RESEARCH DATA ALLIANCE

Data Foundation & Terminology WG
Data Fabric IG

Peter Wittenburg
Co-Chair

research data sharing without barriers
rd-alliance.org

Content

1. Principles & Trends
2. Data Practices
3. Data Foundation & Terminology WG
4. Data Fabric IG

Research is changing

- nr. of researchers increases enormously
- there is a pressure in the direction of Grand Challenges and those topics relevant for societies
- research is increasingly often data intensive
- border-crossing research is a fact (countries, disciplines)

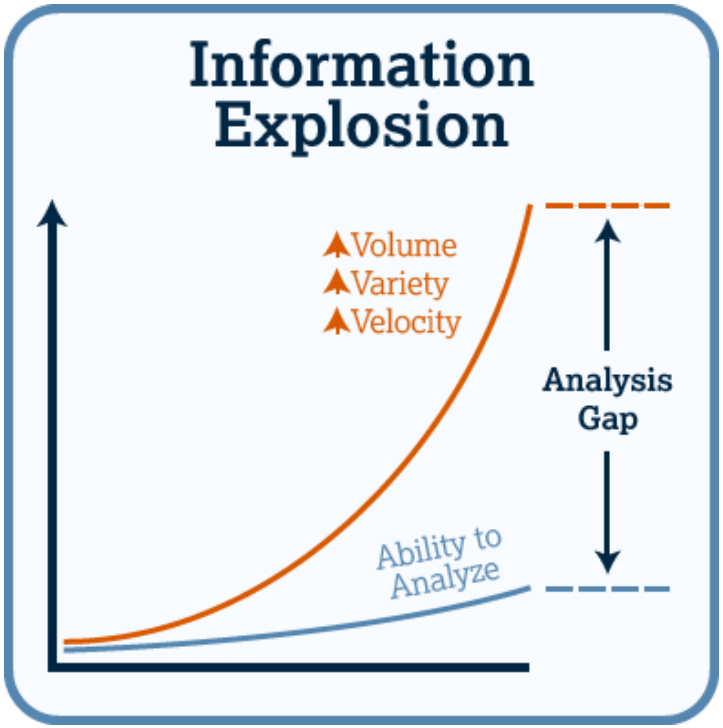
Data are in Focus
Research & Data are global

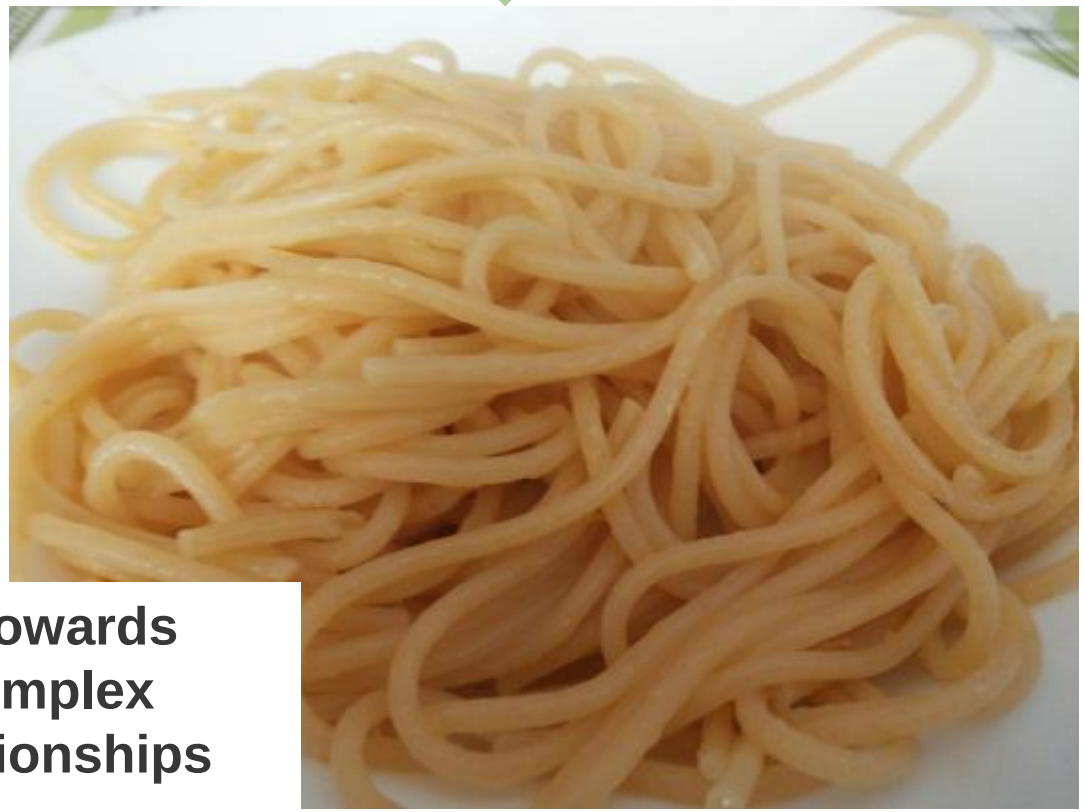
Requirements for Data Science

- let's use the G8 formulations – data should be
  - searchable -> create useful metadata
  - accessible -> deposit in trusted repository and use PIDs
  - interpretable -> create metadata, register schema and semantics
  - re-usable -> provide contextual metadata
  - persistent -> provide persistent repositories
- Funders request Data Management Plans?
- What are the consequences of these principles?
- How to design the necessary infrastructure?

Trends I – Volume, Complexity

from simple structures ...

