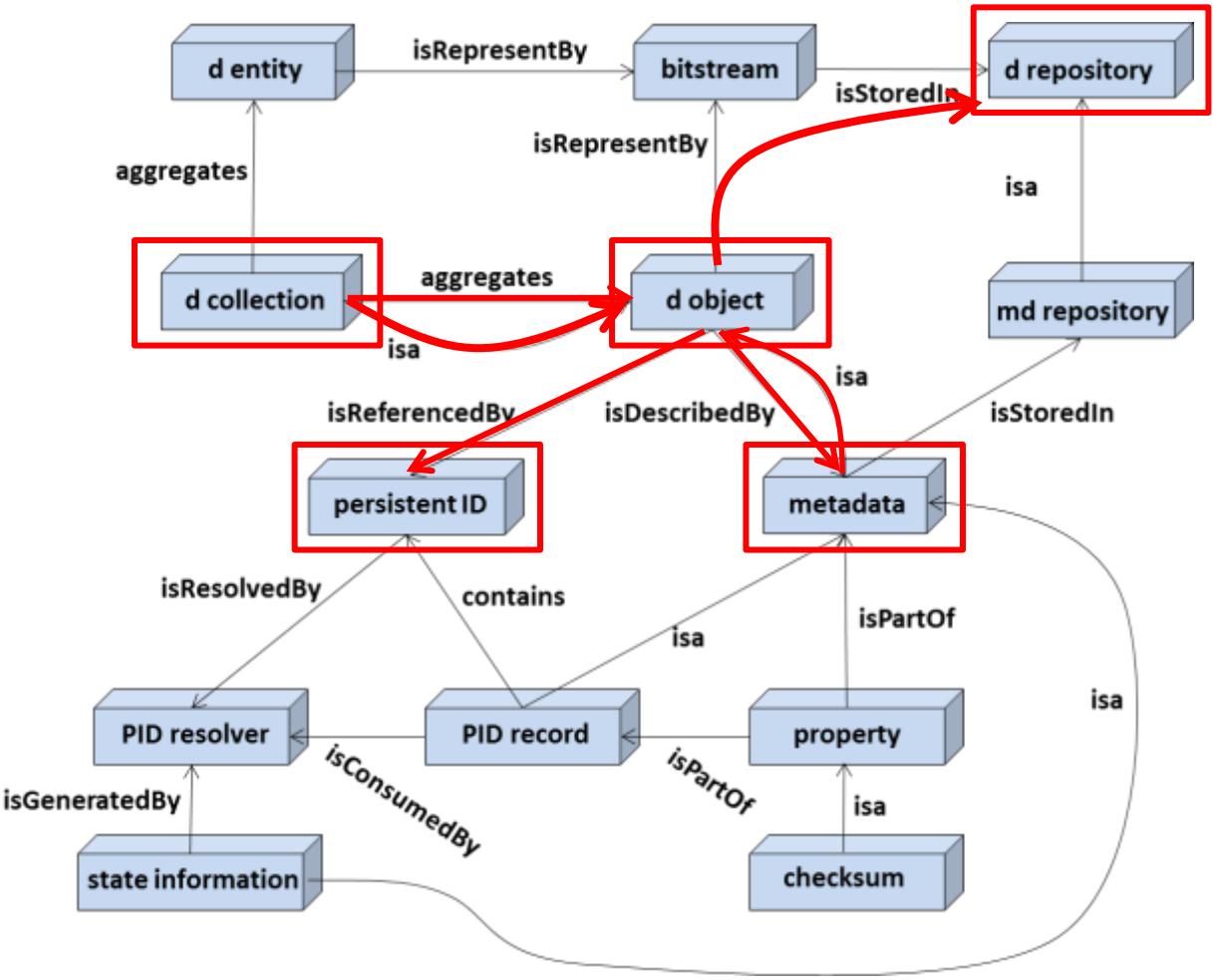
PID



requires to look into the object and thus use content and contextual information

taken from Larry Lannom

RDA DFT – simple powerful data model



Data Foundation and Terminology

Core model is very simple.

If all software developers would implement this model, we would get an enormous increase in efficiency.

Deviations can become very expensive.

FAIR principles

To be Findable:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.

F2. data are described with rich metadata.

F3. (meta)data are registered or indexed in a searchable resource.

F4. metadata specify the data identifier.

