RDA - FAIR – and now?

Peter Wittenburg

Max Planck Society, Max Planck Computing & Data Facility

RDA co-founder

- Is there a problem in data-driven science?
- What is RDA?
- What are typical results of RDA?
- How further?

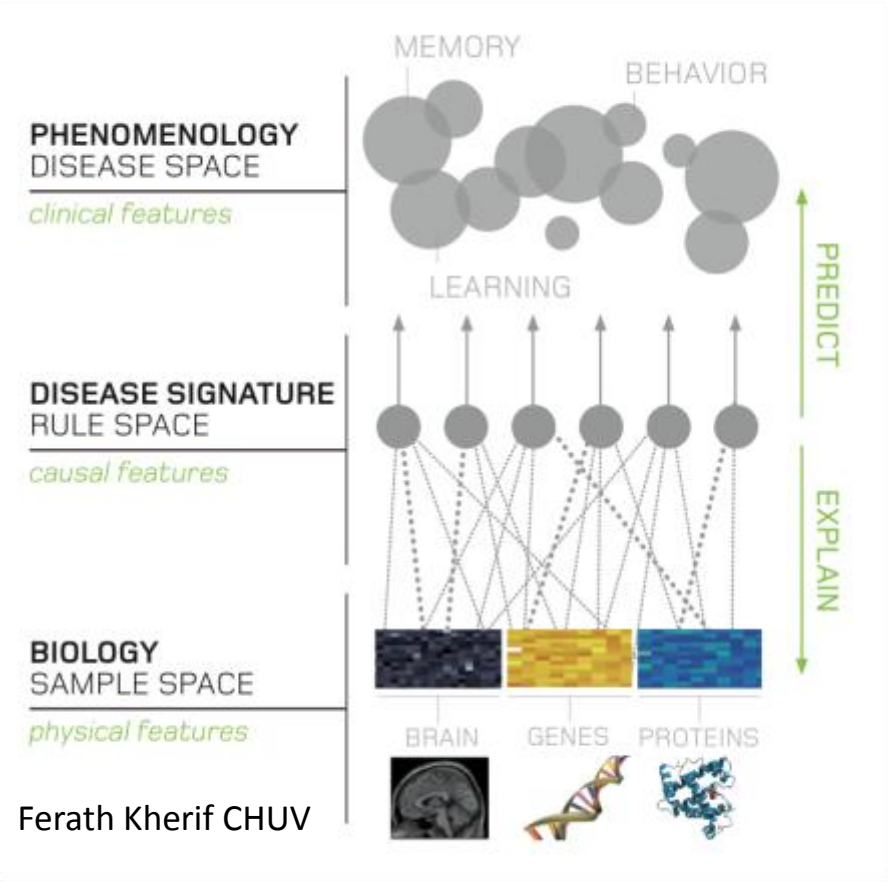
DOBES – Humanities/Languages (2000-12)

http://dobes.mpi.nl/

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

- 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" compared to others
- ***Revolution in humanities: scientific paper is not only goal anymore – it's about repurposing data***

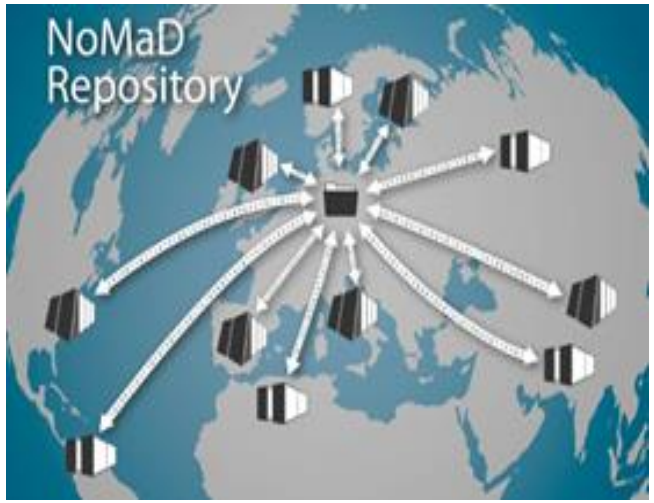
Brain Research – Detect Disease Patterns

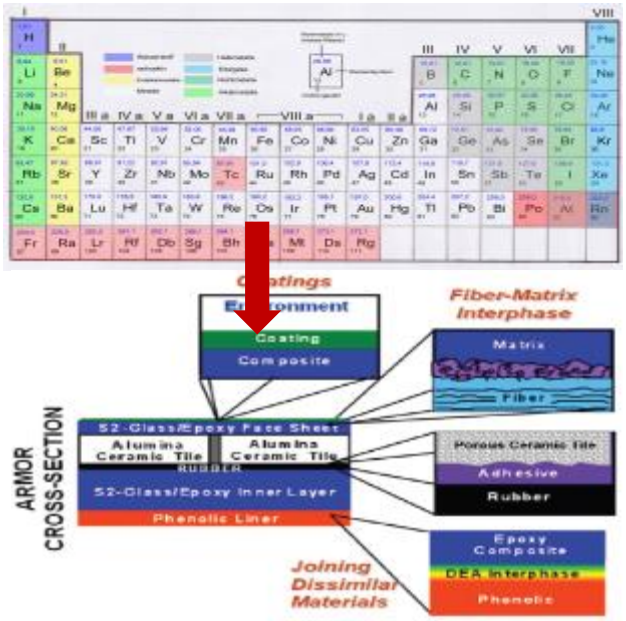


- early detection of causal basis of brain diseases
- machine learning to correlate patterns in data with phenomena
- much data from various specialized labs and hospitals is required

- ***Revolution in biomed world:***
  - ***sharing data outside of the hospitals for new purposes***
  - ***solving rights & ethical problems***

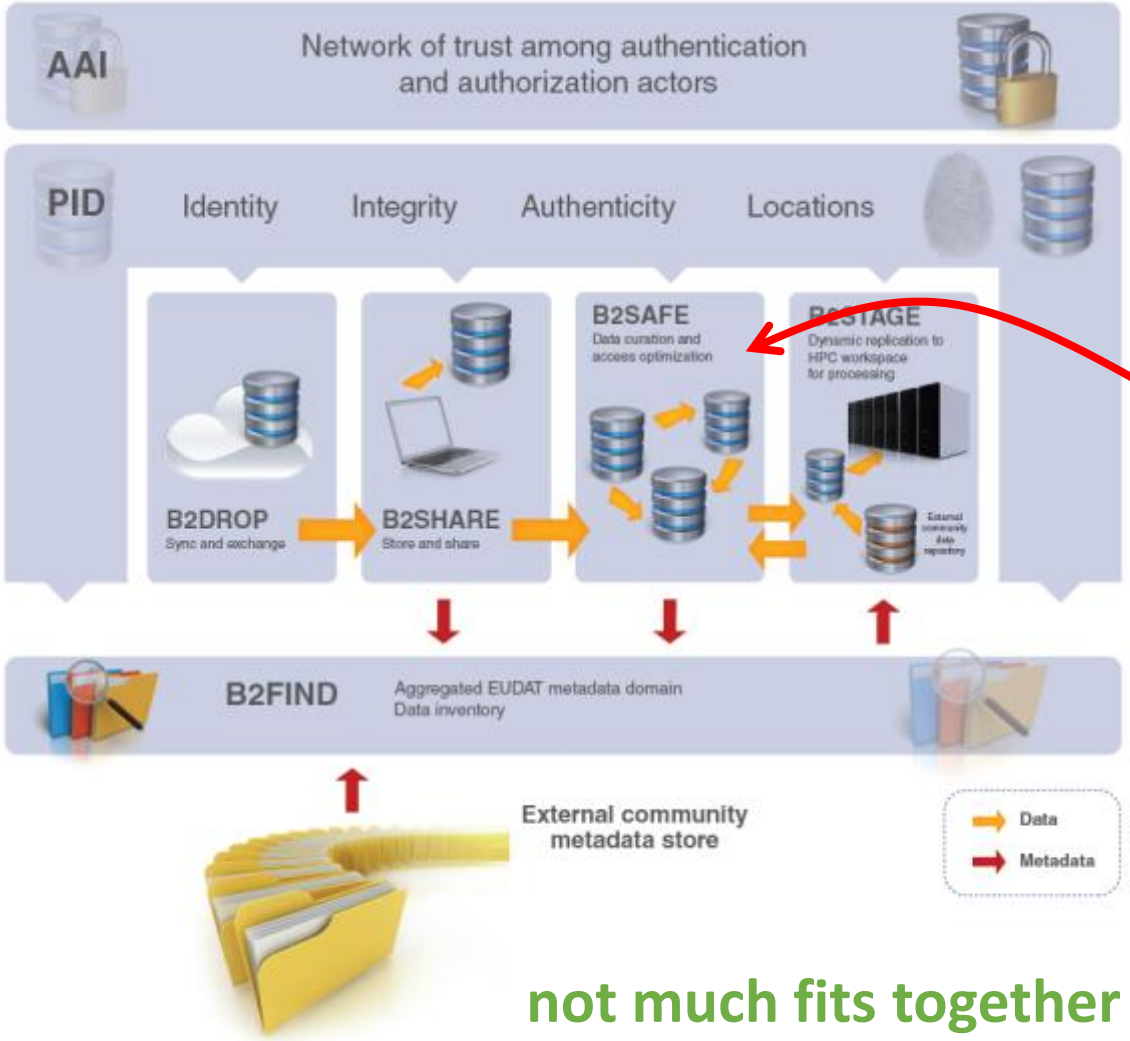
NoMaD - Material Science





- many Labs create data about materials and compounds (experiments + simulations)
  - space of chemical compounds is almost infinite
  - let's categorise this space to quickly find useful compound materials?
  - > 3 Mio aggregated entries now
- categorisation via Machine Learning etc.
- ***Revolution: writing paper is not the only scientific goal anymore – it's repurposing data***

EUDAT Data Infrastructure



- some data centres
- 5 initial communities in the driving seat
- definition of services and task forces to build them

- **B2SAFE** was and is difficult
  - no single API would do to easily replicate data
  - data organisations are different
  - metadata semantics not ready for machine processing
  - etc.

not much fits together yet

did not change data practices in the labs

Is there a problem?