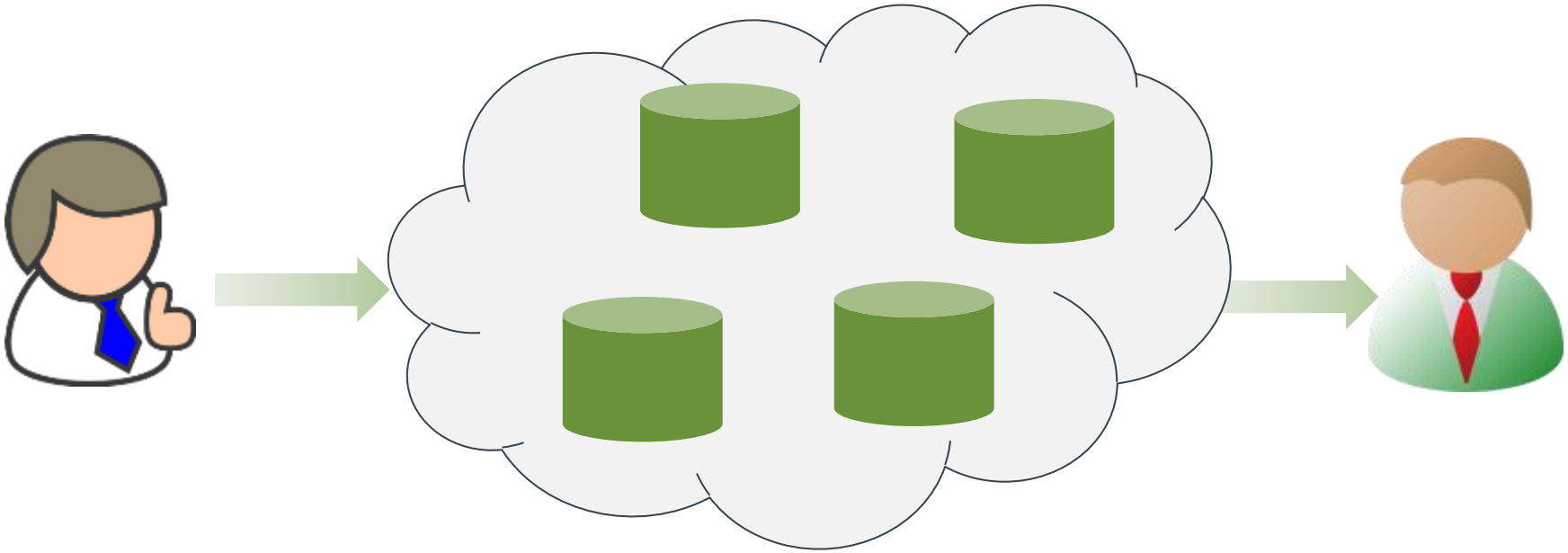




... towards complex relationships

Trends II - Anonymity

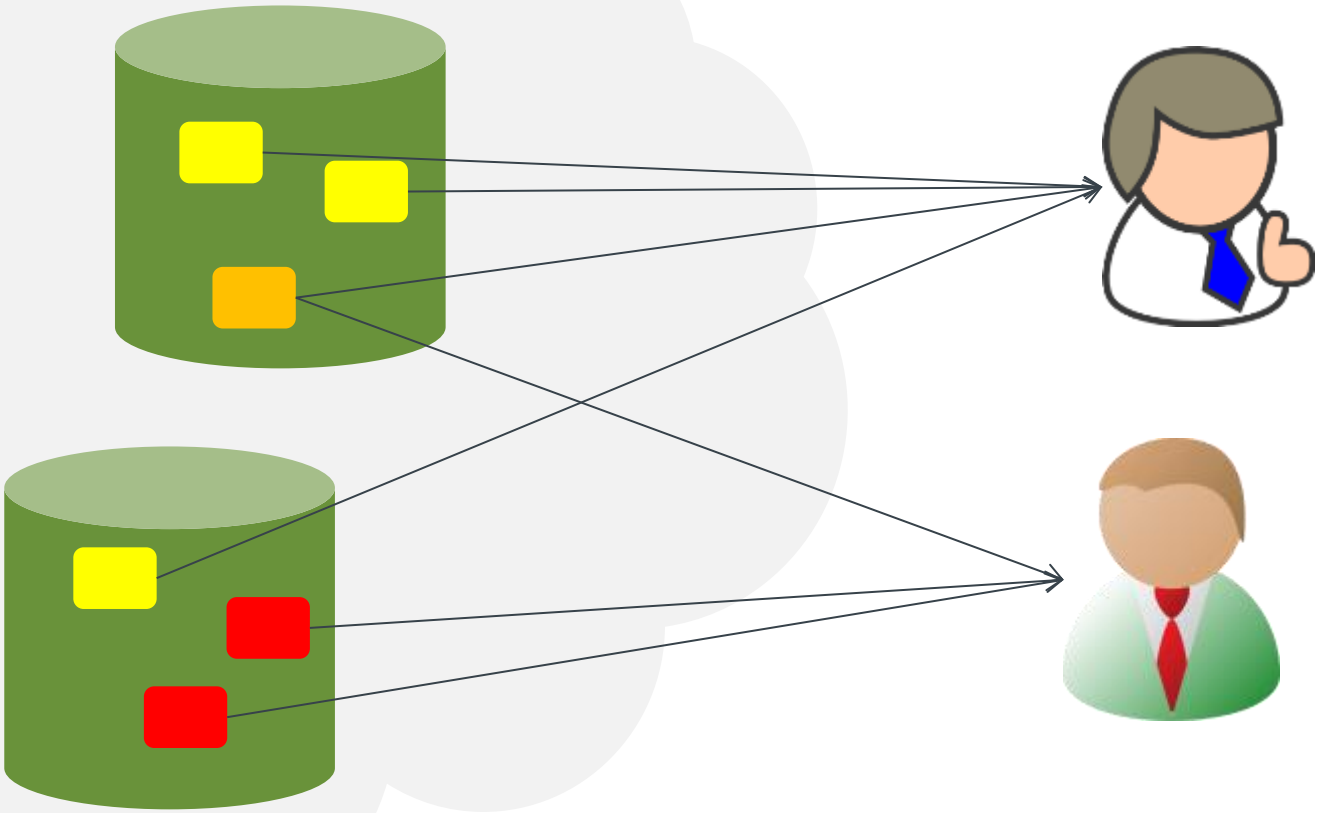
direct exchange between known colleagues



Domain of Repositories

new mechanisms of building trust needed

Trends III – Re-Usage



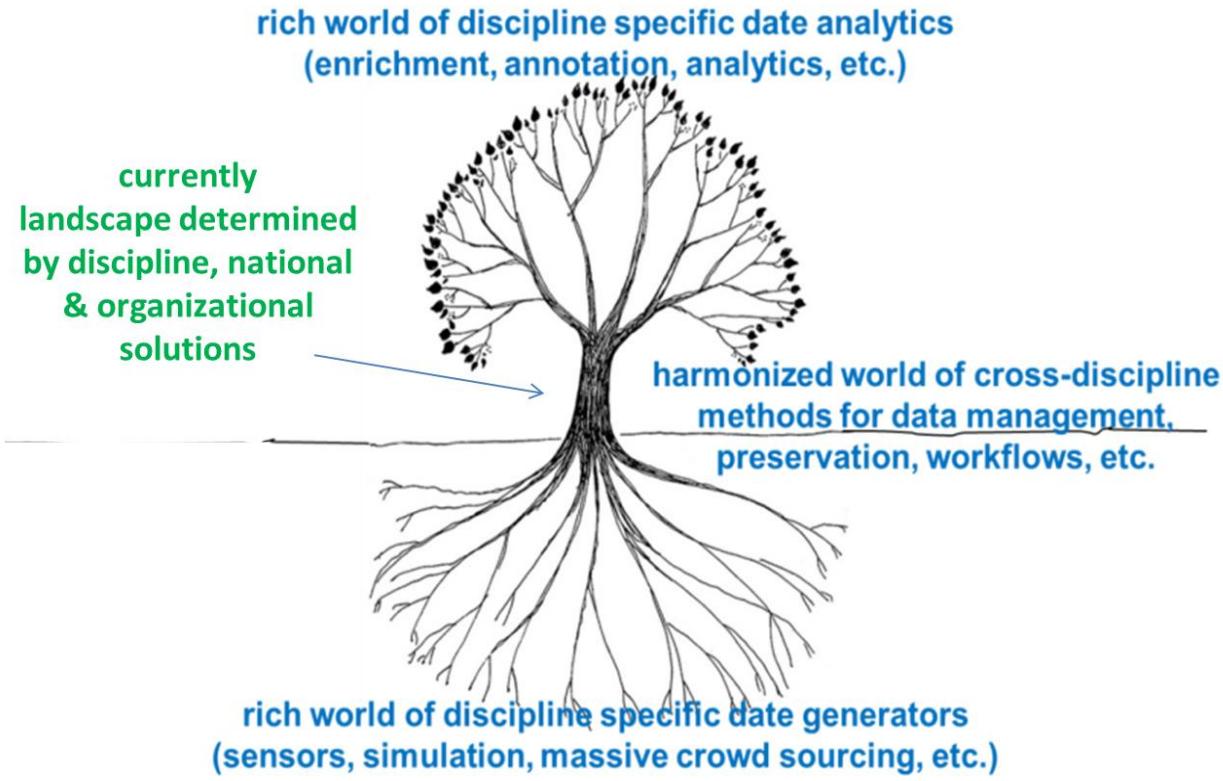
Data will be re-used in different contexts

