Digital Objects as anchor points in a complex data Domain

Peter Wittenburg

Max Planck Society, Max Planck Computing & Data Facility

RDA

Two Major Challenges

volumes
50 Mio smart devices producing continuous data streams

dynamics
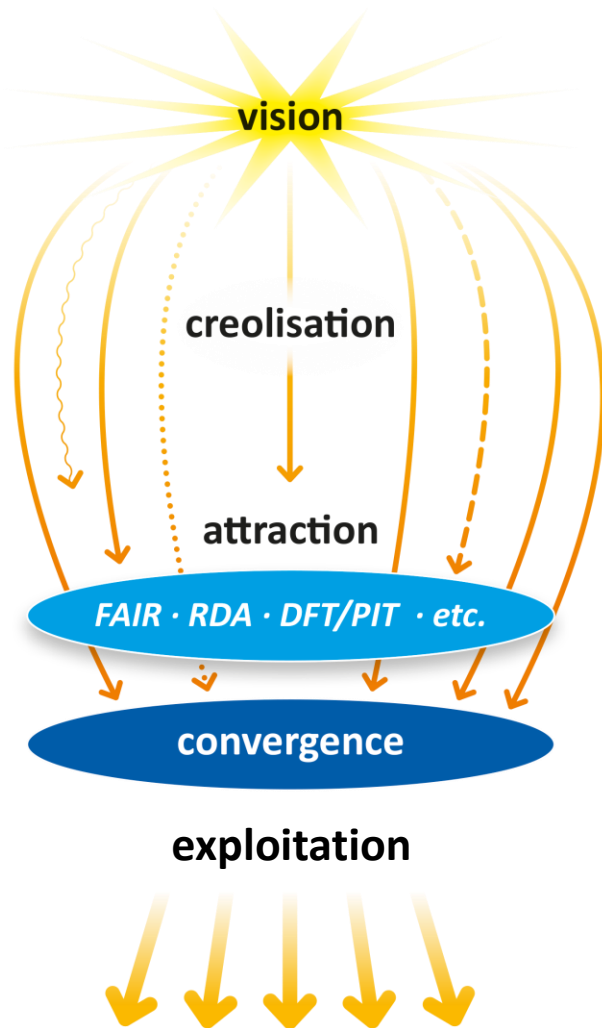enormous acceleration of dynamics and heterogeneity

Abdel Labbi (Dist. Eng. IBM) at RDA Plenary 2018:

Universe of data is changing faster than tools can be built.

are we lost in complexity?
how to organise data exchange (market)?

Creolisation in Data Domain

Creolisation
- enormous solutions space – 1000 flowers …
- thus huge fragmentation

Consequences
- **80% waste of time in data intensive projects**
- **60% data projects fail**
- **there is no open data market**
- **high costs and many are excluded**
- **low degree of automation**

Some Attraction now
- wide agreement on **identification**
- **FAIR** principles as a global and common language
- **Ref. Architectures**, **Frameworks/Platforms**
- grass-roots **RDA** to work out specifications of **components** (characteristics, interfaces)

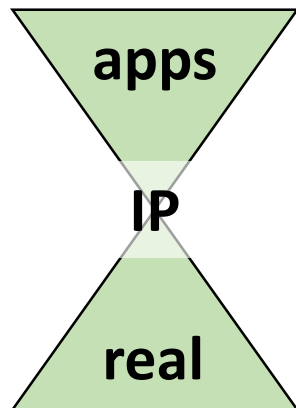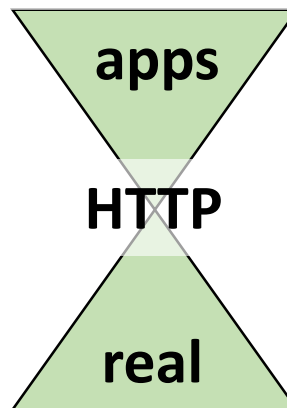## Convergence needed – but how?

