Data Fabric and GEDE push towards broad agreements on PID Systems

Peter Wittenburg

Max Planck Computing and Data Facility

Motivation for activities

- the solution space in the data domain is huge
  - all infrastructure projects have creative people and „know better“
  - let‘s not forget industry having the same problems (-> Alex)
    - IoT with its billions of cyberphysical systems each creating huge amount of data objects changes the game
    - in China massive usage of Handles in food chain supply control
  - hampers interoperability
  - hampers effective software development and thus tool support
  - the costs are too high and no one can maintain all solutions

- senior researchers and funders ask for change

- how to do it? how to get tool builders on board?
- FORCE11, DFIG, GEDE, etc. are attempts to give impulses
  - need to learn to be inclusive and not special!
  - nevertheless rough consensus

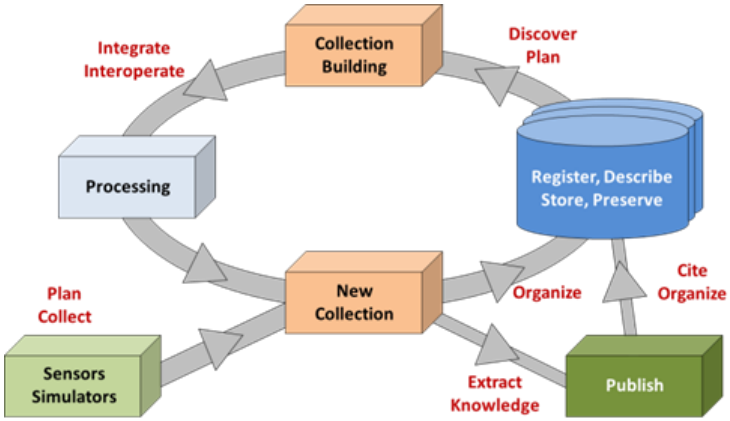
20 Years of PID Discussion – is it enough?

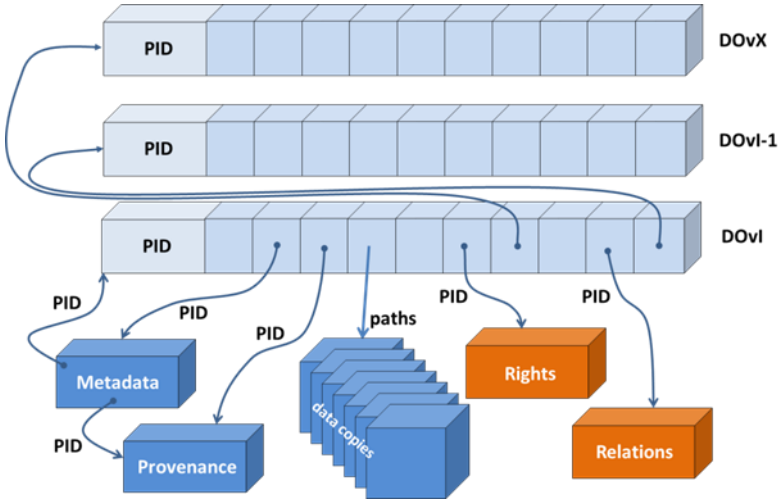
- 1995: Kahn & Wilensky

  A Framework for Distributed Digital Object Services

- recent Scientist Workshop in RDA
  - unclear what PID state is and what possibilitis are
- over the 20 years many suggestions, implementations & tests from different points of views (publications, labs, sem-web)
- effectively reduction of solution space towards DOIs and Handles
  - learned a lot from other approaches
  - semantic web still on different track
  - but there are other „big“ ID systems (bar codes, etc.)
  - China is going to set up a national PID system (kind of baseline) (huge performance/robustness/security issues to be solved)
  - it will have an effect on academic domain

Data Fabric IG: Guidelines & Testing



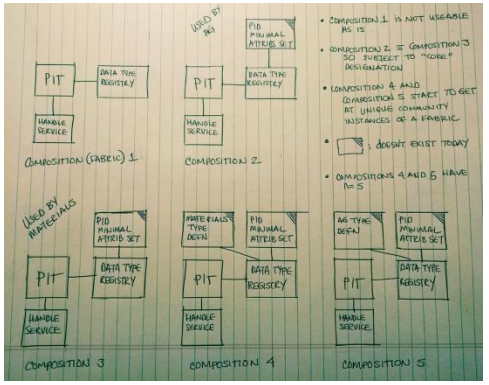


- identify essential components to make lab work more efficient & specify them
- stimulate testing
- stimulate configuration building by bringing components together
- with respect to PID usage different approaches in DFIG
  - few groups who just want an identifier
  - most groups who want to add useful information to Digital Objects
  - proof identity
  - check location
  - etc.
  - it‘s all for machine use