Trends VI – large federations

taken from EUDAT

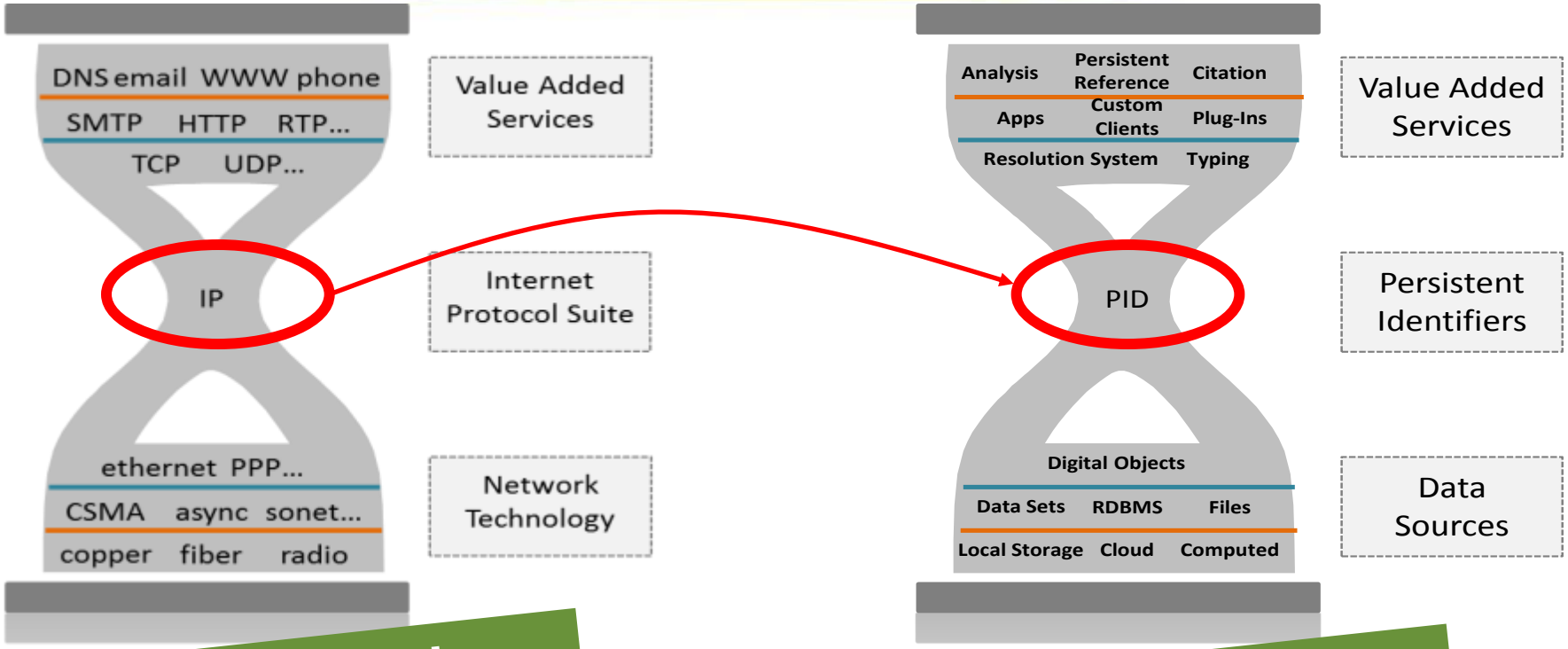
- domain of registered data
- various common data services (across countries & disciplines)

Trends V – unified Data Management



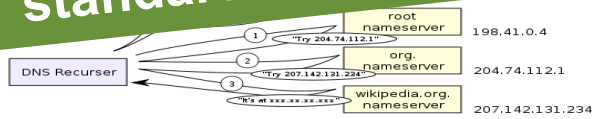
management of data objects is widely type and discipline independent

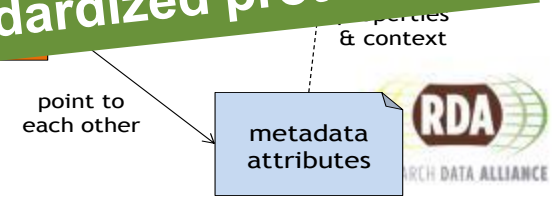
Trends VI – world-wide PID system



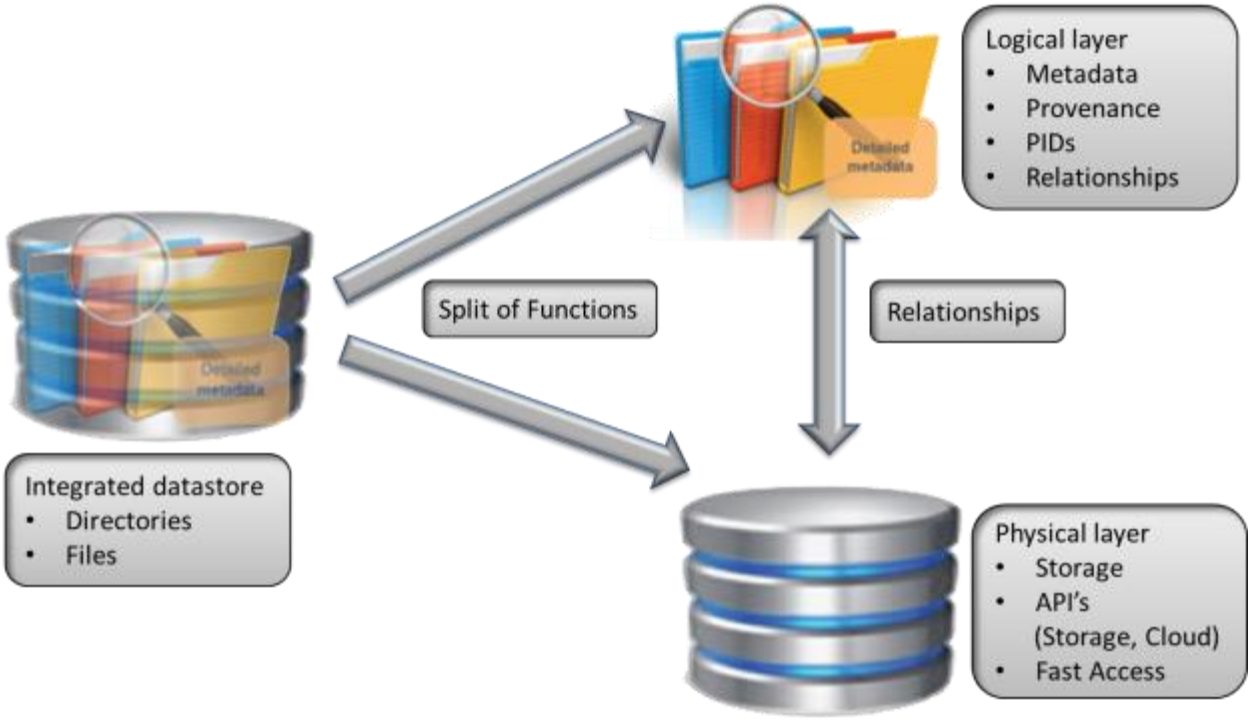
Internet Domain
nodes with IP numbers
packages being exchanged
standardized protocols

Data Domain
objects with PID numbers
objects being exchanged
standardized protocols





Trends VII – split of functions



"logical layer" operations are complex due to relations, etc.

