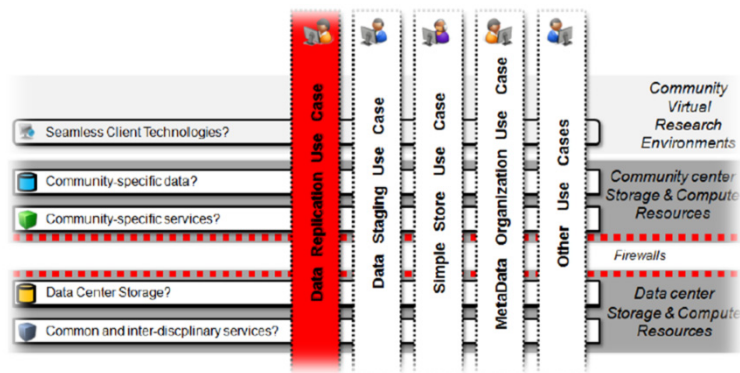# 'Big Data' in GLUE2

Thoughts using GLUE2 in Data-Driven Infrastructures
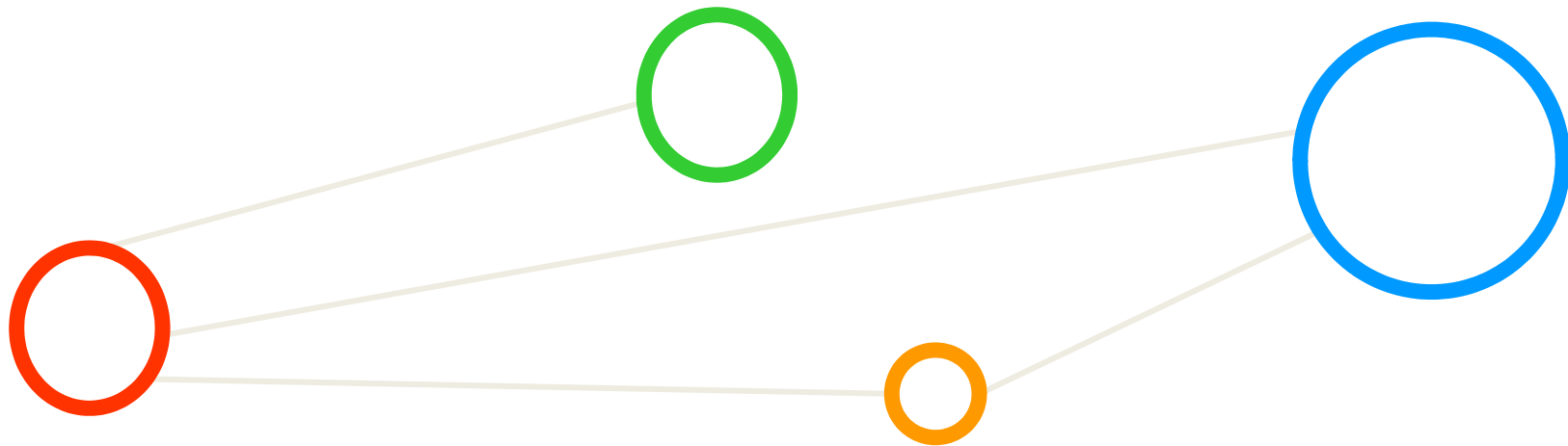

Morris Riedel et al.
Juelich Supercomputing Centre
EUDAT Scientific Community Coordinator