Data Fabric IG: configuation building

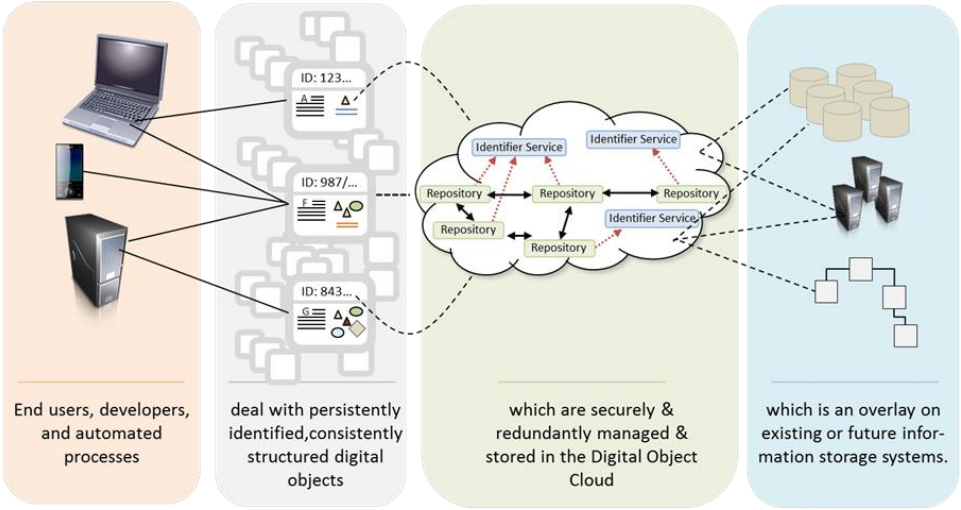
- PID Centric Data Management and Access approach – 2 tracks

1. bring current results together, identify gaps and build
   -> Beth



2. Global Data Object Cloud
   - identify urgently needed components
   - use what is there
   - and build

   -> Tobias



all documented here:

https://rd-alliance.org/group/data-fabric-ig/wiki/df-configuration-pid-centric-data-management-and-access.html

Is that sufficient?

- **NO** – we also need an agreement finding process incl. guidelines at the end
- started this in DFIG with restricted scope
  - Identify a number of Bundles https://rd-alliance.org/group/data-fabric-ig/wiki/recommendations.html):
    - structure issues
    - discovery/finding bundle
    - repository bundle
    - persistent identifier bundle
    - data bundle
    - data access bundle
    - data processing bundle
- Define characteristics and requirements for each bundle by building Charrettes that will take responsibility
- first: collect all relevant assertions about a bundle

Is that sufficient?

- **NO** – we cannot just discuss this in the RDA realm
- need to include the practitioners building infrastructures!!!
  - therefore in Europe we started the GEDE group
- need to include industry
  - therefore we started to make links with IoT, ITU, etc.
- need a structured approach with some commitment
  - **R**esponsible
  - **A**ccountable
  - **C**onsulting
  - **I**nformed

RACI Model

GEDE RDA
RESEARCH DATA ALLIANCE
Group of European Data Experts in RDA

MPCDF

CHARTER: The aim of the **Group of European Data Experts in RDA (GEDERDA)** 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.

- platform to engage European data professionals in a cross-disciplinary exchange of principles, recommendations and standards in the data domain;
- discuss messages coming from RDA Working & Interest Groups and other initiatives working on data issues;
- discuss core components and concrete infrastructure configuration building and testbeds:
- discuss and develop aggregated guidelines/recommendations and contribute to the RDA Atlas of Knowledge.

