Peter Wittenburg
RDA Europe Director
Max Planck Computing and Data Facility

Background

- 75% (RDA EU) - 80% (Michael Brodi MIT) of data scientist time is lost for typical data management & access (DMA) tasks - what a waste of time and money
- due to thousands of initiatives the solutions space is huge for all aspects of DMA
  - it creates big interoperability problems
  - it hampers investments, i.e. less innovation
  - no one can maintain solutions and pay for all the mappings
- it is time to change where possible
- W3C, IETF, RDA etc. create recommendations
- how can we create momentum to agree on basics?
  - need a discussion process involving the large RIs
  - need to have a comprehensive view on "bundles"
  - need to see where we can agree

GEDE Basis

- in Europe we have the ESFRI and other large initiatives
- sent 65 emails based on advice from EC
  - within few weeks 45 declared interest to participate and nominate delegates
  - 2 send emails questioning the need and the chances
  - some did not react, but we did not want to push and first consolidate
- GEDE exists since June 2016
- agreed on a Charter and Procedures
  - charter defines scope of GEDE activities
    *The aim of the **Group of European Data Experts in RDA (GEDE-RDA)** is to promote, foster and drive the discussions and consensus forming on creating guidelines, core components and concrete data fabric configuration building based on a bottom-up process.*
  - we should be careful and not blurr the mission
- participation is voluntarily and thus per definition weakly based
- do we need to include others as well??

GEDE Members

ACTRIS	ENVRI Plus	LIBER
AGRO	E-RIHS	LifeWatch
AnaEE	ESRF	METROFOOD
Asterics	ESS	MIRRI
BBMRI	ESS - Social	MYRRHA
CESSDA	EST	NIDI
CLARIN	EUCALL	PARTHENOS
DARIAH	EU-OPENSCREEN	RDA
EATRIS	EURO-ARGO	SCK•CEN
ECRIN	EURO-BIOIMAGING	SERISS
EISCAT	FAIR	SIOS
ELI	HBP	SKA
ELIXIR	IAGOS	VAMDCground
eLTER	ICOS	
EMBRC	INSTRUCT	
EMFL	KM3NeT	
EMPHASIS		
ENES		

GEDE Way

- had 2 virtual meetings - intention is about one VM every 3 months
- yesterday had the first f2f meeting

- main group
  - discuss bundles (compact thematic areas)
  - identify relevant initiatives that can contribute
  - discuss about subgroup results

- subgroups with experts to work on "bundles"
  - operate inclusive - aggregate all assertions from relevant initiatives
  - identify agreements beyond all terminology differences etc.
  - summarize agreements & disagreements

- the role of RDA people is the one of moderators and perhaps pace makers

Possible Bundles

- discovery/finding bundle
- repository bundle
- persistent identifier bundle
- data bundle
- data access bundle
- data processing bundle
- long term preservation bundle

- which ones to start with
- bundles and bundle definitions will change over time

Relationship with RDA

- RDA has a discussion process on recommendations as well
  - building charettes to find agreements
- RDA discussions are global

- GEDE discussion states are visible worldwide
- GEDE discussions are input for the RDA
- GEDE can take profit of RDA discussions

- will it all work? - don't know yet but what is the alternative

European Infrastructure for Persistent Identifier

Peter Wittenburg
RDA Europe Director
Max Planck Computing and Data Facility

Development in Science & Data Publishing

- 20 years of discussion about PID resolution systems
  - started with a paper about Digital Objects from B.Kahn and B. Wilensky in 1995 and building the Handle System!!!
  - is it necessary since we already have DNS, BarCodes, IPv6 etc?
    - common answers:
      - don't mix concepts (billions of devices – trillions of data objects)
      - resolution to meaningful state information is crucial (like passport)
      - securicy scheme is important
  - various solutions from which we learned a lot (AWKs, PURLs, LSID, Handles, URNs, etc.)
  - but also PID Zombies or almost Zombies (PURL, LSID, …)

- MPG decision in 2009
  - Handle Service for MPG scientists
  - but only if CNRI independence, security, reliable support
- and DOIs of course and layered services (DataCite, Crossref, etc.)