Creolisation in early Internet
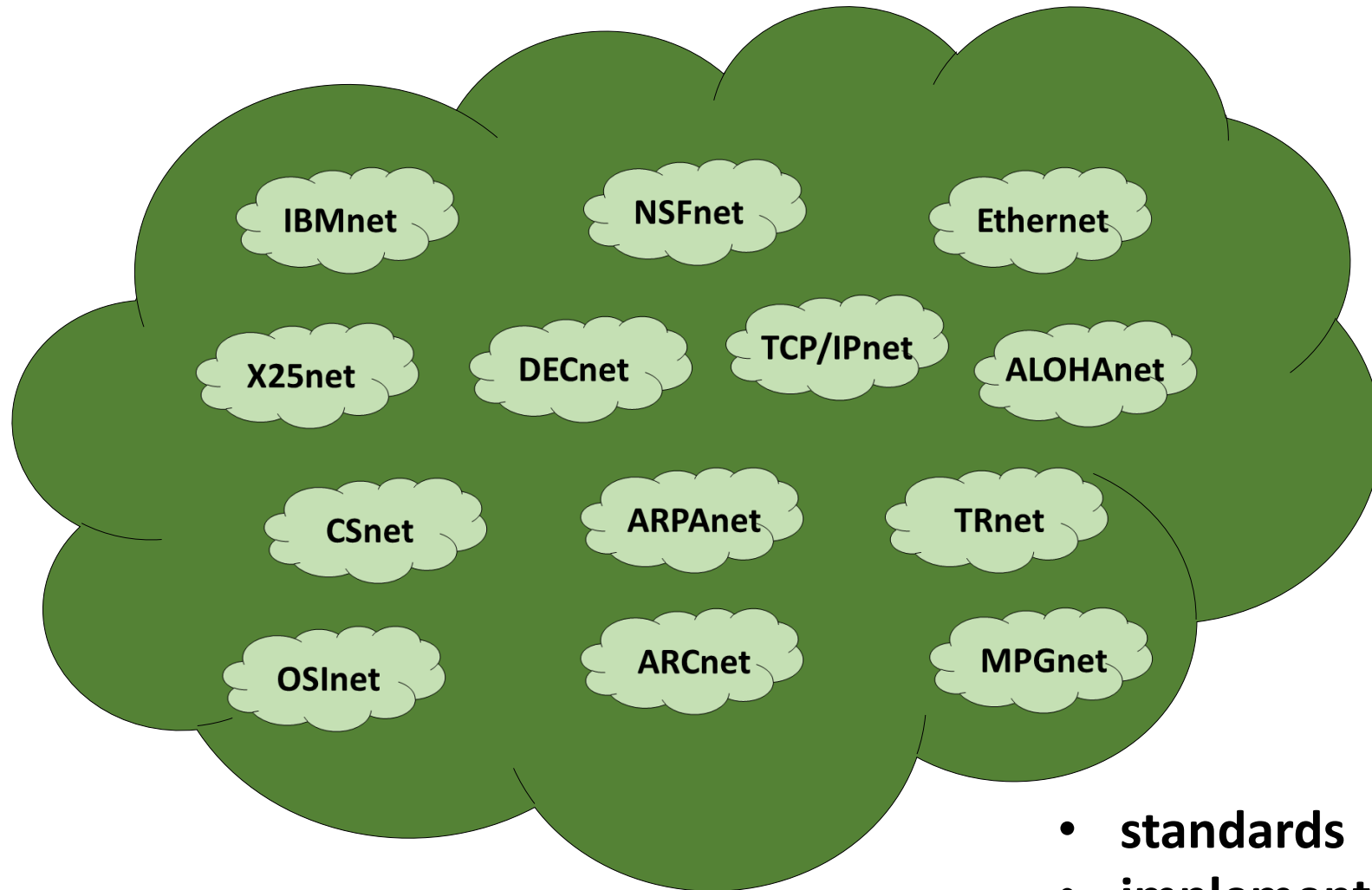
- **standards**
- **implementations**
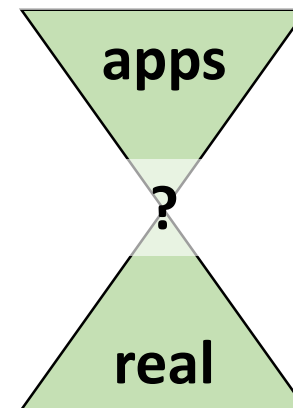
Well-known Hourglass Shapes