- data intensive projects and large data aggregations are a fact
- but ...
  - data is hardly visible and not accessible (only 18% of data in registered repositories is accessible)
  - data domain is fragmented – data integration is a costly job (identification, organisation and description of data, etc.)
  - 80% of created data is not accessible any longer after short periods
  - 75-79-80% (RDA EU, CrouwFlowers, MIT) of data scientists' time is lost with data finding/integration/management work (data wrangling)
  - sorting out rights as a never ending story
- much work cannot be done, many are excluded

Will IoT change the game?

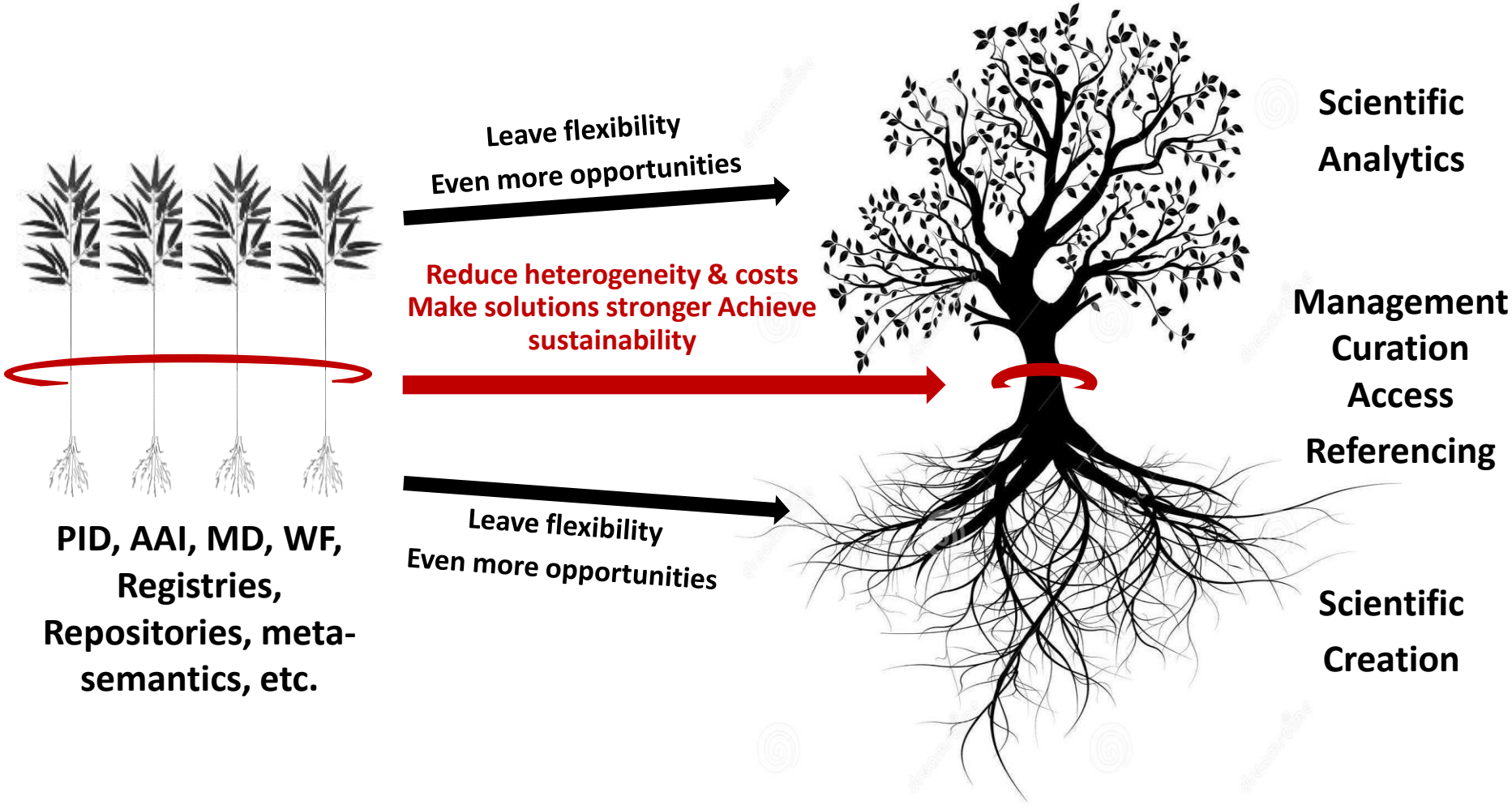
- 50 billion smart devices (Intel) will create true data monsters

- continuous streams with high-granularity
- optimisations and real-time decisions required
- much more re-purposing of data in various contexts

Are we fit for these challenges?
No our methods are not scalable - we need to change!

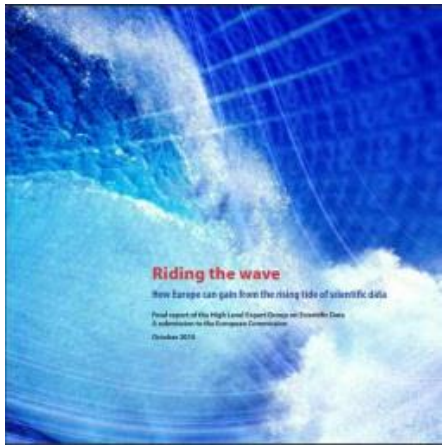
- Is there a problem in data-driven science?
- What is RDA?
- What are typical results of RDA?
- How further?

Fundamental Observation



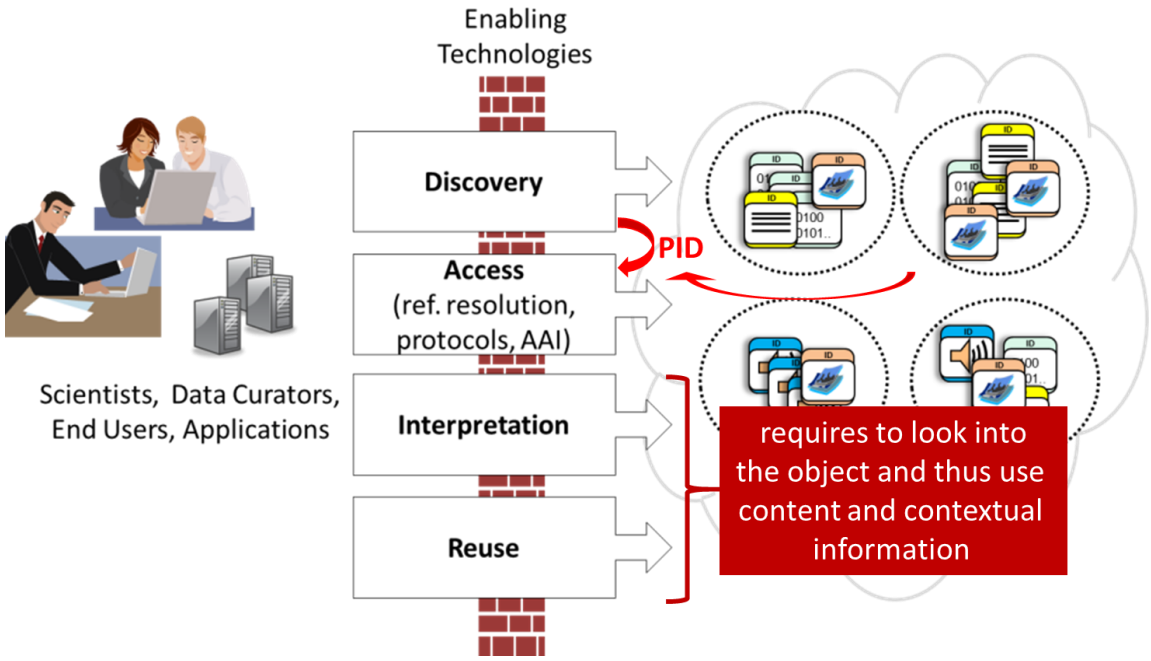
from 1000s of silo solutions → towards tree with stable trunk