Finally: work on reducing the huge solution space

The GEDE Members

MPCDF

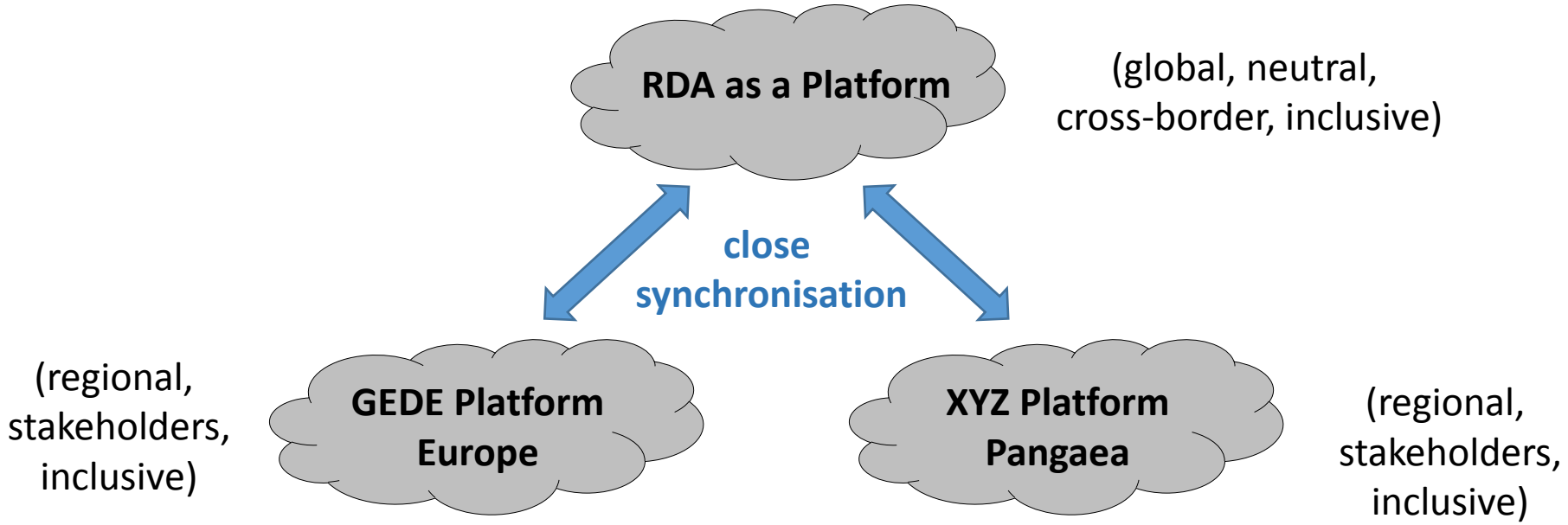
ACTRIS
ACTRIS
AGRO
AnaEE
Asterics
BBMRI
CESSDA
CLARIN
DARIAH
EATRIS
ECRIN
EISCAT
ELI
ELIXIR
eLTER
EMBRC
EMFL
EMPHASIS
ENES
ENVRI Plus
E-RIHS
ESRF
ESS
ESS – Social
EST
EUCALL
EU-OPENSCREEN
EURO-ARGO
EURO-BIOIMAGING
FAIR
HBP
IAGOS
ICOS
INSTRUCT
KM3NeT
LIBER
LifeWatch
METROFOOD
MIRRI
MYRRHA
PARTHENOS
RDA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

GEDE Procedure

- reduce the complex data landscape into coherent chunks
  - let‘s call them bundles
- define bundles
- define responsible chairs (not so easy)
  - RDA must continue to help as moderators, pace makers
- build groups of engaged experts
- aggregate assertions of relevance from different initiatives
  - identify common sense
  - extract agreed messages
  - identify disagreements
  - check overall coherence
  - is it tested??
- submit regular reports to GEDE and RDA
  - obviously some iterations

started with PID Bundle

Discussion Process complicated



- we need to speed up to be prepared for the challenges in N years
- how to include industry in this – as in internet times some just want to impress their standards
- is it all feasable?
  - don‘t know
  - so let‘s try or does someone know better?
- it remains an open interaction based on rough consensus and testing

Possible Directions

- have connected national „systems“ available for everyone will be a good baseline approach
  - there will be specialised services
- need to overcome the Handle/DOI devide
  - give clear messages about the publication of collections
  - but there is the huge domain of DOs associated with Handles
- how to do it – very simple or?
  - establish criteria for accepted service providers and give a seal
  - include the HQ service providers in metrics
- will come back on this at the end

Thanks for your attention.