Computer Networking

apps

IP

real

World Wide Web

apps

HTTP

real

Data Domain

apps

?

real
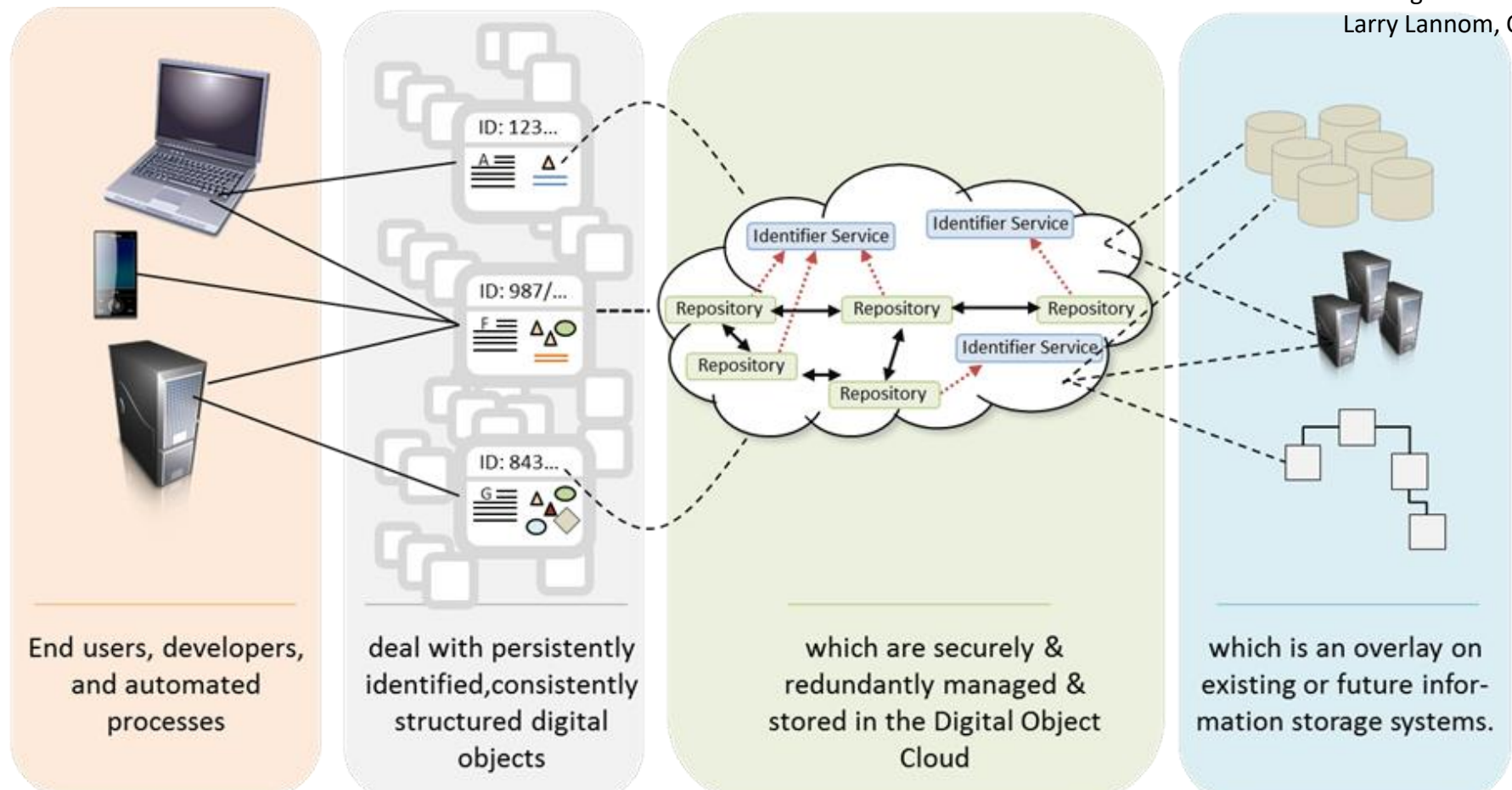
> defining a new stable layer to build on and save investments

> enabling a phase of exploitation based on reduced costs and global agreements

> of course covering only small small aspect of whole software stack but nevertheless huge impact

> some argue that WWW gives us all we need for data

Global Digital Object Infrastructure

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 infor- mation storage systems.