Started at ICRI 2012 Copenhagen



- 75 data experts from 15 countries at DAITF workshop
- discussed the needs of global and cross-disciplinary agreements for the data domain

key was Larry Lannom's layer presentation



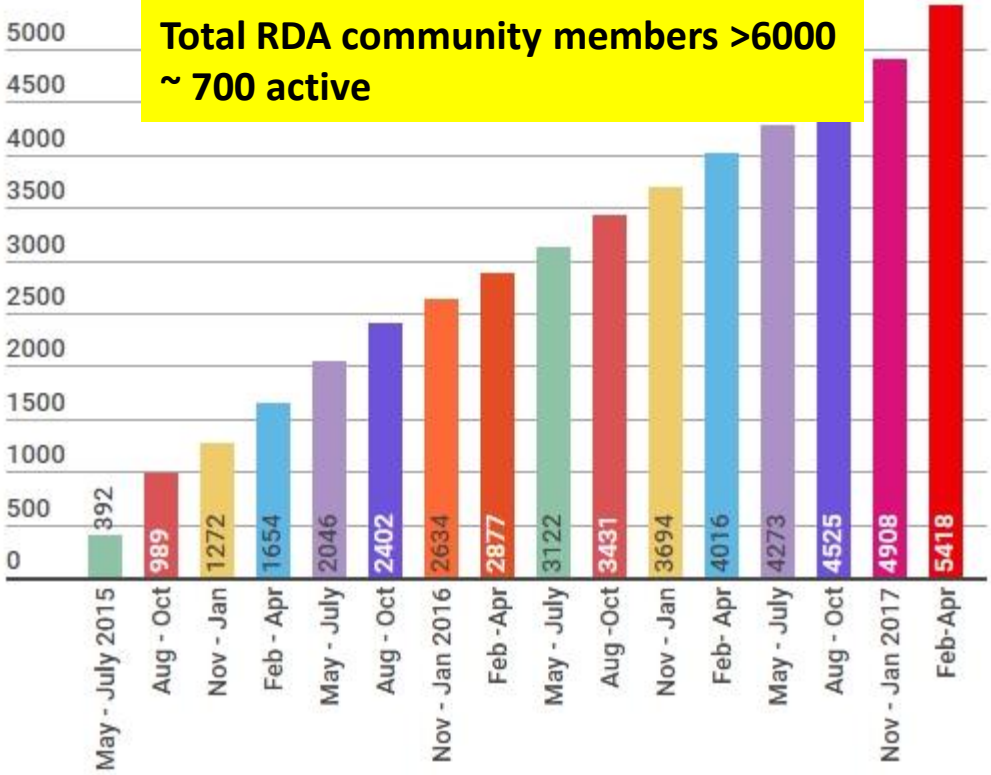
Official RDA start in March 2013

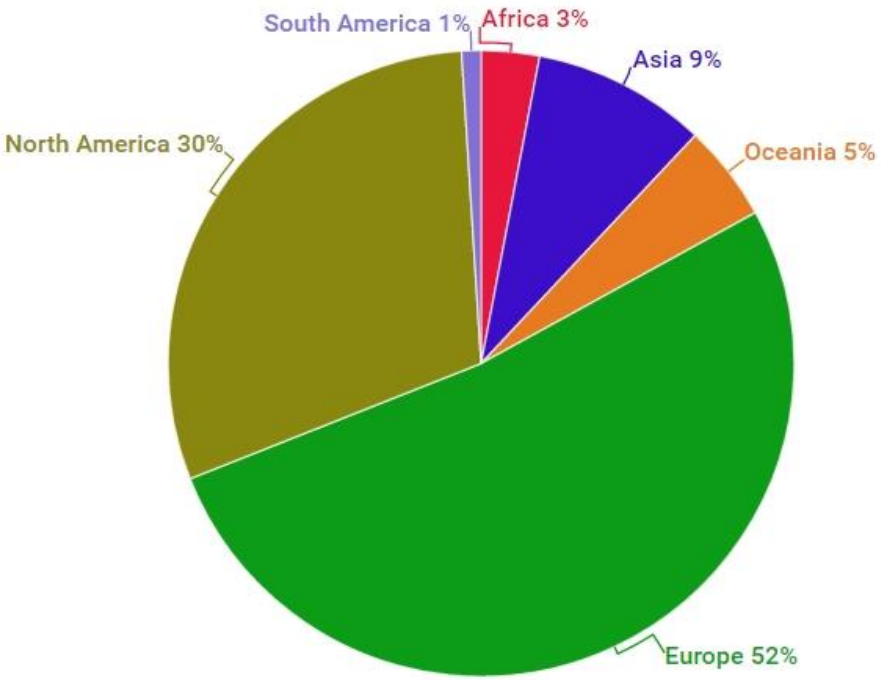
- March 2013 P1 in Gothenburg (2 plenaries per year)
  - Washington, Dublin, Amsterdam, San Diego, Paris, Tokio, Denver, Barcelona, Montreal, Berlin
- agreed on 6 basic principles for work and participation

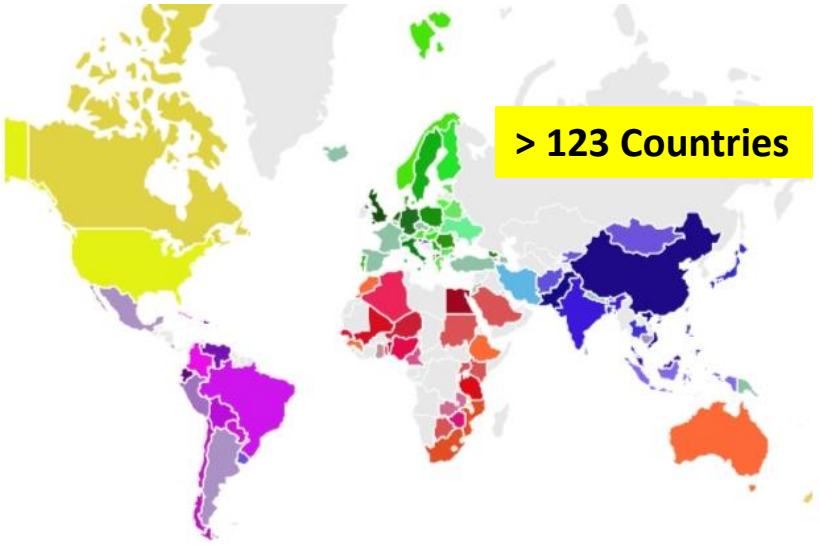
  openness, consensus, balance, harmonisation, community-driven, non-profit
- agreed on grass-roots approach with WGs and IGs instead of reference model/architecture
  - some people doubt whether this can lead to success
  - looks indeed somewhat chaotic, but ...
  - need to accept that some outputs will not be taken up
  - don’t we know this from IETF RFCs

RDA worldwide

Total RDA community members >6000
~ 700 active







> 123 Countries

- practitioners from many disciplines & sectors
- simple governance (WG/IGs, Council, TAB, OAB)
- get funds from NSF, EC, AU
- more countries ready to support us (SA, CA, CN, etc.)
- have >80 organisational members

Organisational & Affiliate Members

43 Organisational Members

8 Affiliate Members

RDA Interest (IG) & Working Groups (WG) by Focus (1)

Total 81 groups:
30 Working Groups & 51 Interest Groups

Domain Science - focused

- ❑ **Agrisemantics WG**
- ❑ BioSharing Registry WG
- ❑ **Fisheries Data Interoperability WG**
- ❑ On-Farm Data Sharing (OFDS) WG
- ❑ **Rice Data Interoperability WG**
- ❑ **Wheat Data Interoperability WG**
- ❑ **Agricultural Data IG (IGAD)**
- ❑ Biodiversity Data Integration IG
- ❑ Chemistry Research Data IG
- ❑ Digital Practices in History and Ethnography IG
- ❑ Geospatial IG
- ❑ Global Water Information IG
- ❑ Linguistics Data Interest Group
- ❑ Health Data IG
- ❑ Mapping the Landscape IG
- ❑ Marine Data Harmonization IG
- ❑ Quality of Urban Life IG
- ❑ RDA/CODATA Materials Data, Infrastructure & Interoperability IG
- ❑ Research data needs of the Photon and Neutron Science community IG
- ❑ Small Unmanned Aircraft Systems' Data IG
- ❑ Structural Biology IG
- ❑ Weather, Climate and air quality IG

Community Needs - focused

- ❑ Certification and Accreditation for Data Science Training and Education WG
- ❑ **RDA/CODATA Summer Schools in Data Science and Cloud Computing in the Developing World WG**
- ❑ **Teaching TDM on Education and Skill Development WG**
- ❑ Archives & Records Professionals for Research Data IG
- ❑ Data for Development IG
- ❑ Development of Cloud Computing Capacity and Education in Developing World Research IG
- ❑ Education and Training on handling of research data IG
- ❑ Ethics and Social Aspects of Data IG
- ❑ International Indigenous Data Sovereignty IG