To be Accessible:

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.

To be Interoperable:

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.

To be Re-usable:

R1. meta(data) have a plurality of accurate and relevant attributes.

R1.1. (meta)data are released with a clear and accessible data usage license.

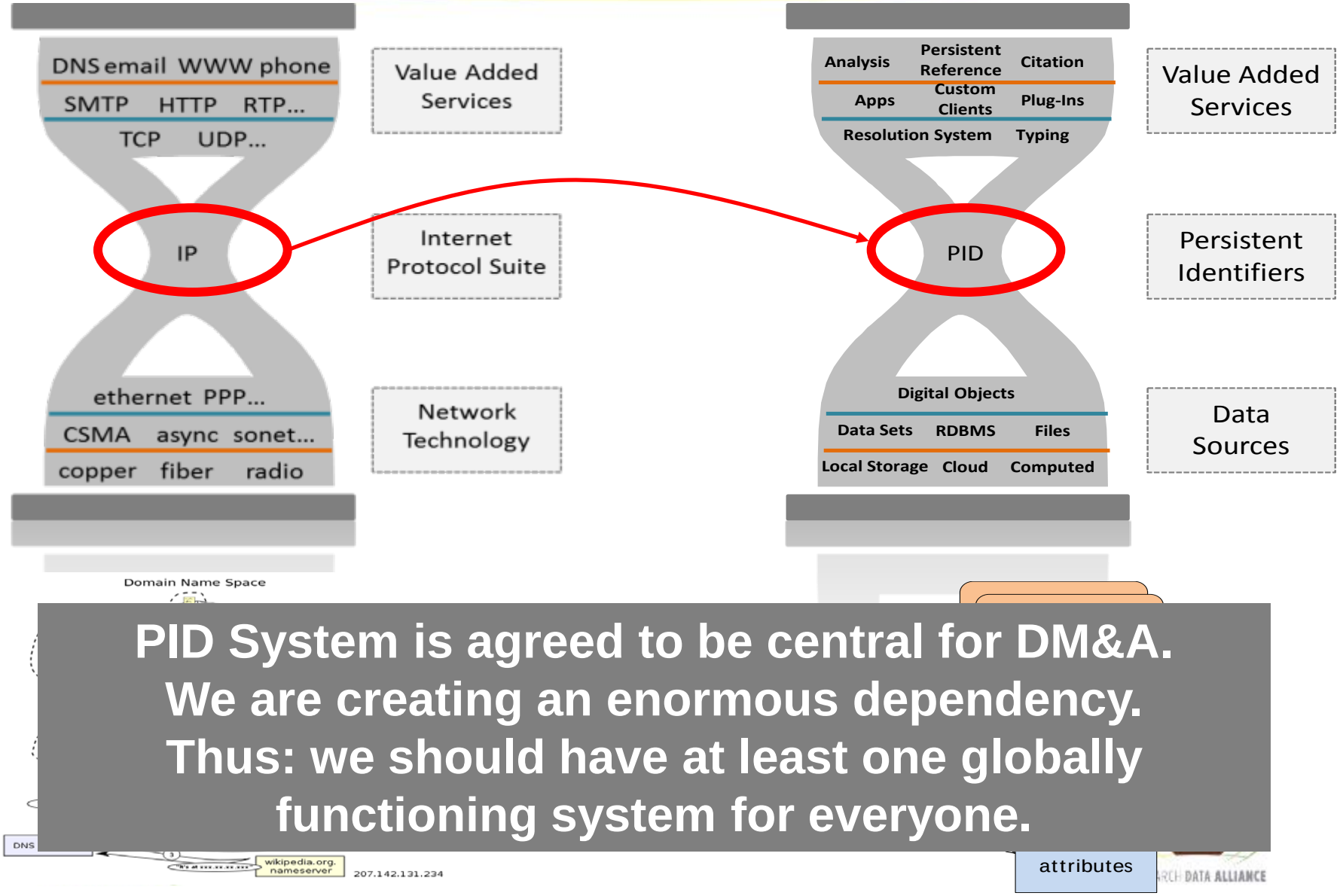
R1.2. (meta)data are associated with their provenance.

R1.3. (meta)data meet domain-relevant community standards.