- user only deal with Digital Objects (Metadata and **PIDs**) - virtualisation
- components already being implementing, Clouds based on virtualisation
- working now on global testbed project (C2CAMP): DO operators, workflows, etc.
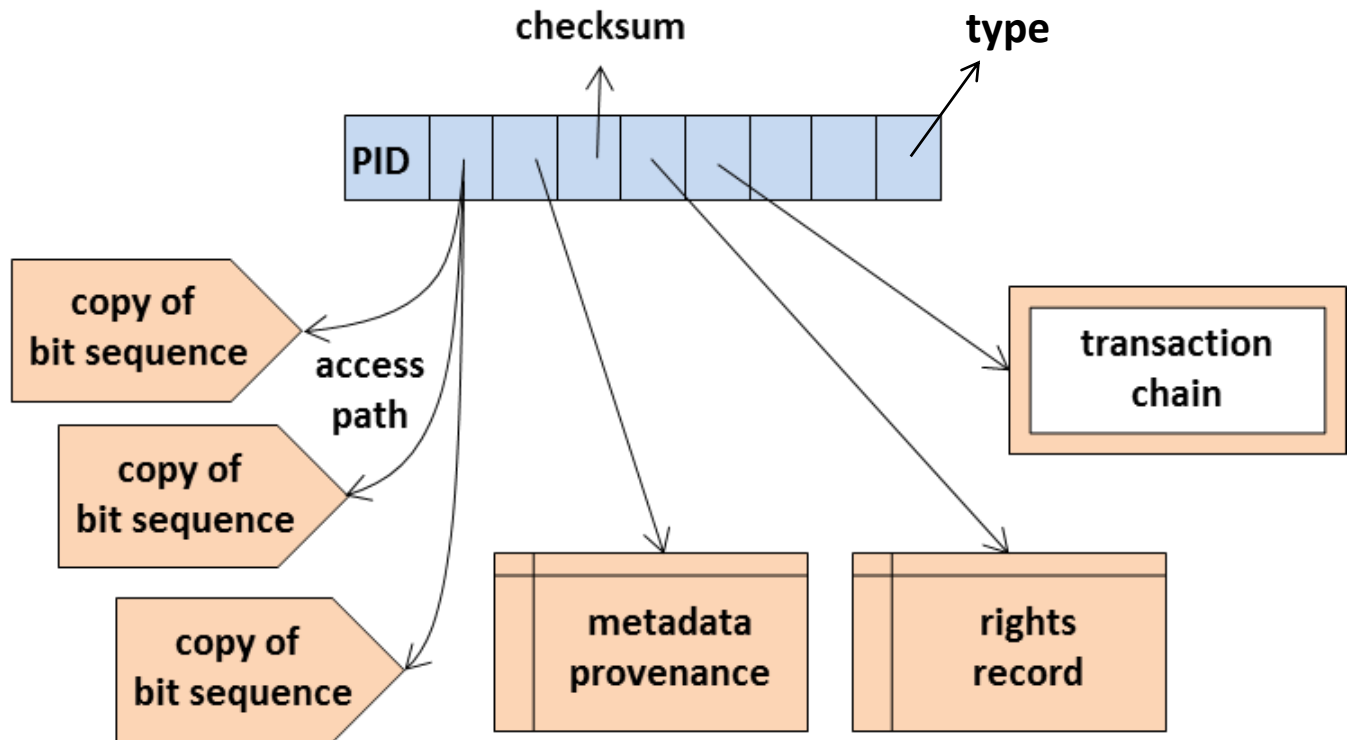
Core Data Model and PIDs are central
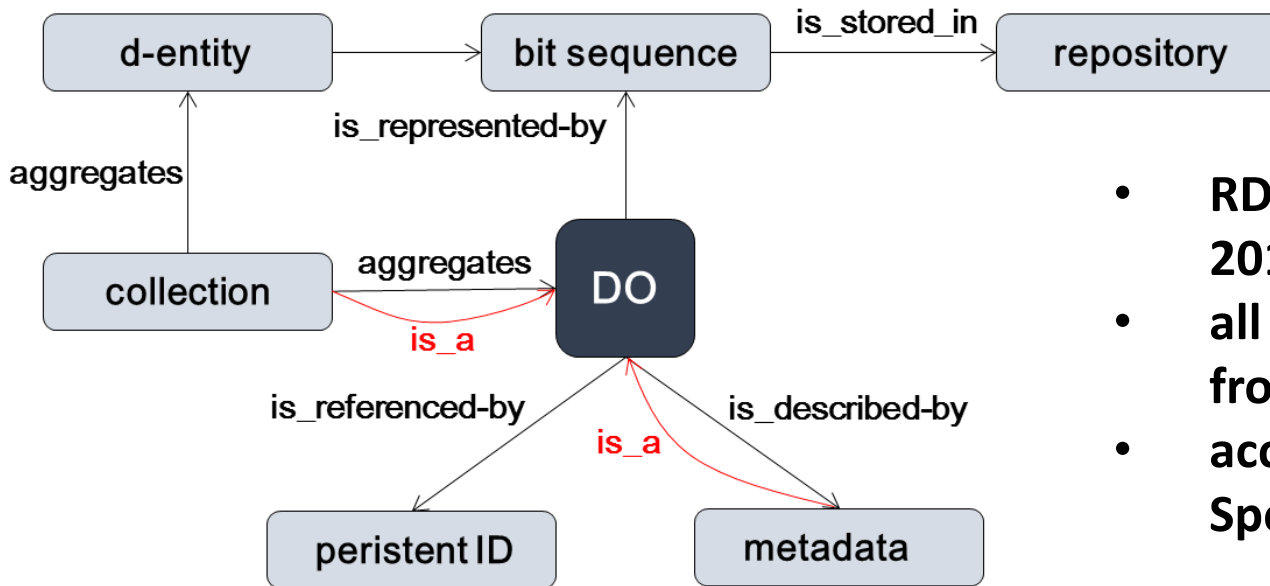
- **RDA standardisation since 2013 (DFT Core Model)**
- **all based on >20 use cases from various disciplines**
- **accepted now as EU ICT Specification**