RDA Interest (IG) & Working Groups (WG) by Focus (2)

Data Stewardship and Services – focused

- **Brokering Framework WG**
- WDS/RDA Assessment of Data Fitness for Use WG
- RDA / WDS Publishing Data Workflows WG
- Active Data Management Plans IG
- Data in Context IG
- Data Rescue IG
- Data Versioning IG
- Domain Repositories IG
- Libraries for Research Data IG
- Long tail of research data IG
- Preservation e-Infrastructure IG
- Preservation Tools, Techniques, and Policies IG
- RDA/WDS Certification of Digital Repositories IG
- RDA/WDS Publishing Data Cost Recovery for Data Centres IG
- Repository Platforms for Research Data IG
- Research Data Provenance IG
- Virtual Research Environments IG

Base Infrastructure – focused

- Array Database Assessment WG
- **Data Type Registries WG**
- **Metadata Standards Catalog WG**
- Metadata Standards Directory WG
- **PID Kernel Information WG**
- **Data Fabric IG**
- **Data Foundations and Terminology IG**
- Big Data IG
- Brokering IG
- Federated Identity Management IG
- Metadata IG
- PID IG
- Vocabulary Services IG

RDA Interest (IG) & Working Groups (WG) by Focus (3)

Reference and Sharing - focused

- **Data Citation WG**
- Data Description Registry Interoperability WG
- Data Security and Trust WG
- Empirical Humanities Metadata WG
- **Provenance Patterns WG**
- RDA / WDS Publishing Data Bibliometrics WG
- **Research Data Collections WG**
- QoS-DataLC Definitions WG
- International Materials Resource Registries WG
- National Data Services IG
- RDA/CODATA Legal Interoperability IG
- Reproducibility IG
- Data Discovery Paradigms IG
- **Repository Core Description WG**
- Research Data Repository Interoperability WG

Partnership Groups

- RDA / TDWG Metadata Standards for attribution of physical and digital collections stewardship WG
- RDA/NISO Privacy Implications of Research Data Sets IG
- RDA/WDS Scholarly Link Exchange Working Group
- RDA/WDS Publishing Data IG
- ELIXIR Bridging Force IG

will that ever converge?
is that relevant?
do we need reference architectures?

RDA Europe Project - RDA DE/UK/ES/FI/etc.

- Community Building
  - lots of interactions – bringing data professionals together
    - GEDE group – 47 ESFRI projects participating
    - eIRG Interactions with e-Infrastructures
  - many national meetings in Europe
    - RDA DE – community of about 200 data experts
  - discussing RDA results – stimulating new groups
  - doing training, hackathons, etc.

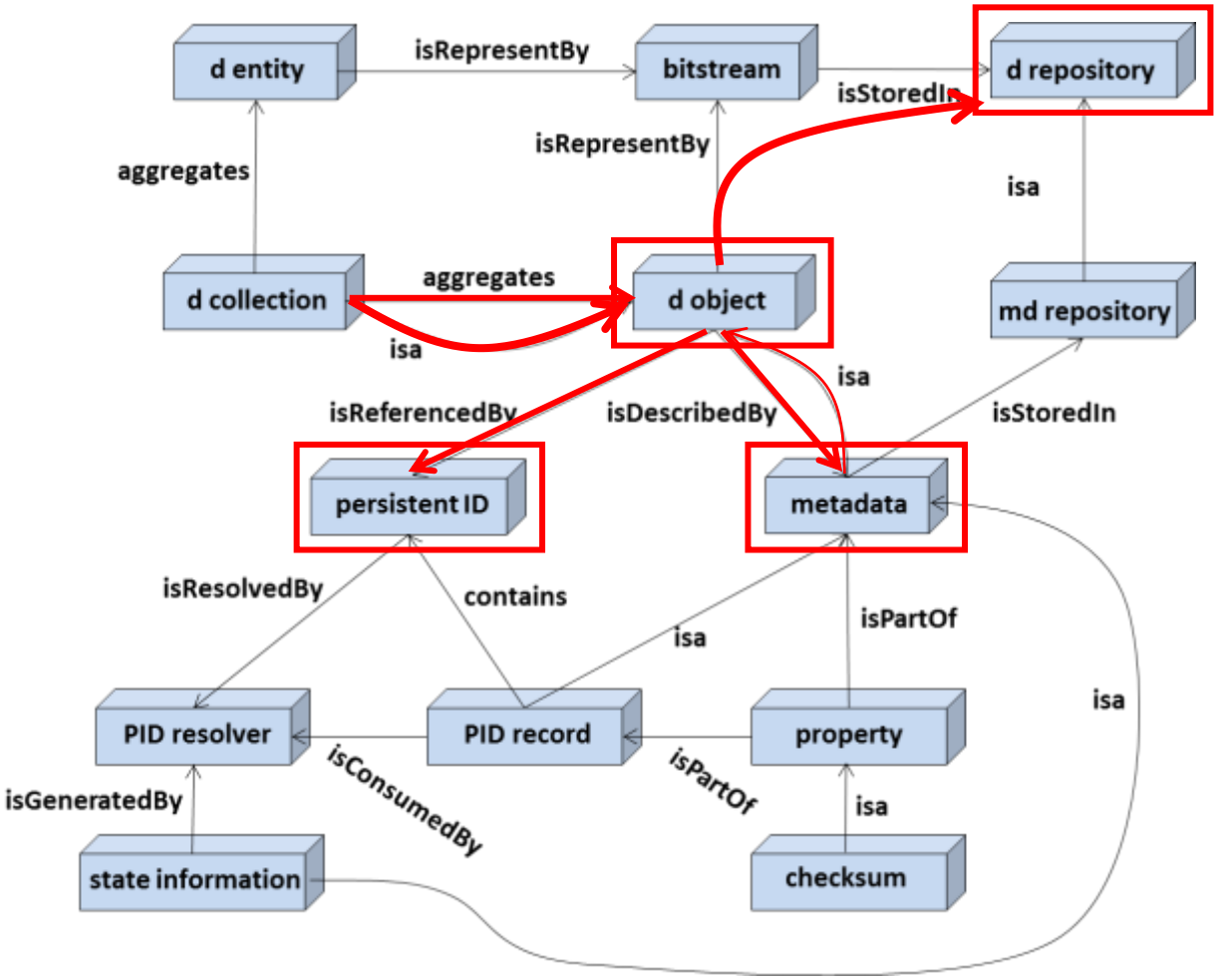
  - intensifying discussions with industry

1. RDA DE Mitgliederversammlung: 19.3. ab 17.00 TU Berlin

- 3 full days **RDA Plenary** (21-23 March)
- 2 days **Co-located and Associated Events** (19-20 March)
- 2 days **Industrial Side Event** (19-20 March) TU Berlin

- Is there a problem in data-driven science?
- What is RDA?
- What are typical results of RDA?
- How further?

RDA DFT – Simple Core Data Model

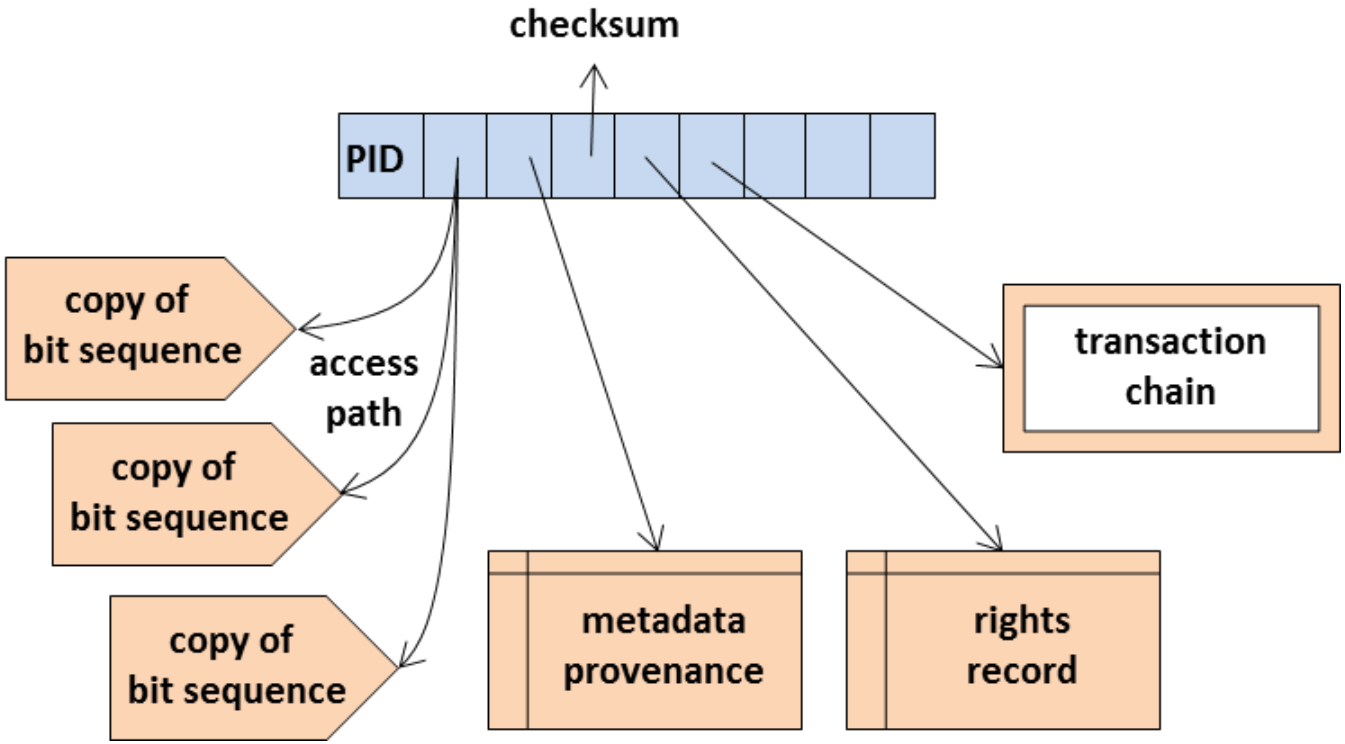


Data Foundation and Terminology WG

Core model is very simple.