Content

1. Principles & Trends
2. Data Practices
3. Data Foundation & Terminology WG
4. Data Fabric IG

Data Practices I – Survey

- ~120 Interviews/Interactions
- 2 Workshops with Leading Scientists (EU, US)

- too much manual or via ad hoc scripts
- too much in Legacy formats (no PID & MD)
- there are lighthouse projects etc. but ...
- DM and DP not efficient and too expensive
  (Biologist for 75% of his time data manager)
- federating data incl. logical information much too expensive
- hardly usage of automated workflows and lack of reproducibility

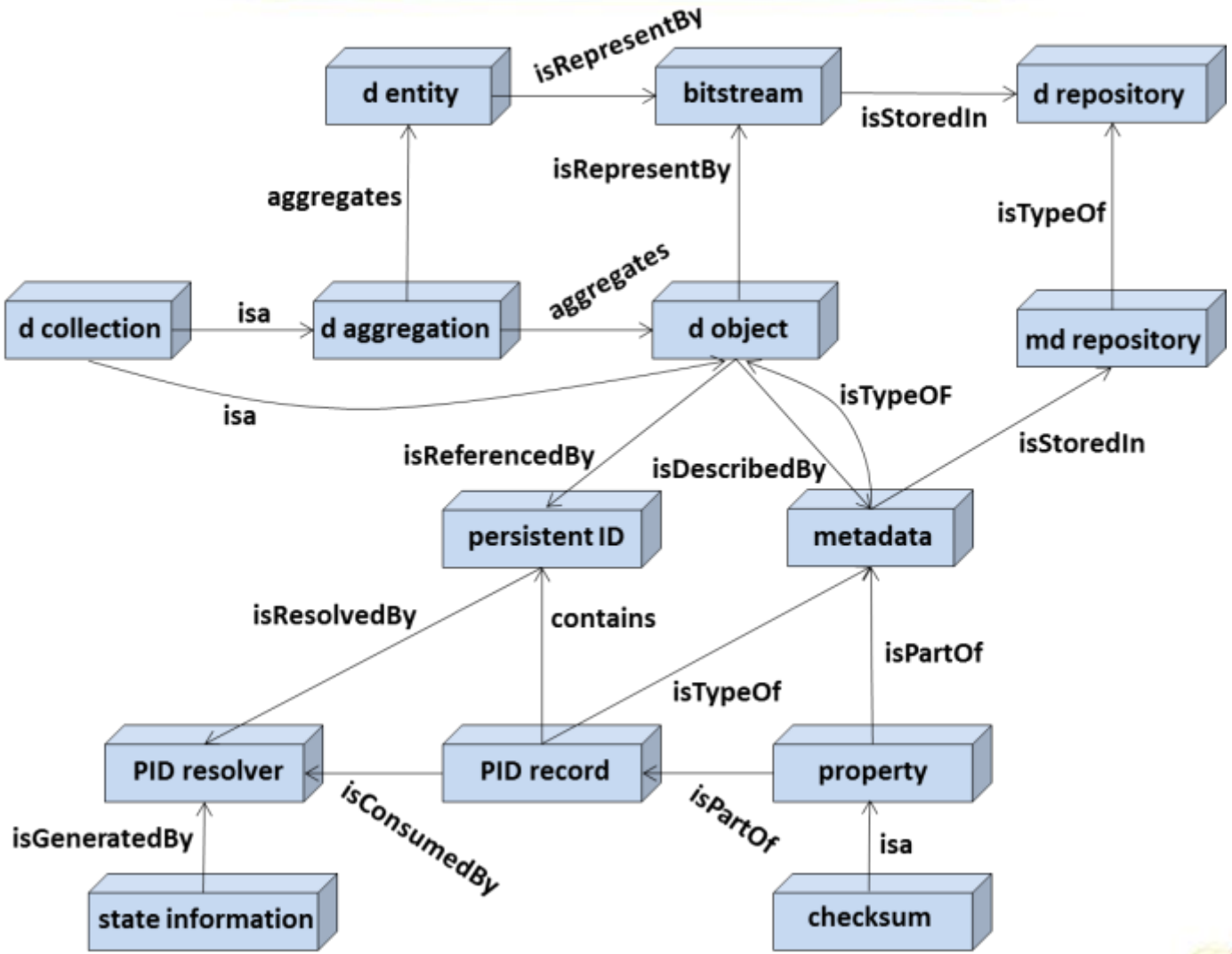
Data Practices I – Survey

- ~120 Interviews/Interactions
- 2 Workshops with Leadi
- DI research only available for Power-Institutes
- pressure towards DI research is high, but only some departments are fit for the challenges
- **Senior Researchers: can't continue like this!**
- need to move towards proper data organization and automated workflows is evident
- but changes now are risky: lack of trained experts, guidelines and support
- [illegible] much too expensive
- Hardly usage of automated workflows and lack of reproducibility