- **Kahn: We need to know where we are talking about!**
- **getting dependent on functioning PID resolution system**
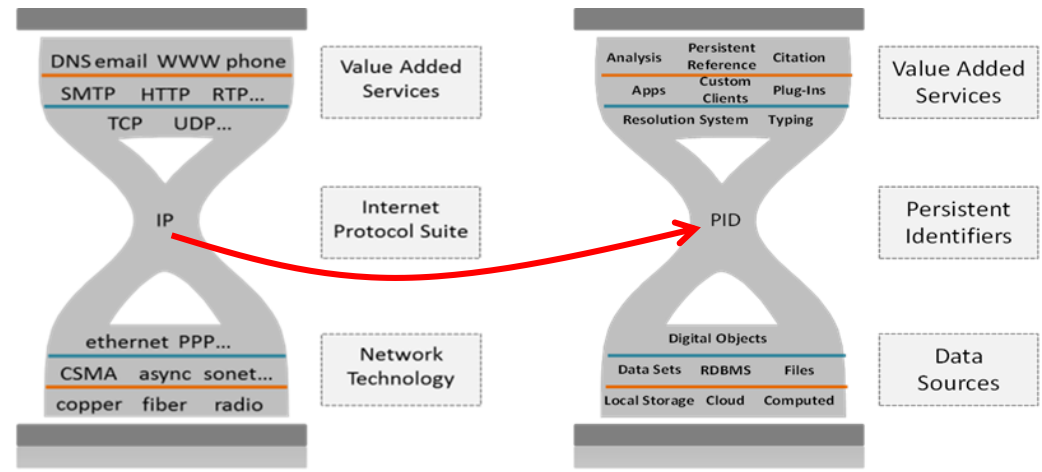
PID for binding (standardisation in RDA)

if we rely on PIDs we can do even more with them!