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

PID as binding glue



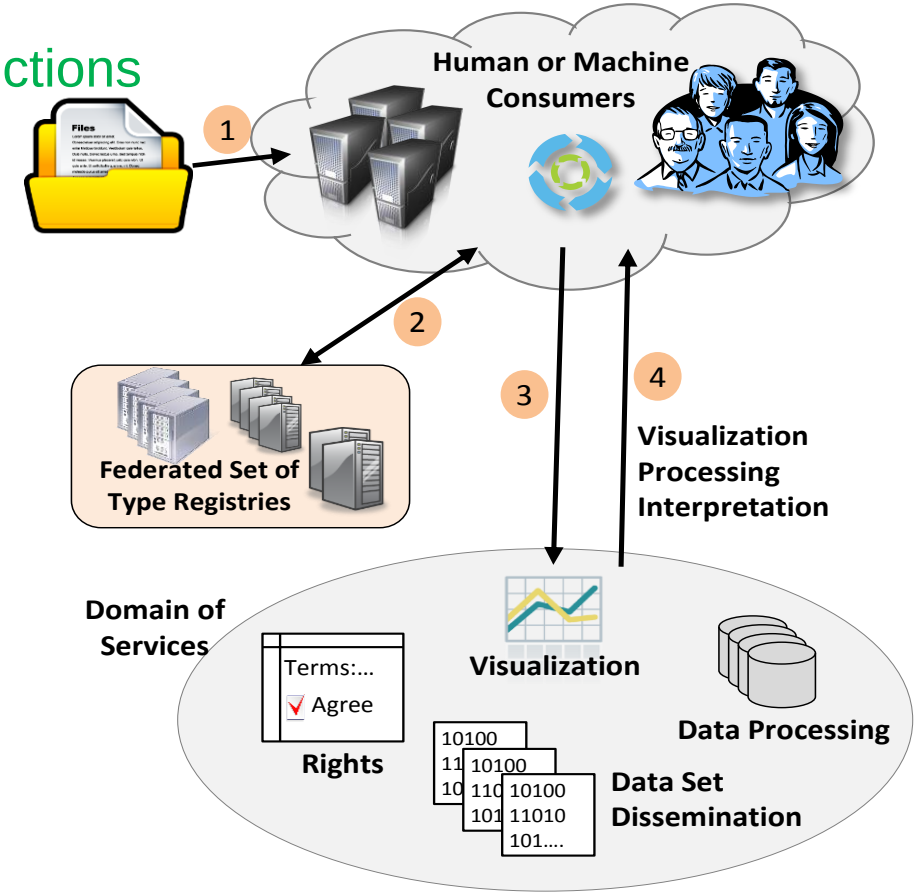
PID Information Type WG & PID Kernel Information WG

- specify principles of interoperability
- specify core types such as „checksum“ to allow machine interpretation
- technology is there (DONA Foundation for Handles & DOIs)

Data Type Registry

- result: a registry for data types
- Linking structure/semantics with functions
- you get an unknown file,

  pull it on DTR and content is being visualized
- You find a tag and know how to interpret
- no free lunch: someone needs to register and define type
- Various sciences make use of it



- Is there a problem in data-driven science?
- What is RDA?
- What are typical results of RDA?
- How further?

We all feel that something bigger will happen.

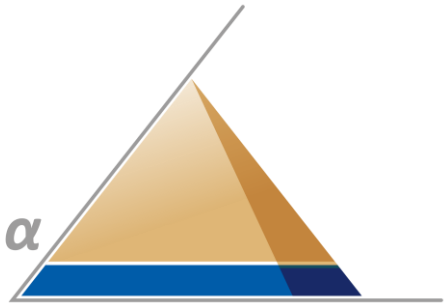
All what follows happened just recently.

Recent US Workshops: Impulse is Needed

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

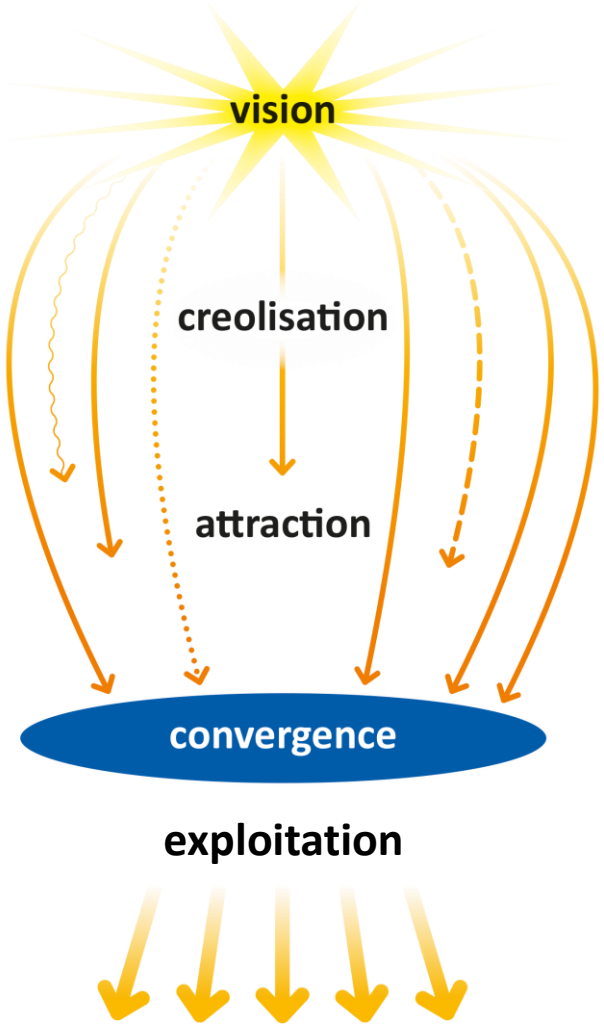
- Define a solid basis for future developments and heavy investments

J. Hendler (W3C): need a new commonality layer again to enable growth
G. Strawn (BRDI): seem to be at a point comparable to the moment where Internet was born
Pyramids: *get a few basic principles right and solve logistics systematically to build giants*



→ **C2CAMP**

Comparing revolutionary Infrastructures

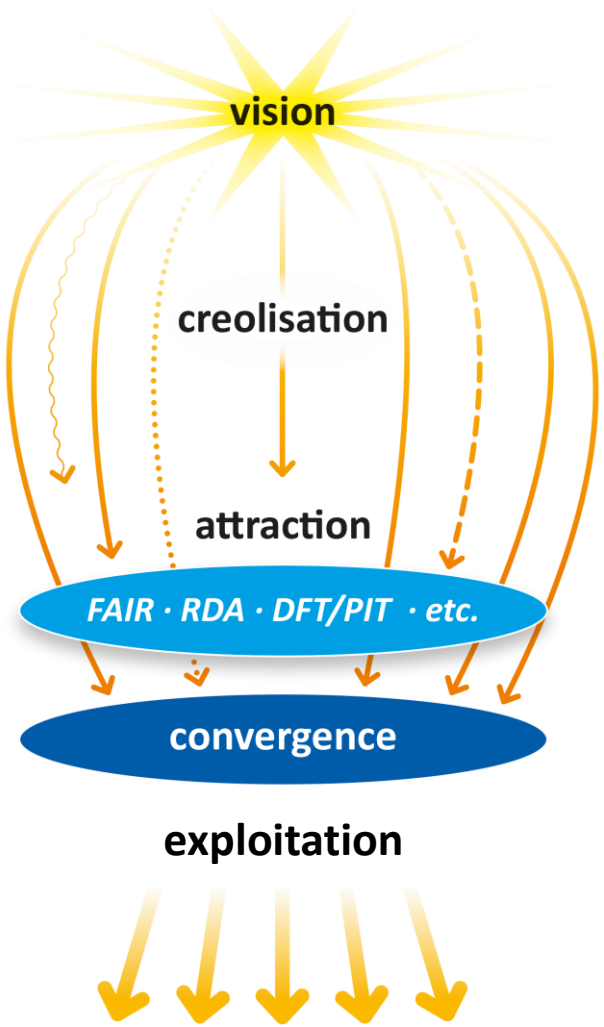


George Strawn (US Academy, BRDI, driver of Internet) & Myself (paper almost ready)

- Electrification
- Internet
- Web (slightly different)

- all share typical evolution dynamics
- all include disruptive elements
- all come to a covergence step defined by simple principles
  - ac, 110/220 V, 50/60 Hz
  - TCP/IP
  - HTTP/HTML (on top of TCP/IP)

Where are we with data?