From principles to usable components

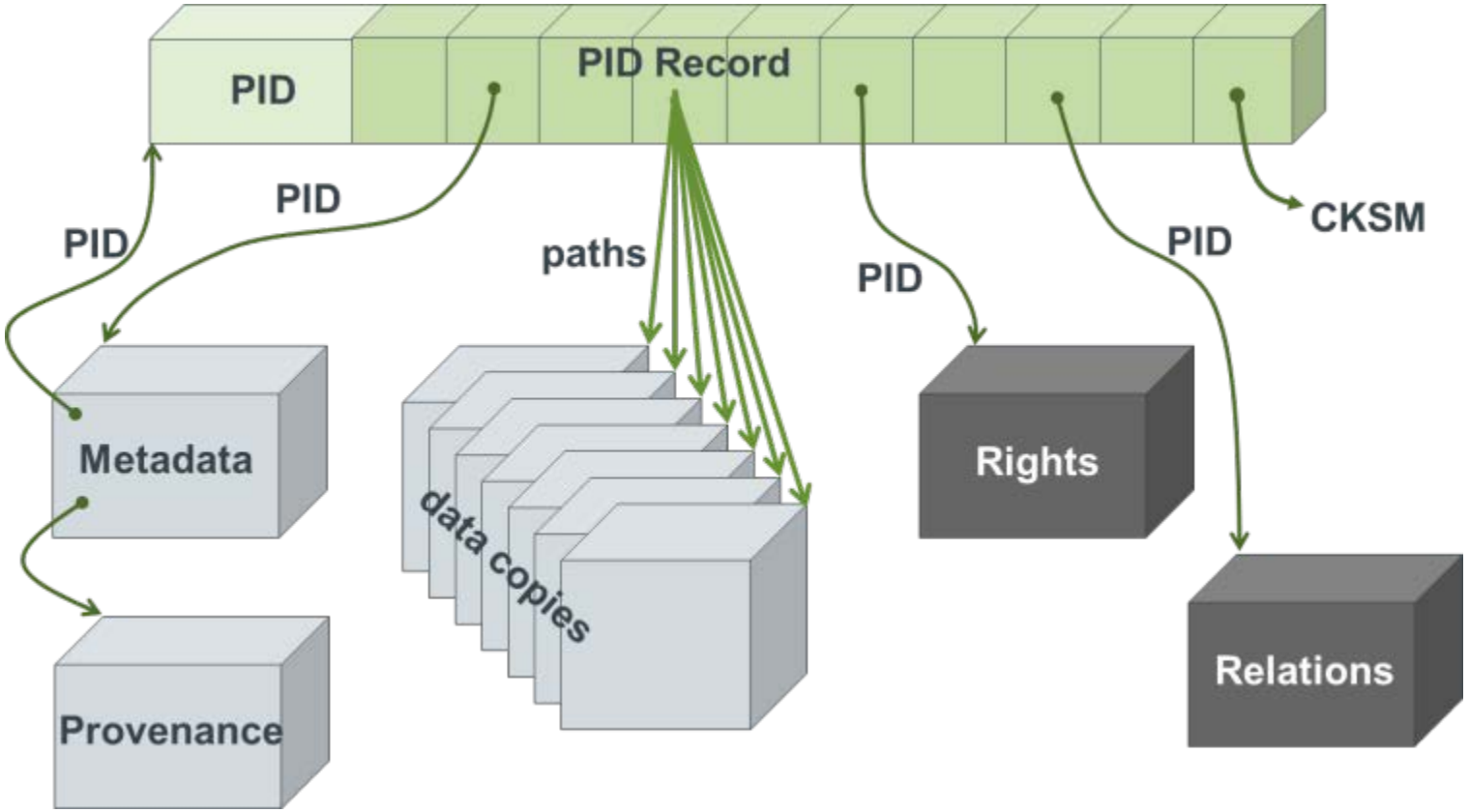
- agreed principles are so important to change minds and data practices
- they are not sufficient to reduce the huge solutions space and thus fragmentation (social and technological aspects)
- CODATA, W3C, RDA, DONA, etc:
  - organise cross-border (disciplines, countries, projects) interaction platforms to define
    - policies
    - components, interfaces, procedures
- Let‘s have a look at one basic component - there are many others ...