RDA PID Information Type WG & PID Kernel Information WG
- specify core attributes such as „checksum, type" to facilitate machine interpretation
- specify principles of interoperability

Worldwide Handle System

Connecting Repositories
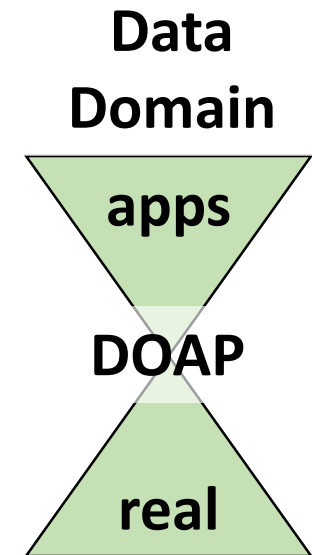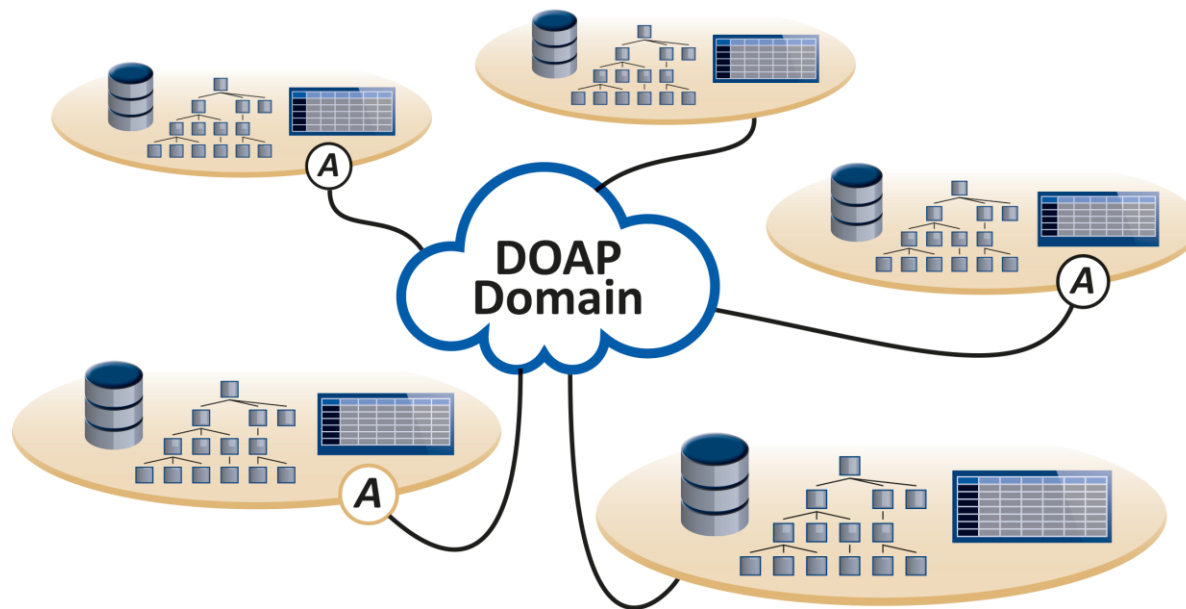
- at least interoperability between repositories whatever data model and data organisation they use