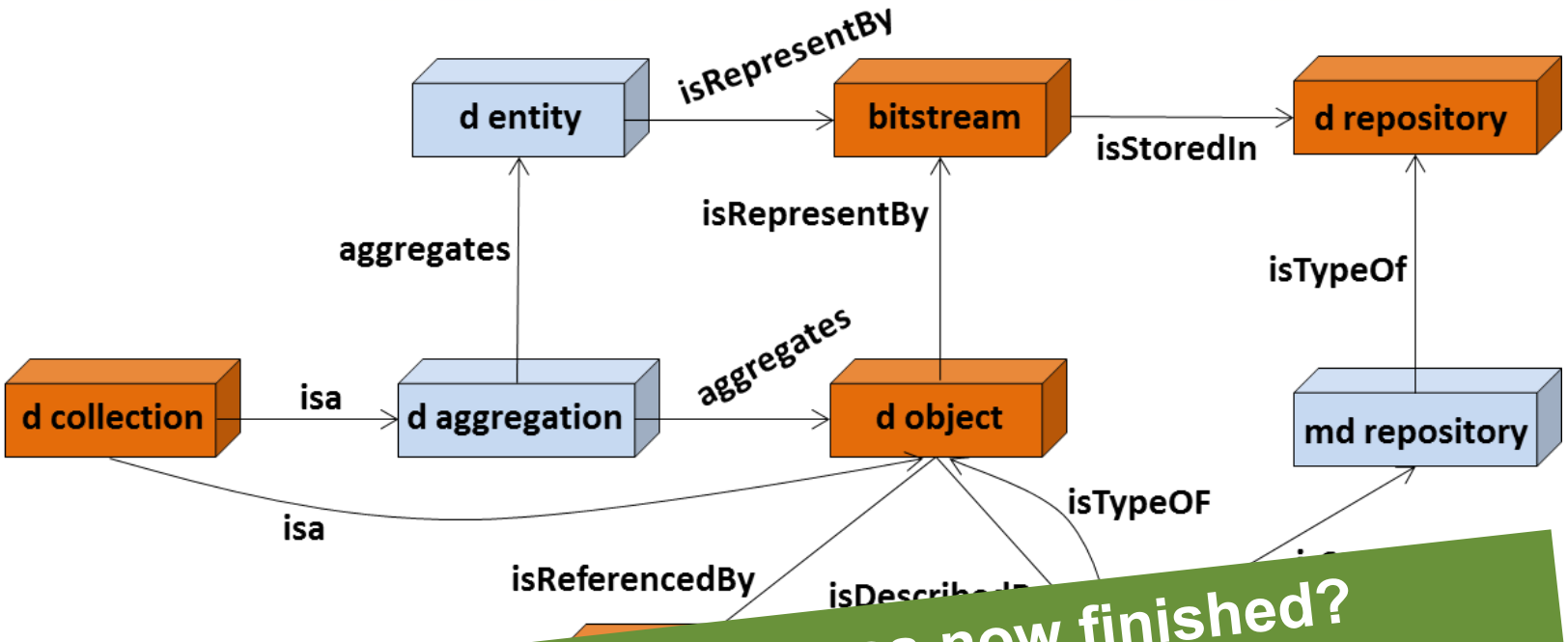
Content

1. Principles & Trends
2. Data Practices
3. Data Foundation & Terminology WG
4. Data Fabric IG

DFT Results: simple common data model



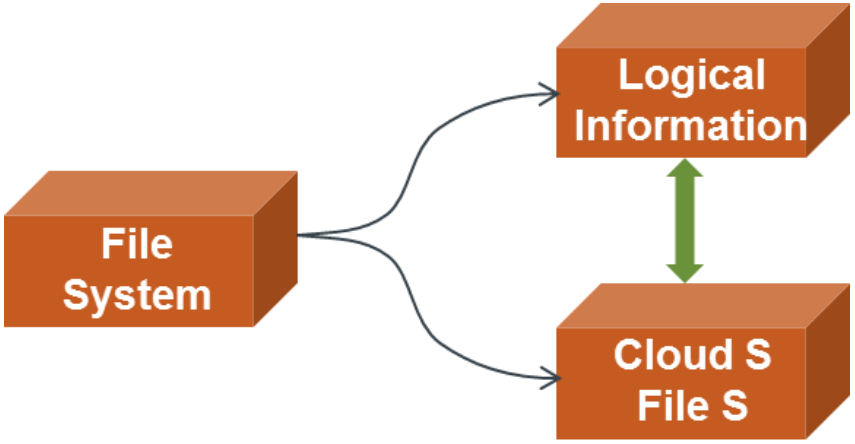
DFT Results: simple common data model

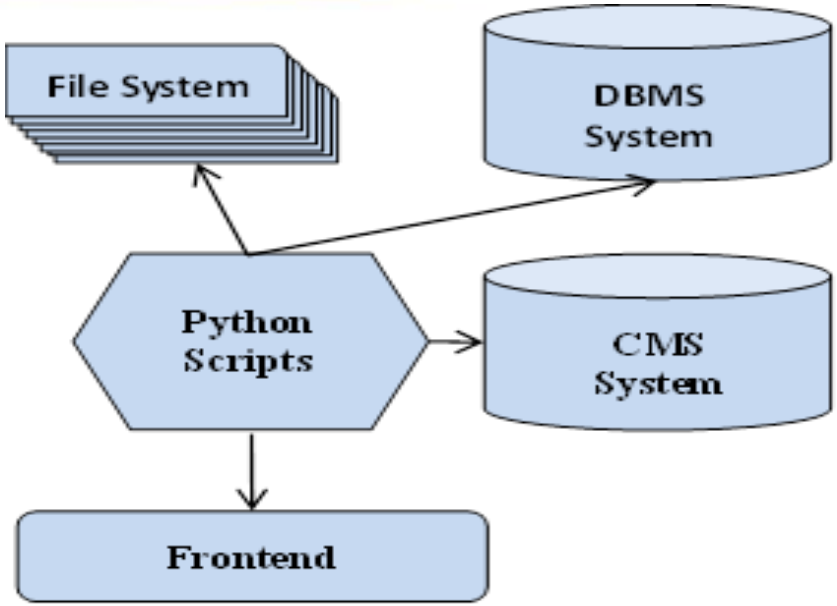


Is this definition process now finished?
No – term tool can be used to discuss

D
A ... identifier is a long-lasting ID represented by a string that uniquely identifies a DO and that is intended to be persistently resolved to meaningful state information about the identified DO.
Note: We use the term Persistent Resolvable Identifier as a synonym.

Impact of DFT Result





"logical" info (PIDs, MD, provenance, rights, relations) not harmonized & no guidance

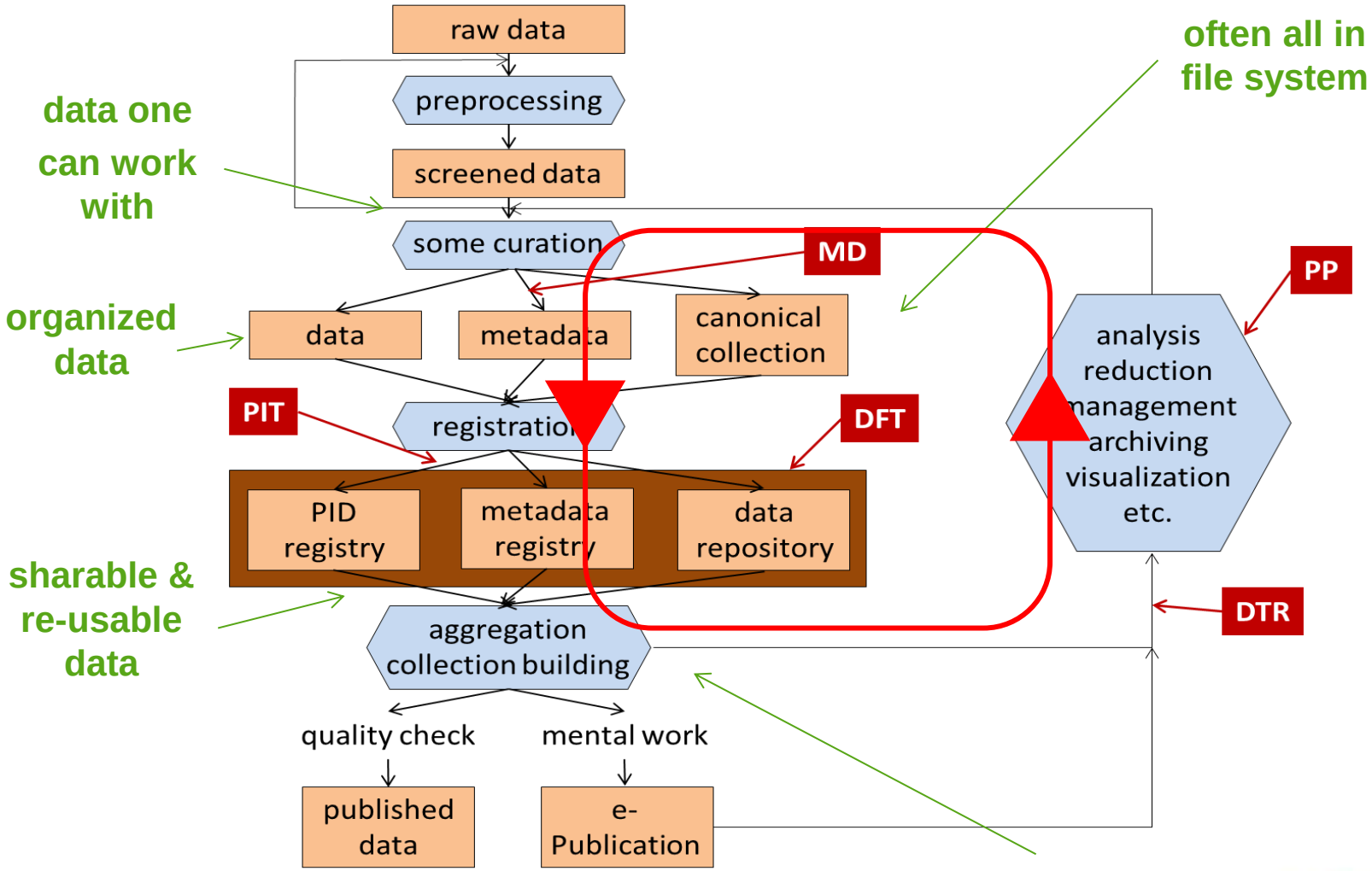
data federation including "logical" information not scalable and not cost-effective