Development in Industry

- ITU X.1255 Standard (2013) - Framework for discovery of identity management information
- ITU Decision to support Digital Object Architecture (DOA)
- ITU Comparison Papers: Handles, BAR Codes, etc. (BAR codes as Handle suffixes, etc.)
- Invention of the Clouds as „Object Stores" with Internal IDs http://s3.amazonaws.com/bucketname/**objectid**
- Discussion in IoT domain about identification methods and suitability of DOA
- China: set up of a Child Food Supply Chain Control System based on Handles
- China: coming decision about a decision about an identification scheme for the country (science, industry, governement, etc.)

Recent PID Workshop

- experts from many relevant stakeholders (DOI, Handle, DataCite, CrossRef, ITU, etc.)

- Proper PID Usage and support will become key for competitiveness ...
- PIDs need to be used by all parties dealing with data professionally ...
- **International and national steps to be taken urgently to offer a sustainable, structured and mature PID service landscape ...**
- Efforts to be taken to offer services across sectors and communities
- Urgently need to come to a structured and integrated domain of Handle Service providers
- Service providers need to ensure that these two interoperable domains are part of one integrated landscape of rich services.

https://www.rd-alliance.org/sites/default/files/attachment/20160901_RDA_PID_event_Garching_report_final.pdf

PID Bundle

PID1. RDA DFT1.1: A digital object is ... referenced and identified by a <u>persistent identifier</u> ...

PID3. RDA DFT1.3: A <u>PID record</u> contains a set of attributes stored with a PID describing DO properties.

PID5. FAIR-F1: (meta)data are assigned a globally unique and eternally <u>persistent identifier</u>

PID6. FAIR-A1: (meta)data are retrievable by their <u>identifier</u> ...

PID7. RDA-PIT1: <u>PID systems</u> should support the generic PIT API where Information Types (*properties of DOs)* are openly registered and defined.