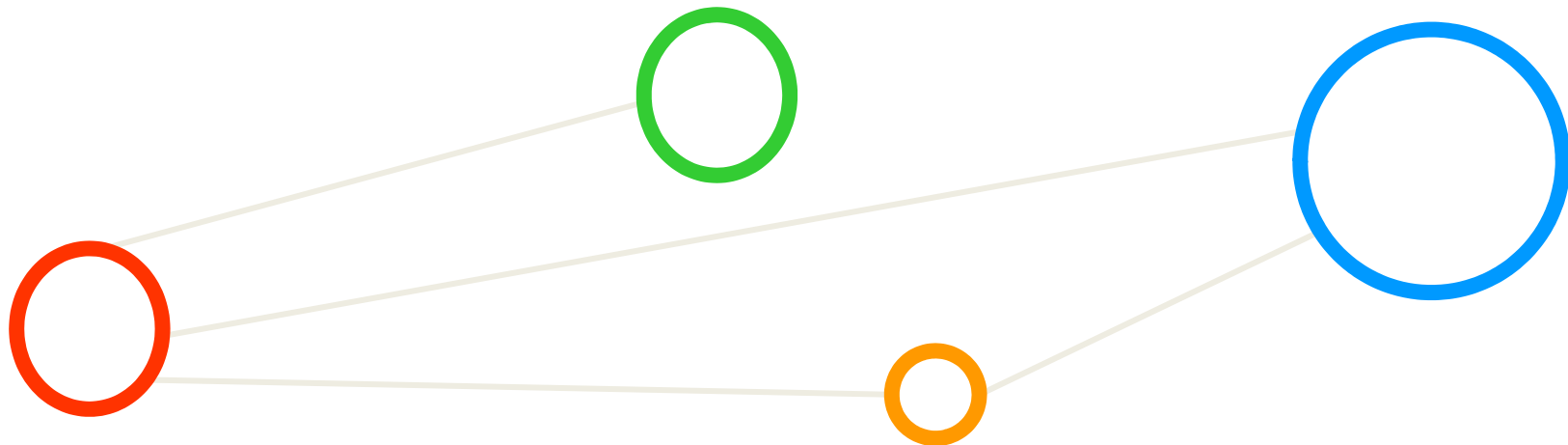OGF37, Charlottesville, 12th March 2013

Outline

Outline

- EUDAT 101
- Thoughts about using GLUE2

EUDAT 101

European Collaborative Data Infrastructure

- European Commision Project
- Start: 10/2011, Duration: 3 years
- Funds: 16,3 M€ (9,3M€ EC)

- Create a long-term oriented collaborative federated data infrastructure for science
- Simple, secure, and registered policy-based storage of data
- Enable data access for geographically-dispersed teams
- Support multi-disciplinary service re-use for science
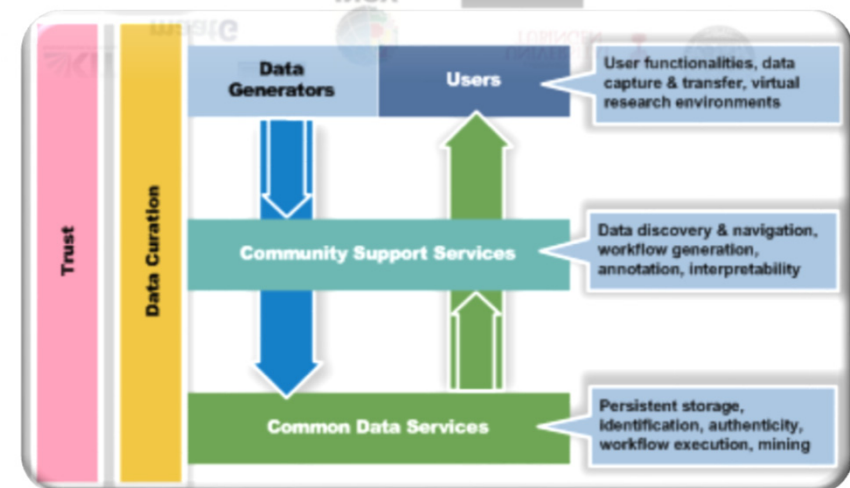
Selected Scientific Communities

EPOS – European
Plate Observing System

CLARIN – Common
Language Resources and
Technology Infrastructure

ENES – Service for
climate modeling

VPH – The Virtual
Physiological Human

- Distributed data sensors
- About 200 centers in EU
- About 20 centers in EU
- Pilot project with 5 hospitals