- some have compatible native organisation, others need to write adapters
- my experience: would help a lot to become more efficient, machines know to find all relevant information
- not addressing Semantic interoperability, but facilitating
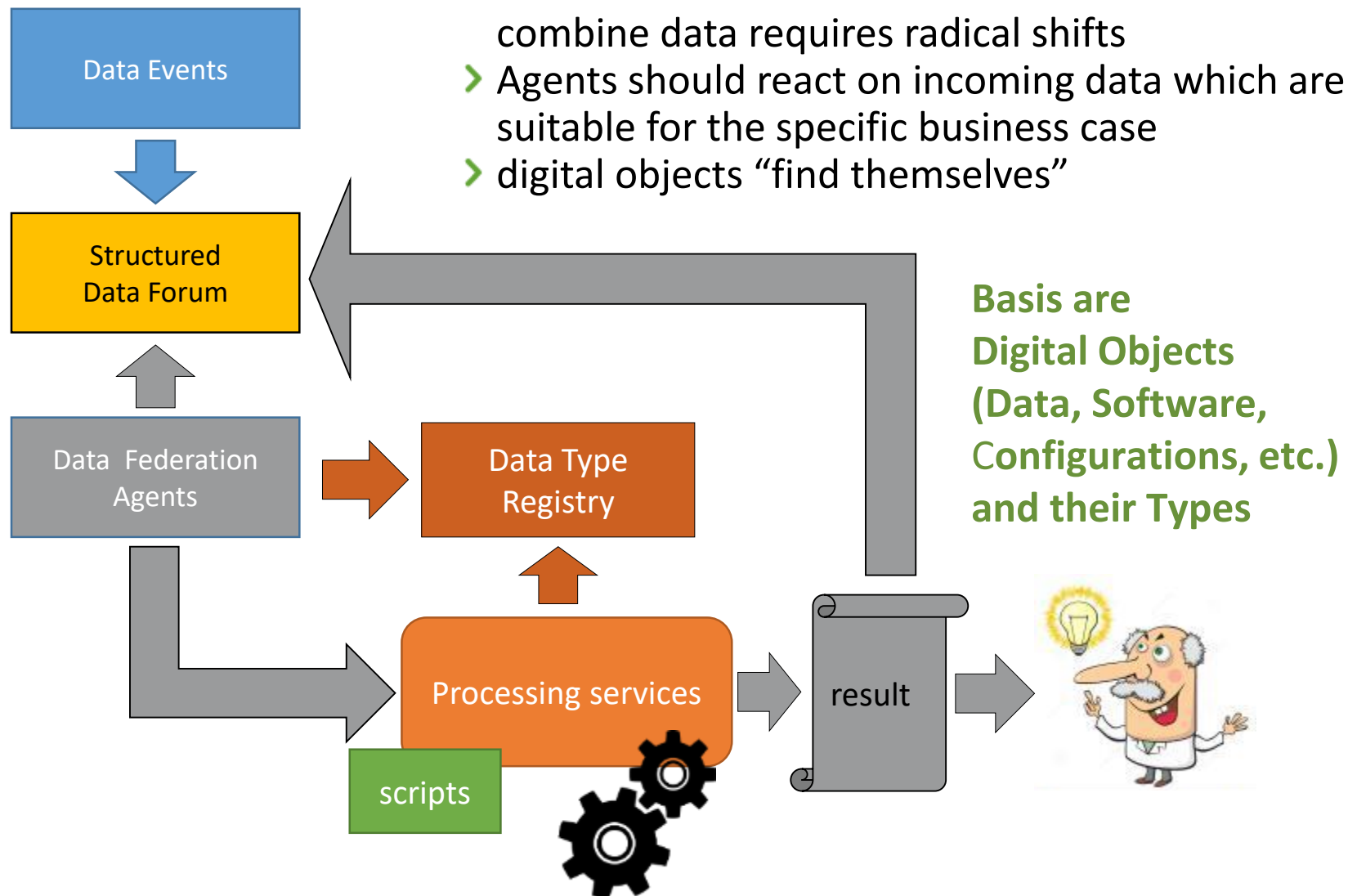- urgent now: specifying the DO Access/Interface Protocol