PID10. DOI: For electronic documents and published digital objects register a <u>digital object identifier</u> (*DOI, which is a Handle with prefix 10*) and associate suitable information with it *(such as citation metadata).*
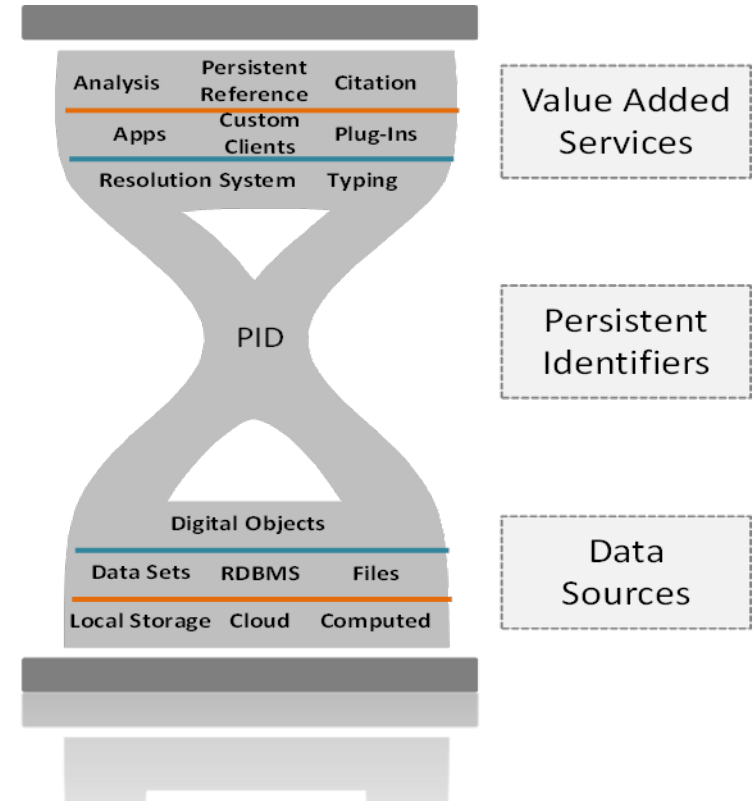
Global Digital Object Cloud
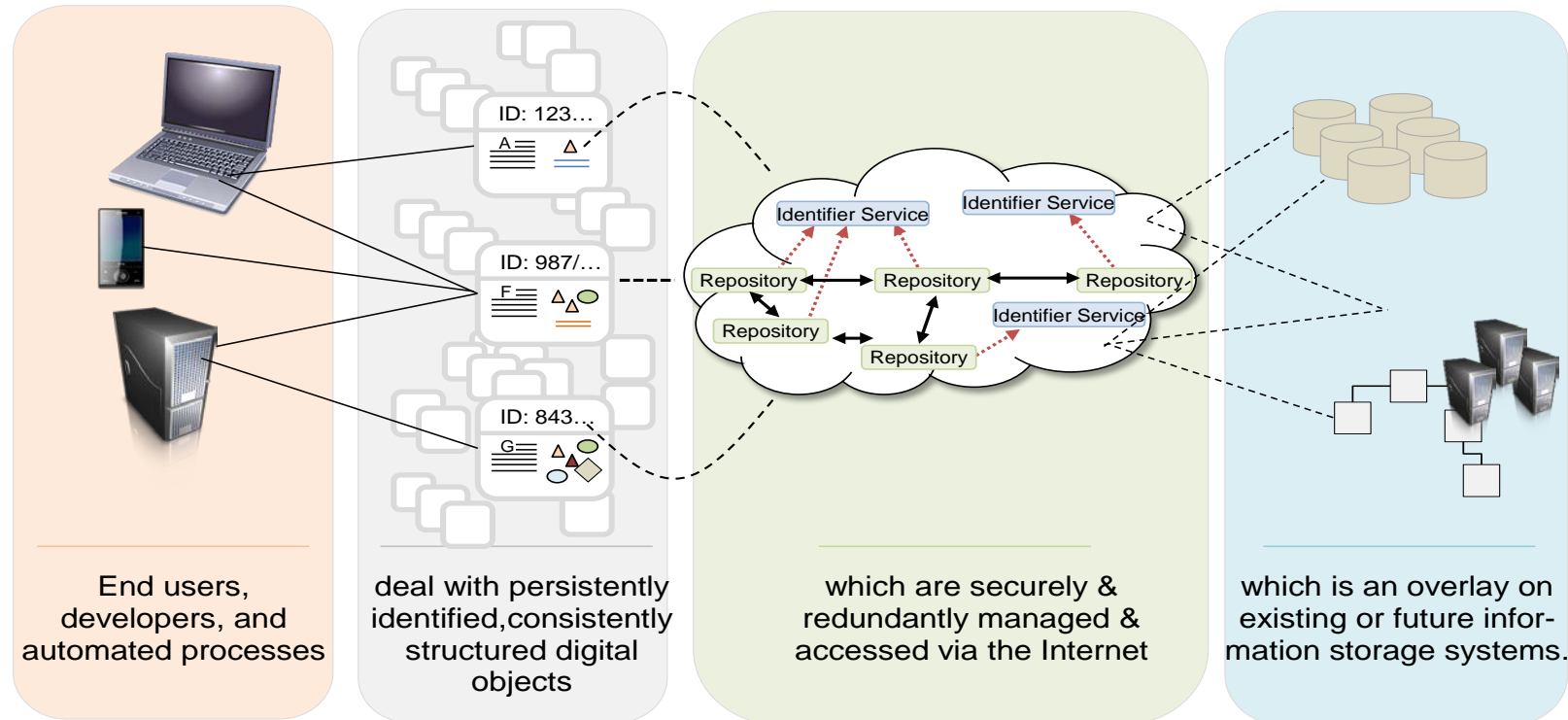
- if we can rely on PIDs we can build a virtualisation layer where users simply rely on representations of DOs (PID and associated infos)
- PIDs can be used to „bind" all relevant information
- is it just dreaming – no some groups are working on it