Creolisation

- enormous solutions space – 1000 flowers
- and thus huge fragmentation

Attraction

- FAIR principles as a global and common language, but not a blueprint for infrastructure building
- grass-roots RDA to work out specifications of components (characteristics, interfaces)
- GOFAIR initiative to network implementers, trainers, etc.
- global C2CAMP initiative
  - building flexible testbeds (no architectures)
  - based on components specified by RDA, OAI, W3C, IETF, etc.)
  - based on the Digital Object concept

Convergence needed – but how?

FAIR Principles (FORCE 11)

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 comm 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:

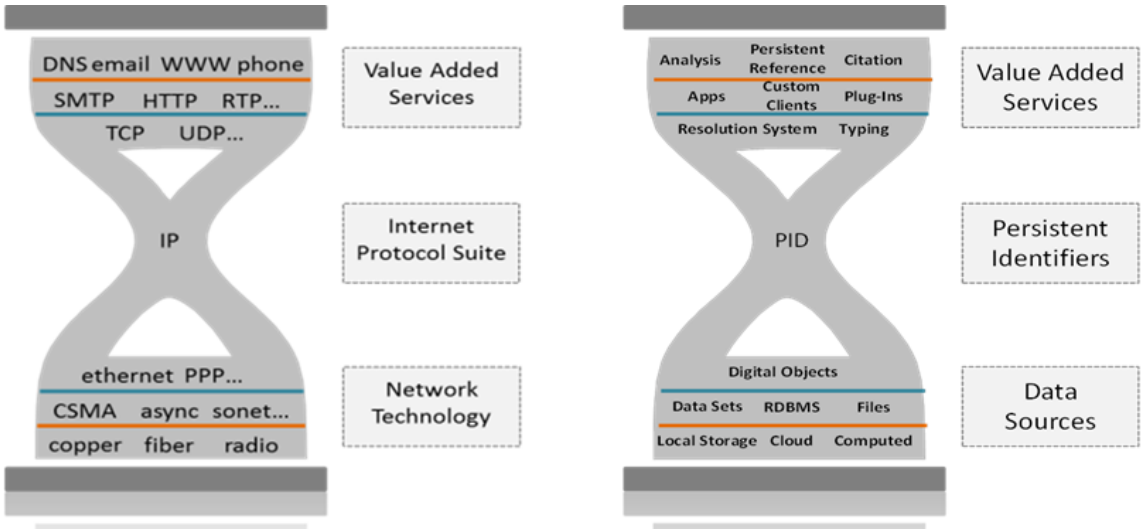
R1. meta(data) have a plurality of accurate and relevant attributes.

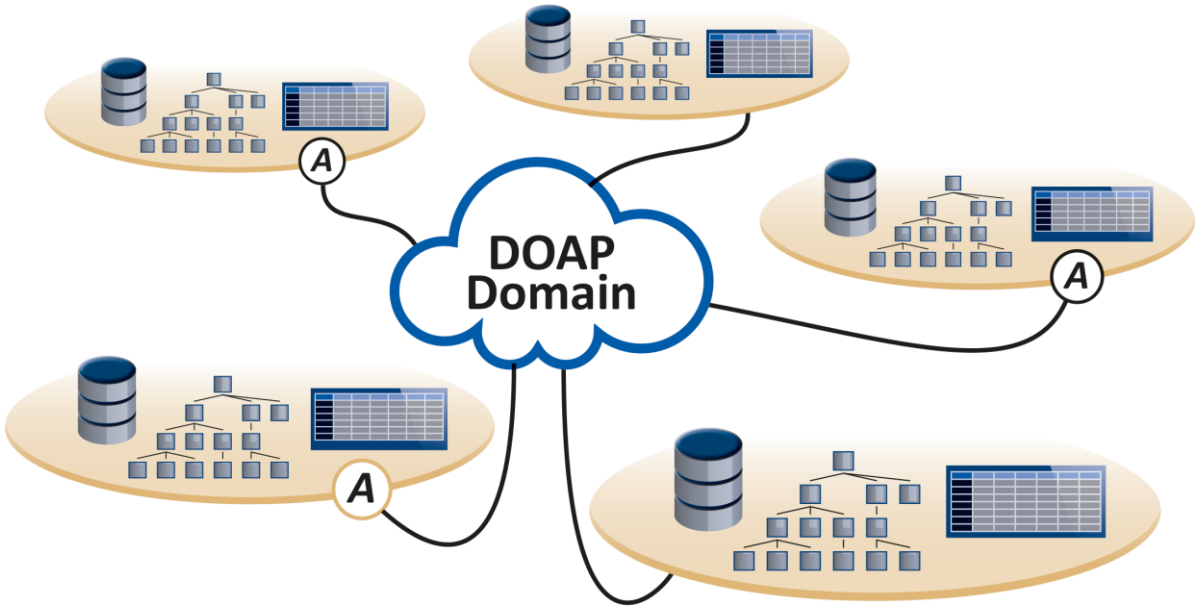
R1.1. (meta)data are released with a clear and accessible data usage license.

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

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

C2CAMP - Interoperability





some similarity to Internet

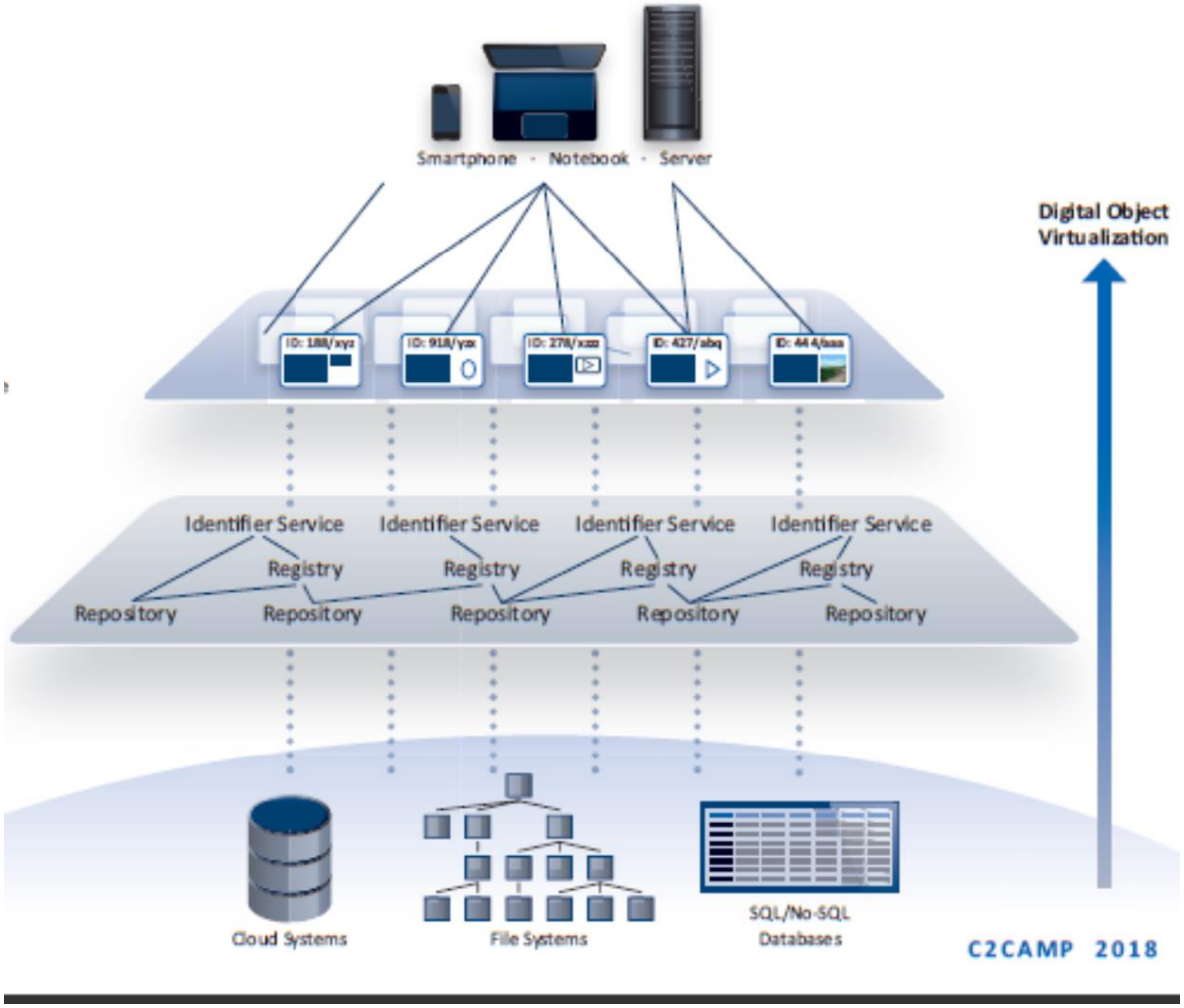
will it work?