Global und persistent IDs as anchors



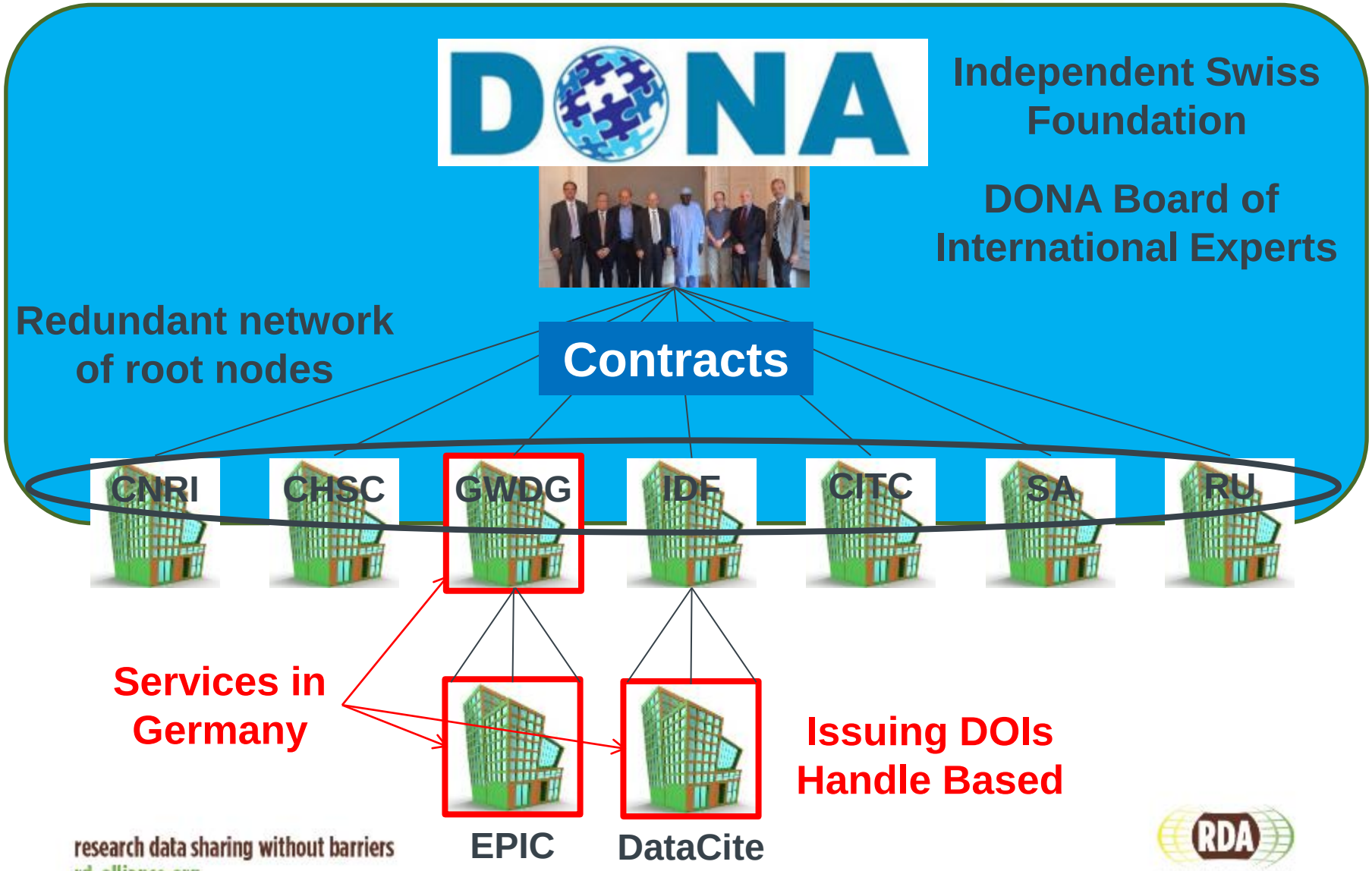
PID System is agreed to be central for DM&A.
We are creating an enormous dependency.
Thus: we should have at least one globally functioning system for everyone.