If we all agree on same simple model
we could write adaptors or include it in development
federating data would become efficient

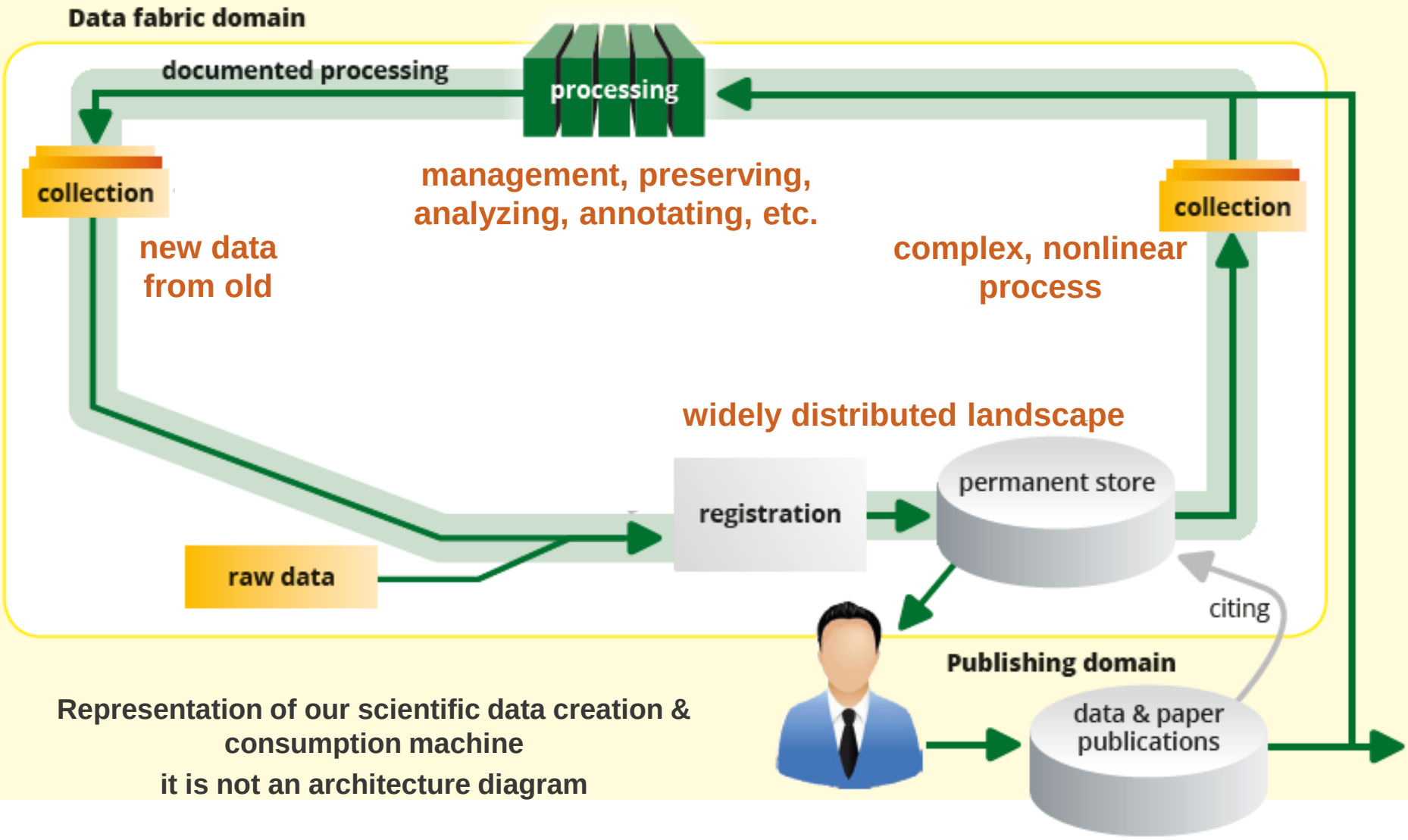
DFIG

1. Principles & Trends
2. Data Practices
3. Data Foundation & Terminology WG