PIDs as Anchors
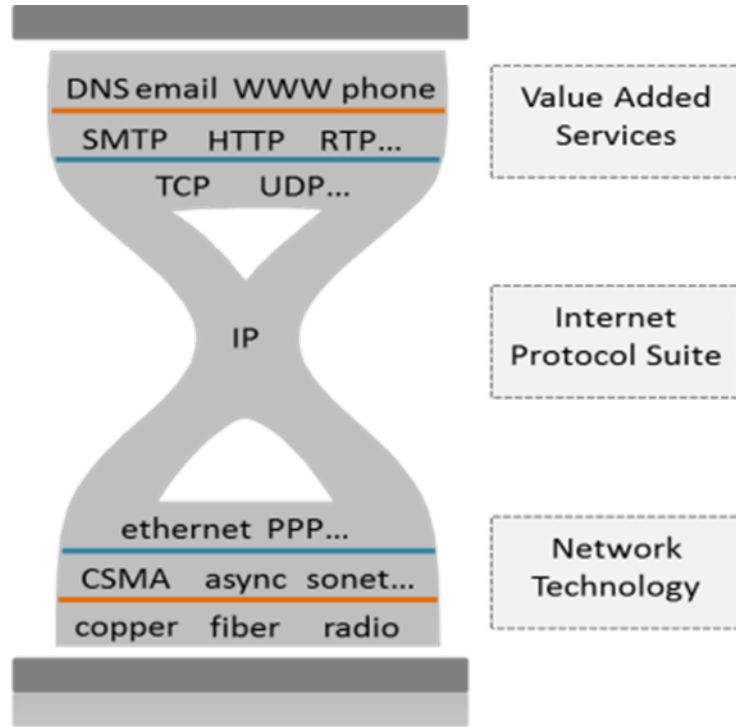
Agreement on IPs and
a few protocols around IPs
were crucial for momentum

Agreement on PIDs and
a few protocols around PIDs
are crucial to achieve
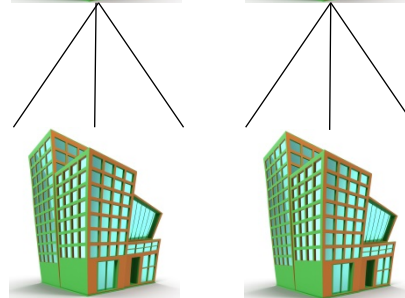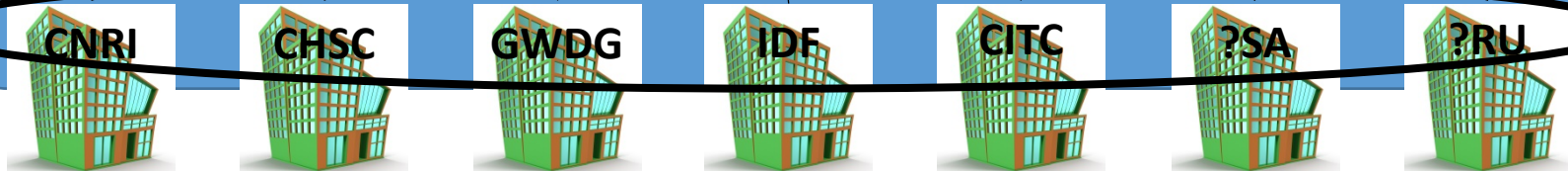momentum

Handle System is Ready

DONA

Independent Swiss Foundation

DONA Board of International Experts

Redundant network of root nodes

Contracts

CNRI CHSC GWDG IDF CITC ?SA ?RU

EPIC DataCite

Thanks for the attention.