what does it offer?

at least interoperability between repositories and data organisations, i.e. you find all relevant entities to be FAIR

not addressing Semantic interoperability, but facilitating

C2CAMP - DO Infrastructure

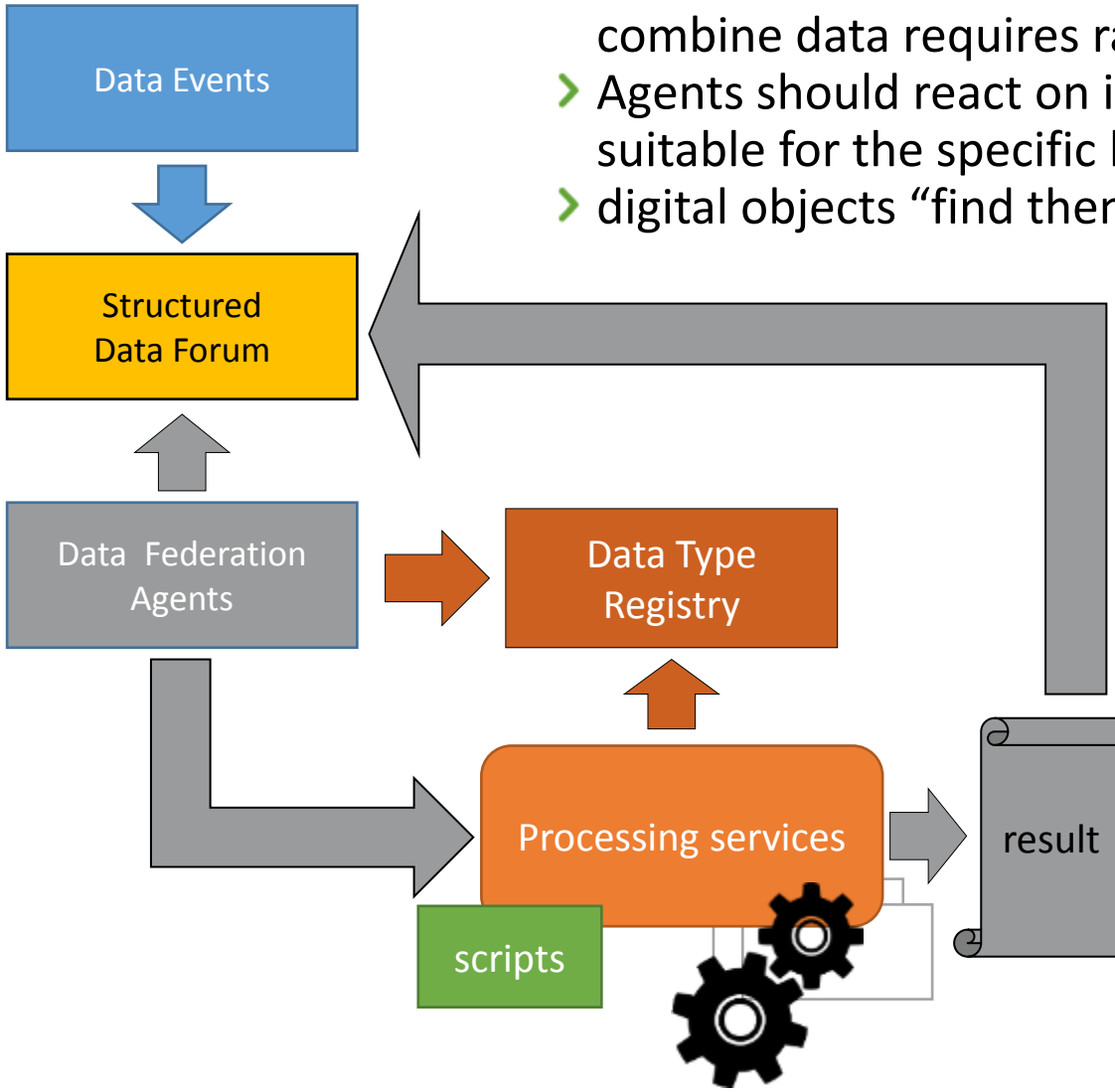


- flexible testbed to build and integrate methods based on DO concept
- come to a systematic and systemic solution for the data challenges
- be prepared for the phase of automatic processing of scientific data
- make it simple for researchers by reducing complexity and let them work in the world of DOs, i.e. working with metadata and PIDs
- integrate components specified by RDA, OAI, W3C etc.
- **Example1**: provide operators to work with DOs (move, delete, etc.)
- **Example2**: build workflows based on a types and proper provenance

Landscape of trustworthy repositories and registries based on interoperable components.

C2CAMP - Type-Triggered 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”

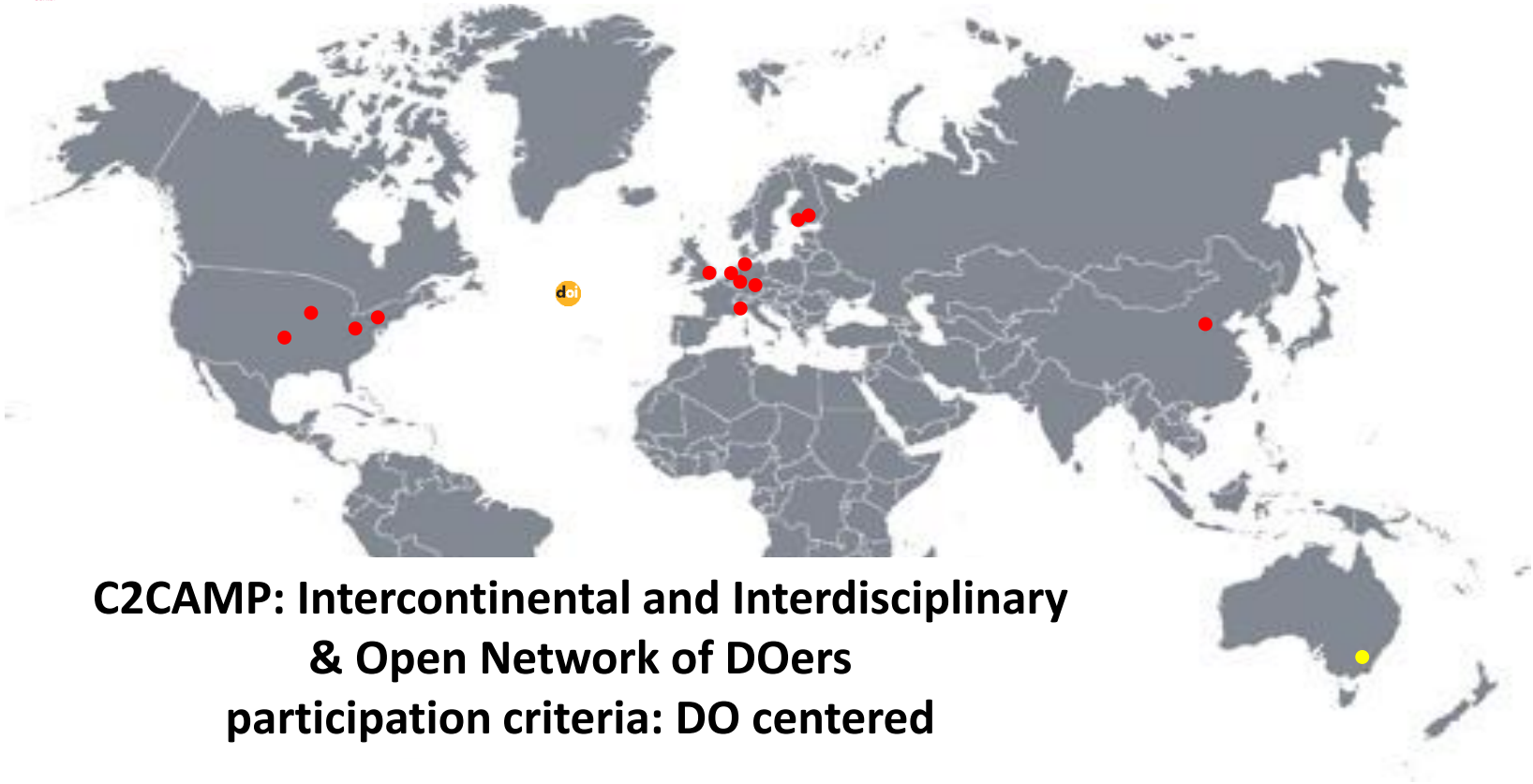


Basis are Digital Objects (Data, Software, Configurations, etc.) and Types

C2CAMP core partners

- some research infrastructures (ICOS, CLARIN, DISSCO, ENES)
- some typical data centers working closely with communities
- some HPC centers working also with communities
- some IT providers