PID System as key component

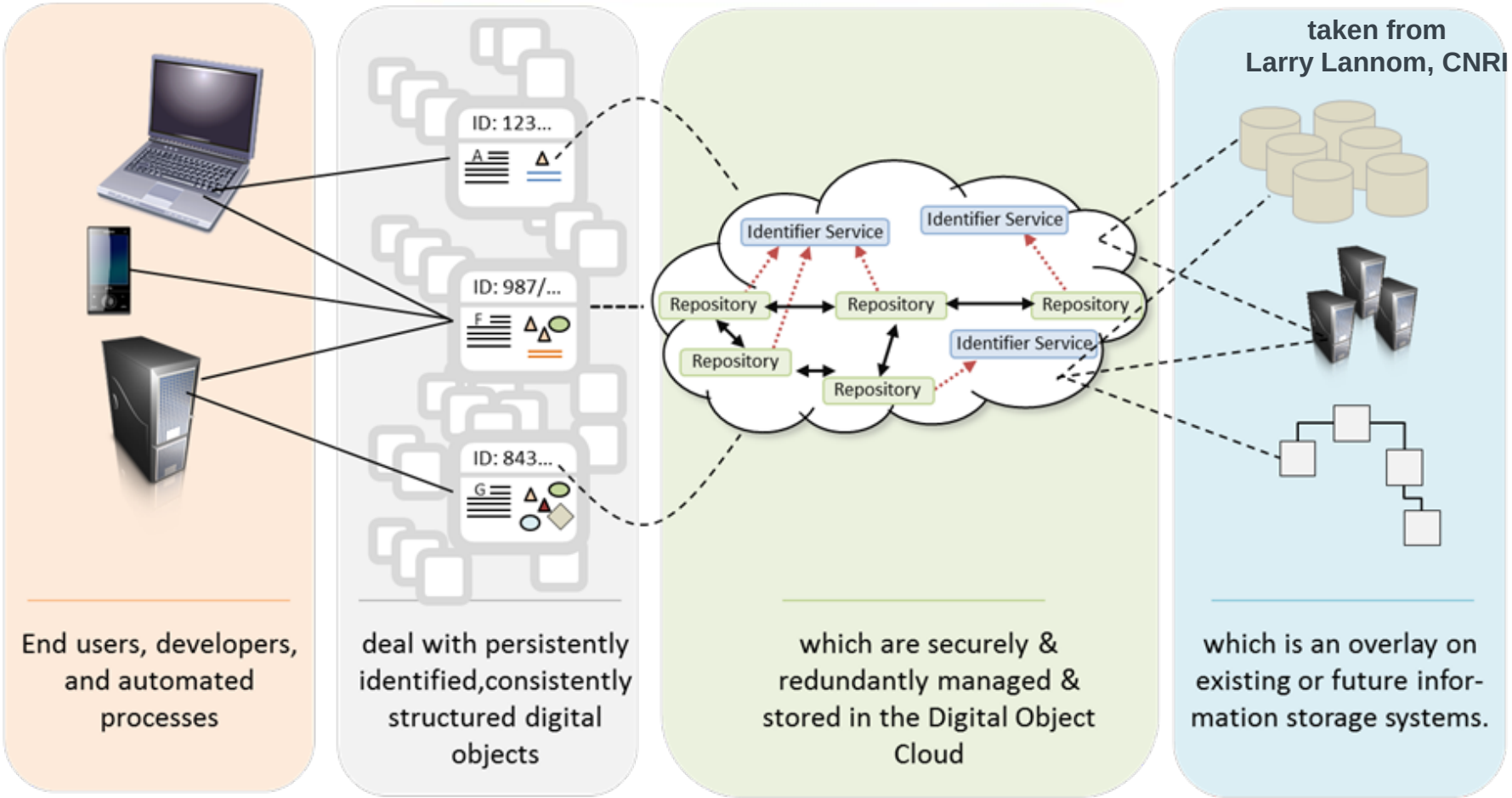


If we rely on a PID system as persistent, let's add relevant information with it.