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

> digital objects "find themselves"

Data Events

Structured Data Forum

Data Federation Agents

Data Type Registry

Processing services

scripts

result

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

Market Types for data

|	Bilateral	Bazar
availability	very selective, silos	many are excluded (SMEs)
metadata	not needed	needed
costs, rights	special deal	explicit
PIDs	not needed	needed
trust	personal relation	broker, certification
roles	creators = dealers	creators =x= dealers
	current situation	to come

Do DOs help to increase trust?

> it gives each digital (data) 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 checked (smart contracts in blockchains -> see talk by Visa)

> is tracing possible?
> > well we could check fingerprint data when data becomes visible
> > can we prevent illegal copying and thus misuse?
> > > some people build closed platforms (Streamr, etc.)
> > > selective customers and not compliant to bazar
> > > we could seal all data – huge effort

Is that all to increase data sharing?

> **NOOOO**

> ownership is not clarified in many cases – different cultures
> GDPR is a big step ahead wrt to rights on personal data
> many fears of data misuse or lack of acknowledgements based on bad experiences
> big lack in data literacy hampering progress
> big lack in data professionals hampering progress
> separation of roles and certification needs to be solved
> licensing difficult (changing value of data, etc.)

MPCDF

Conclusions
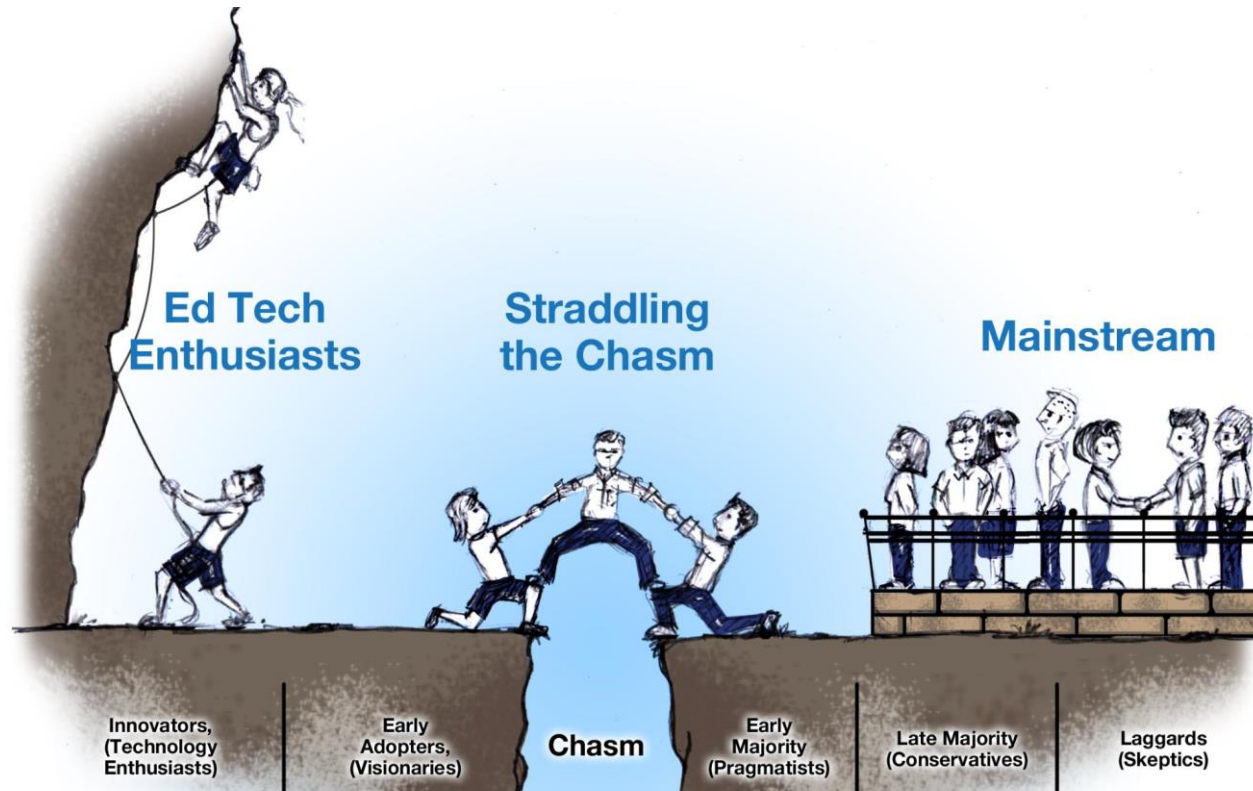
- let's now start building the DO based data domain
- it will solve basic interoperability issues and kick off redesign
- basic components are ready to go
- we need the entrepreneurs

DOs and PIDs are reality in some sectors

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

peter.wittenburg@mpcdf.mpg.de