- all grown from close RDA collaborations and joint priorities
- all agree that scientific cases are in the driving seat
- all have shown commitment and competence



C2CAMP: Intercontinental and Interdisciplinary & Open Network of DOers
participation criteria: DO centered

Join RDA and perhaps also C2CAMP (website coming: c2camp.org)

RDA Global

Email - enquiries@rd-alliance.org
Web - www.rd-alliance.org
Twitter - @resdatall
LinkedIn - www.linkedin.com/in/ResearchDataAlliance
Slideshare - http://www.slideshare.net/ResearchDataAlliance

RDA Europe

Email - info@europe.rd-alliance.org
Twitter - @RDA_Europe
Email – peter.wittenburg@mpcdf.mpg.de

Relevant Activities

- FAIR principles summarised broad discussions into a globally accepted common language and are an important step towards more convergence, but not a blueprint for infrastructure building
- US seems to be ready to fund a small seed project and work out a larger program along C2CAMP core ideas
- China is busy setting up a national PID infrstructures for everyone anticipating the coming challenges (IoT) and enabling systematic approaches
- RDA is a global platform that needs to be used more intensively to work on specifications of components and its characteristics
- GOFAIR is an emerging platform to bring together implementers and trainers, C2CAMP will act as Implementation Network in GOFAIR to synchronise with others
- EU activities
  - EOSC is a great attempt to synchronise minds in Europe, C2CAMP would add the synchronisation concept which is yet missing
  - PRACE/HPC/EDI lack a convincing DO management concept, Exascale without DO management would lack impact on data economy, FENIX as part of C2CAMP
  - FREYA addresses PID usage for schlolarly communication

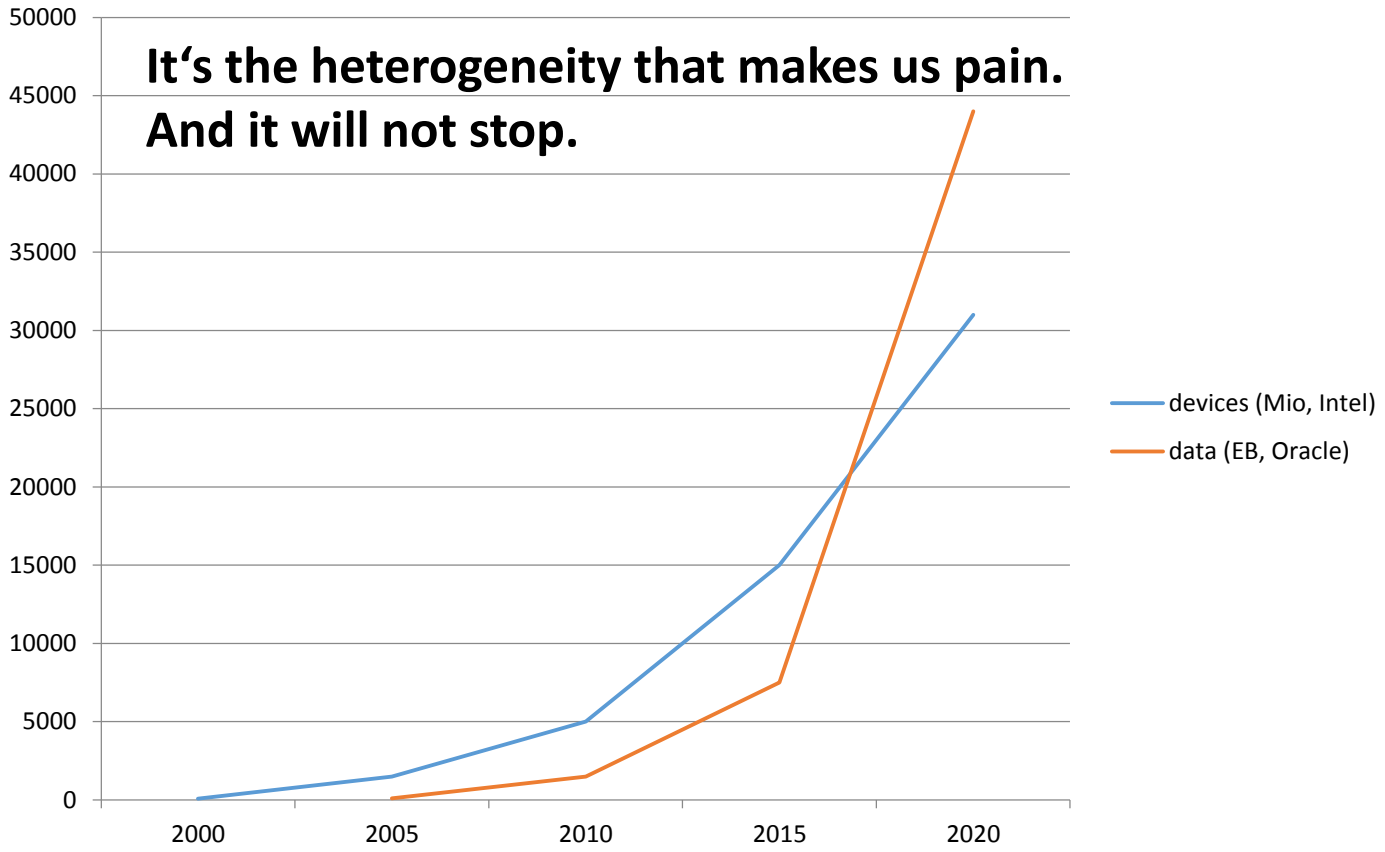
C2CAMP is FAIR compliant and adds an implementation concept missing in EOSC.

Coming to agreed principles

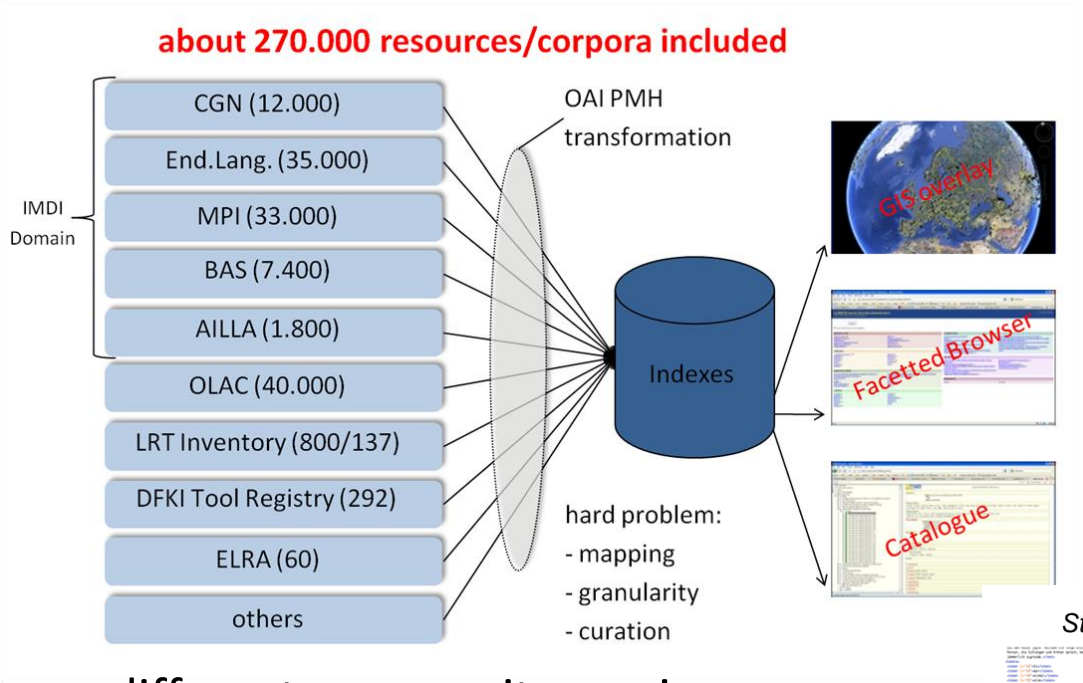
Pre-ICRI Meeting Copenhagen March 2012	G8 Data Group June 2013	Data Foundation & Terminology Sept 2013	FAIR Principles Summer July 2014
discovery	discovery	store data in	findable
access	access	trustworthy	accessible
interpretation	re-usable	repositories	interoperable
re-use	manageable	assign PIDs	re-usable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

increasing convergence & explicitness

Development of devices and data



CLARIN Research Infrastructure (2006-...)



- **bad state & org of data**
- **quality differences**
- **heterogeneity at all levels biggest problem**
- **semantic mapping**
- **metadata for workflows is special**
- **etc.**

- many different community services
  - component-based metadata setup
  - concept registry
  - Virtual Language Observatory
  - distributed Web-Workflow tool
  - etc.
- DSA/WDS compliant centres (repositories, etc.)
- fairly FAIR compliant setup