4. Data Fabric IG

Data Practices – processing steps



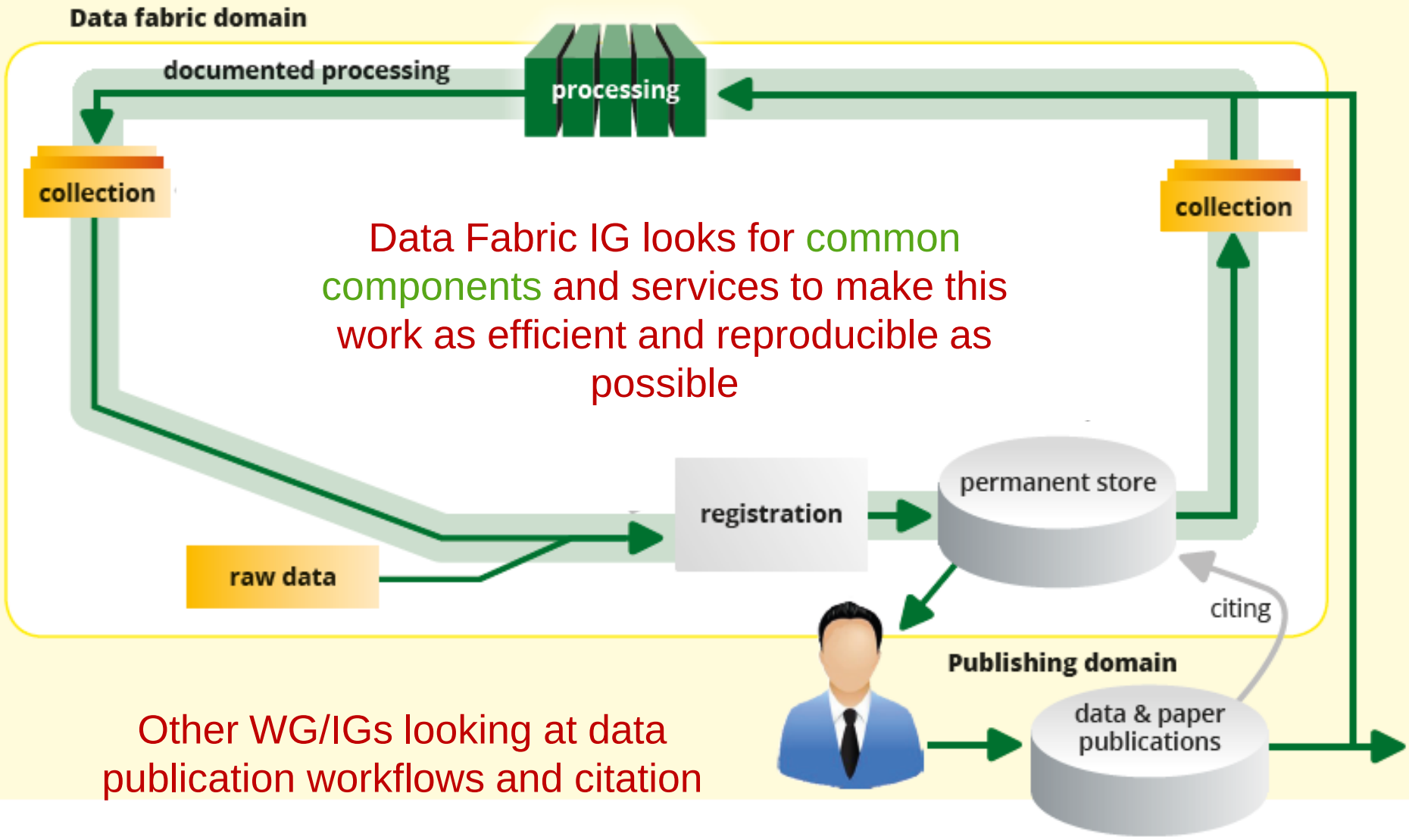
Abstract Data Cycle in the Labs



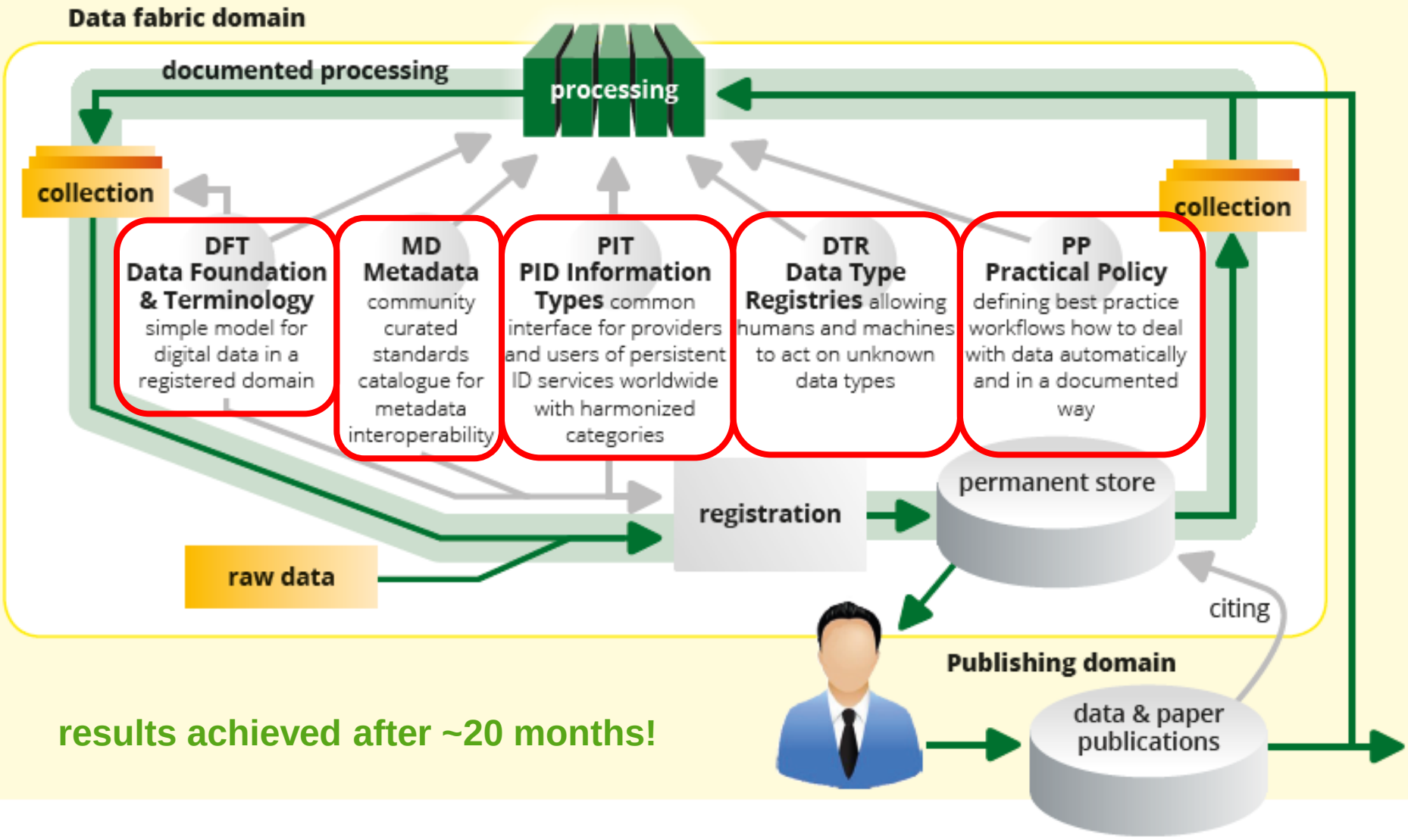
Representation of our scientific data creation & consumption machine