Worldwide Handle System



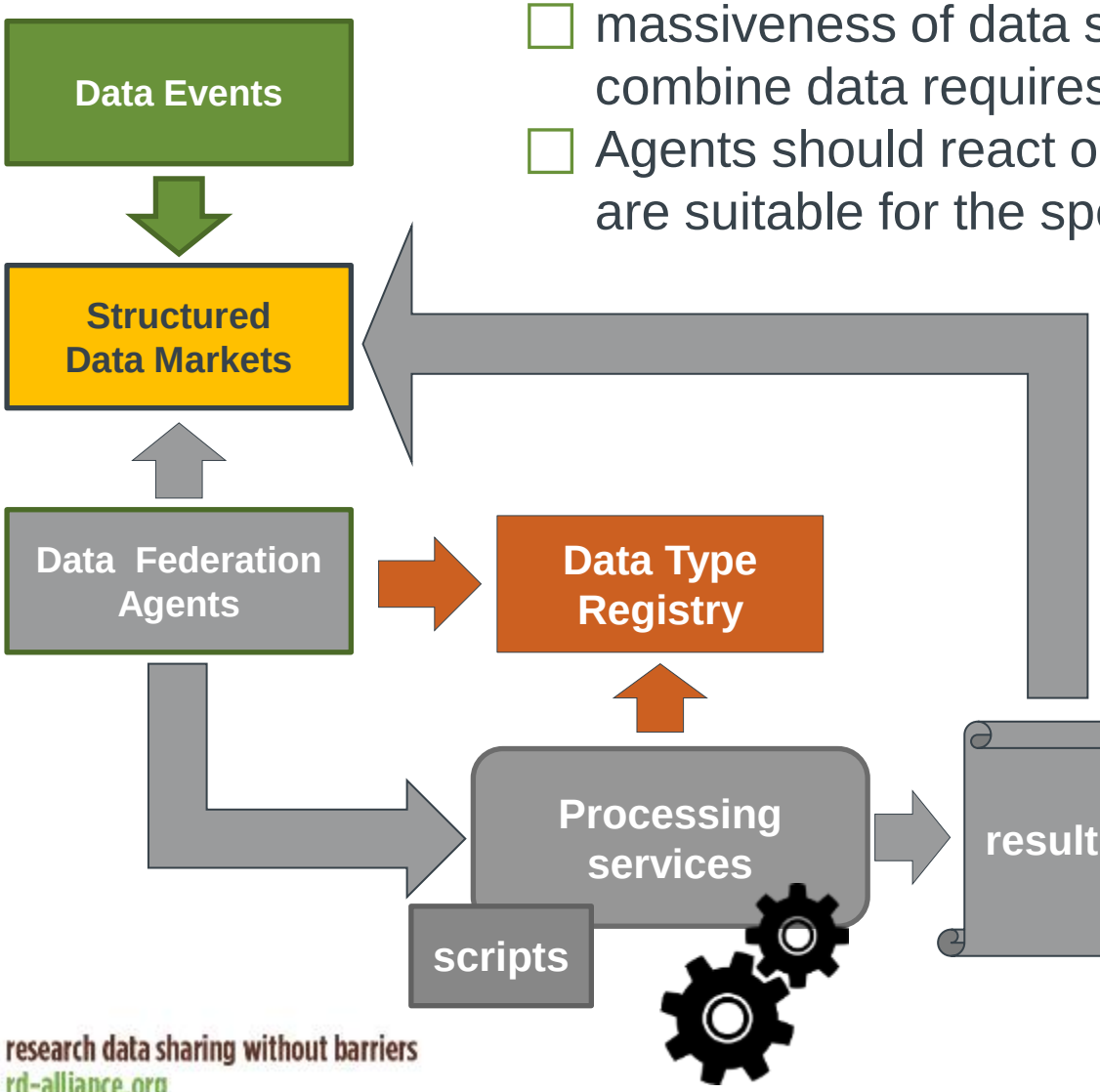
Towards a Global Digital Object Cloud



- if we rely on a PID system we can dream about global virtualisation
- some (climate modelling community) already work on implementing GDOC elements

Towards Type-Triggered 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



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

To do list

- disseminate FAIR principles and follow them
- apply simple DFT Core Model in software
- make use of unifying components where possible to reduce the fragmentation and solutions space
- participate in interaction platforms (RDA, GEDE, etc.)
  - follow state-of-the-art
  - participate in specifying requirements and designing components, i.e. become active in working and interest groups
- train a new generation of experts

References

RDA Global: http://rd-alliance.org
RDA Data Fabric IG: https://www.rd-alliance.org/group/data-fabric-ig.html
GEDE Group: https://www.rd-alliance.org/groups/gede-group-european-data-experts-rda

RDA Deutschland: annual meeting in November in Potsdam

RDA Plenary P9: 5-7.4 2017 Barcelona
RDA Plenary P10: September 2017 Montreal
RDA Plenary P11: April 2018 Berlin (?)

Thanks for your attention.