it is not an architecture diagram

Data Fabric Interest Group

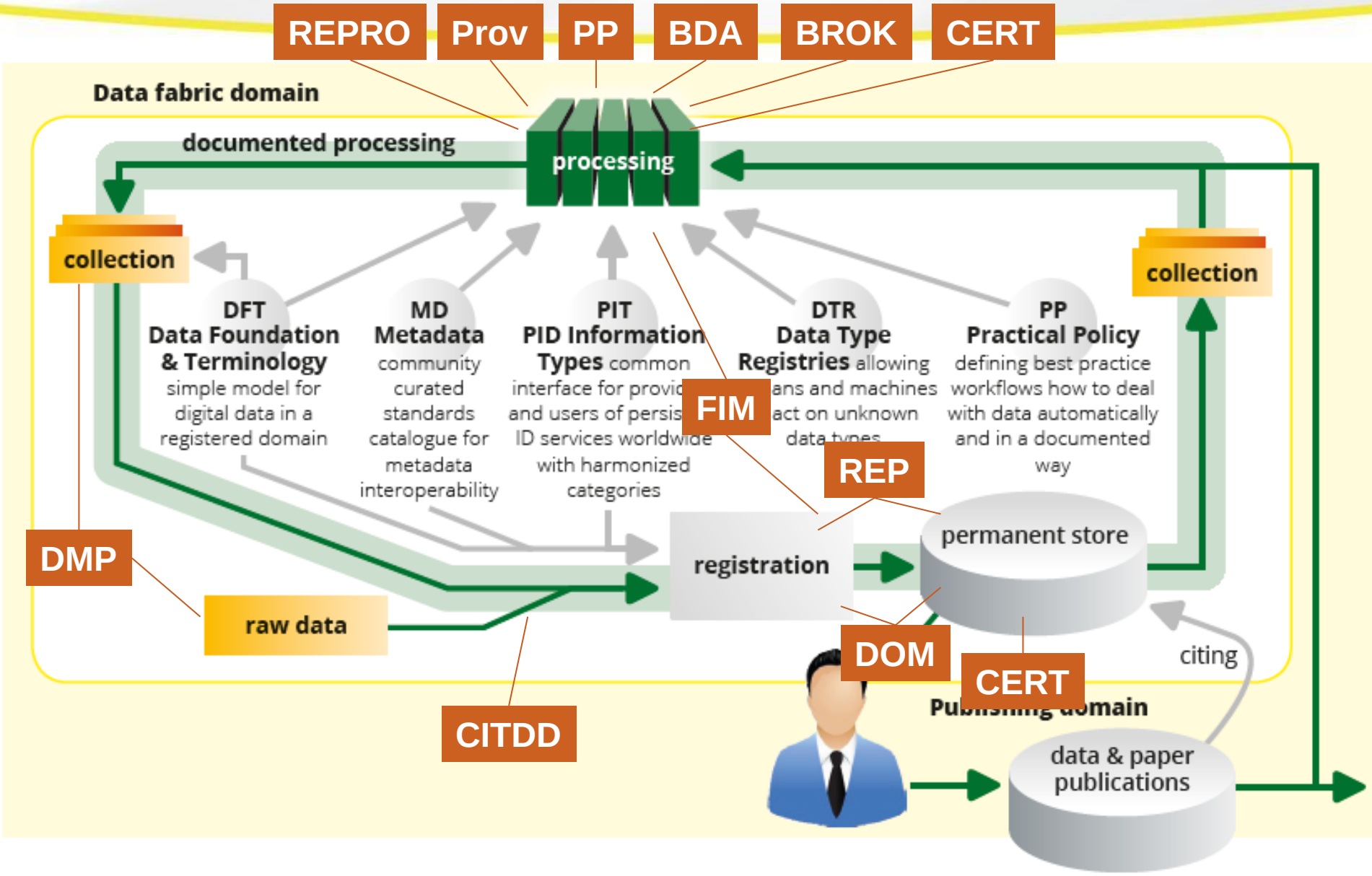


RDA – first Working Group results



results achieved after ~20 months!

DFIG – grouping of WG/IGs



Base analysis on Use Cases - ~20 so far

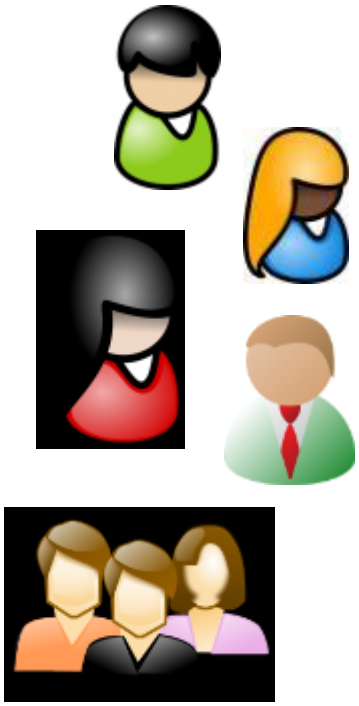
● environmental science ● natural science ● life science ● humanities, soc. sciences ● IT, various

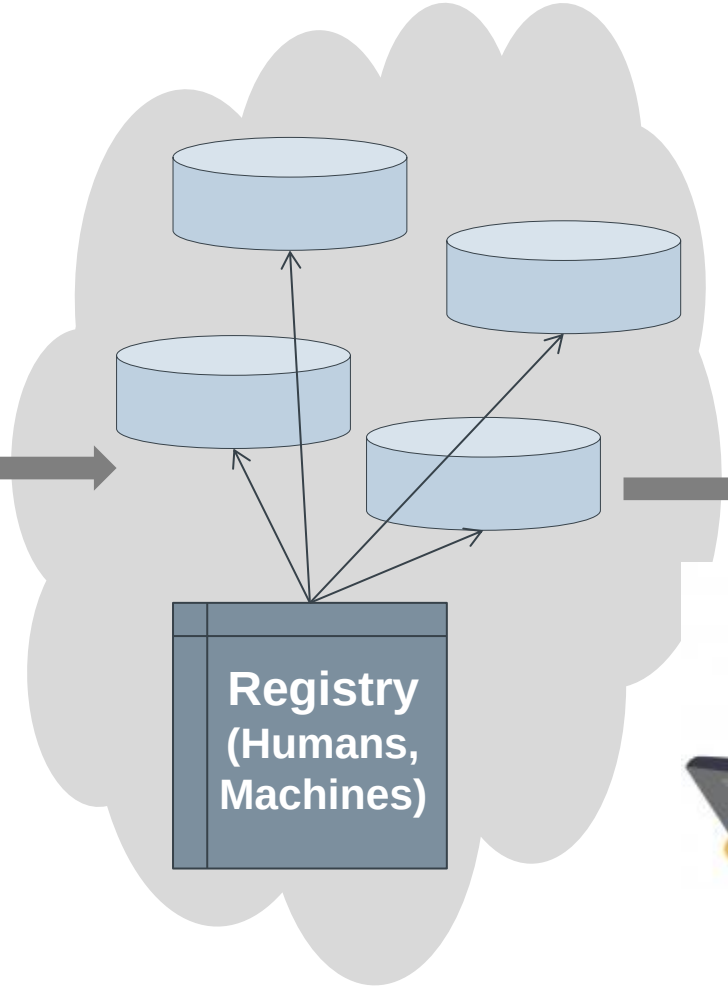
all indicated nodes are centers of national, regional and even worldwide federations

DFIG Spinoff – Repository Registry

Safe Deposit

trusted Re-use
valid References
reproducible Science
machine usage





Scientists

Funders

Publishers

Domain of Trusted
Repositories

State of DFIG

- White Paper open for discussion (incl. what DFIG is about and what not)
- Have received ~ 20 Use Cases so far – on the Wiki
- First analysis of components needed – Position Paper
- Repository and Collection Registry as first Spin Offs

Describe and Upload your Use Case.

Comment on the position paper – tomorrow.

References

- Data Management Trends, Principles and Components - What Needs to be Done Next? V6.1: http://hdl.handle.net/11304/992fe6a0-fe34-11e4-8a18-f31aa6f4d448
- Principles for Data Sharing and Re-use: are they all the same? http://hdl.handle.net/11304/1aab3df4-f3ce-11e4-ac7e-860aa0063d1f
- Living with Data Management Plans http://hdl.handle.net/11304/ea286e5a-f3d1-11e4-ac7e-860aa0063d1f
- RDA Europe: Data Practices Analysis *http://hdl.handle.net/11304/6e1424cc-8927-11e4-ac7e-860aa0063d1f*
- DFT: https://rd-alliance.org/groups/data-foundation-and-terminology-wg.html
- Data Fabric: https://rd-alliance.org/group/data-fabric-ig.html
- Data Fabric Wiki: https://rd-alliance.org/node/44520/all-wiki-index-by-group

Thanks for your attention.

http://www.rd-alliance.org
http://europe.rd-alliance.org