LifeWatch – Biodiversity Data and Observatories

- Distributed data sensors
- Metadata standardisation
- Interoperability reqs
- Involving most nature infrastructures
- Common reference model

http://envri.eu/

http://creative-b.eu/

LifeWatch will construct and bring into operation the facilities, hardware, software and governance structures for all aspects of biodiversity research. Facilities for data generation and processing, data integration and interoperability. A network of observatories, virtual laboratories. A Service Center supporting scientific and policy users.
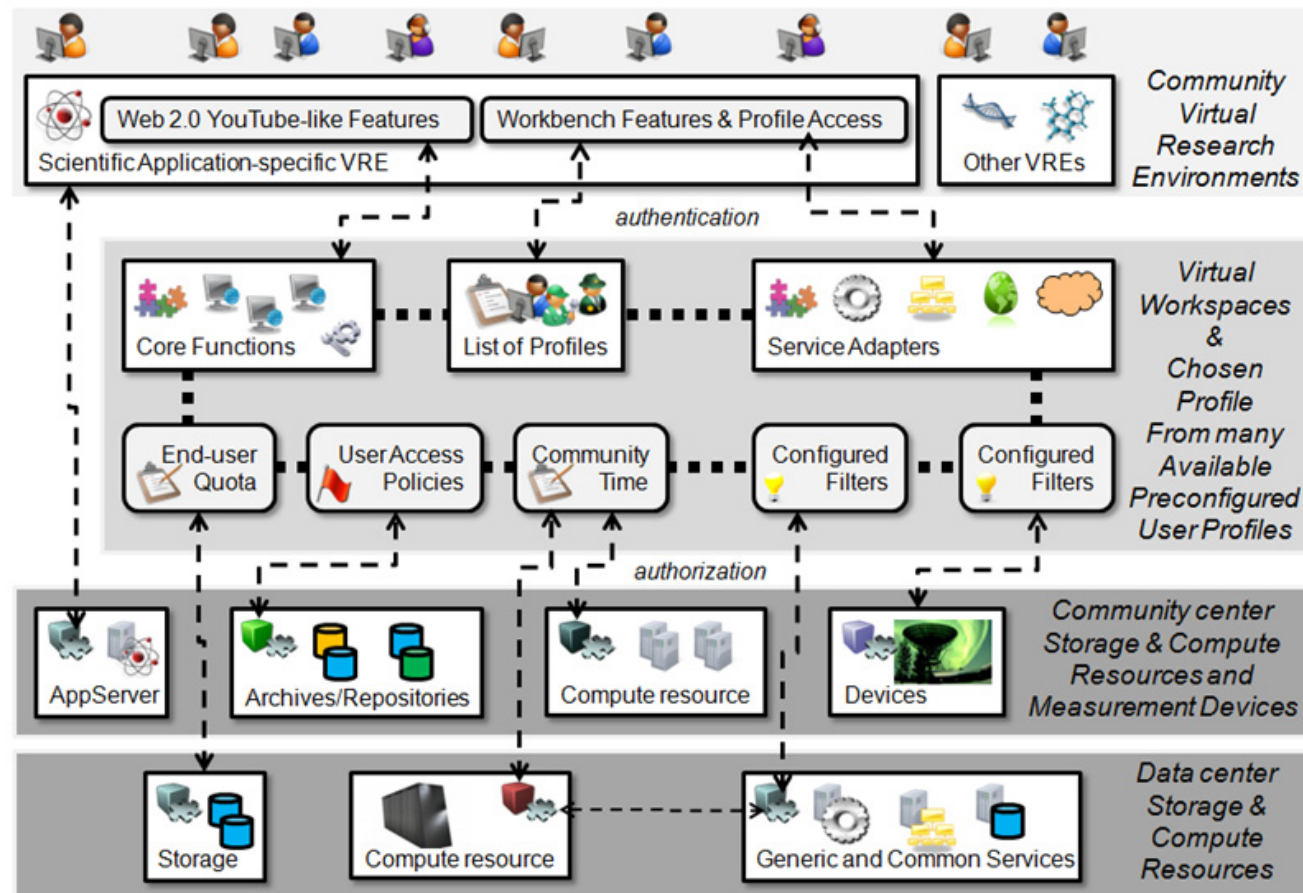
Motivational 'ScienceTube' Vision
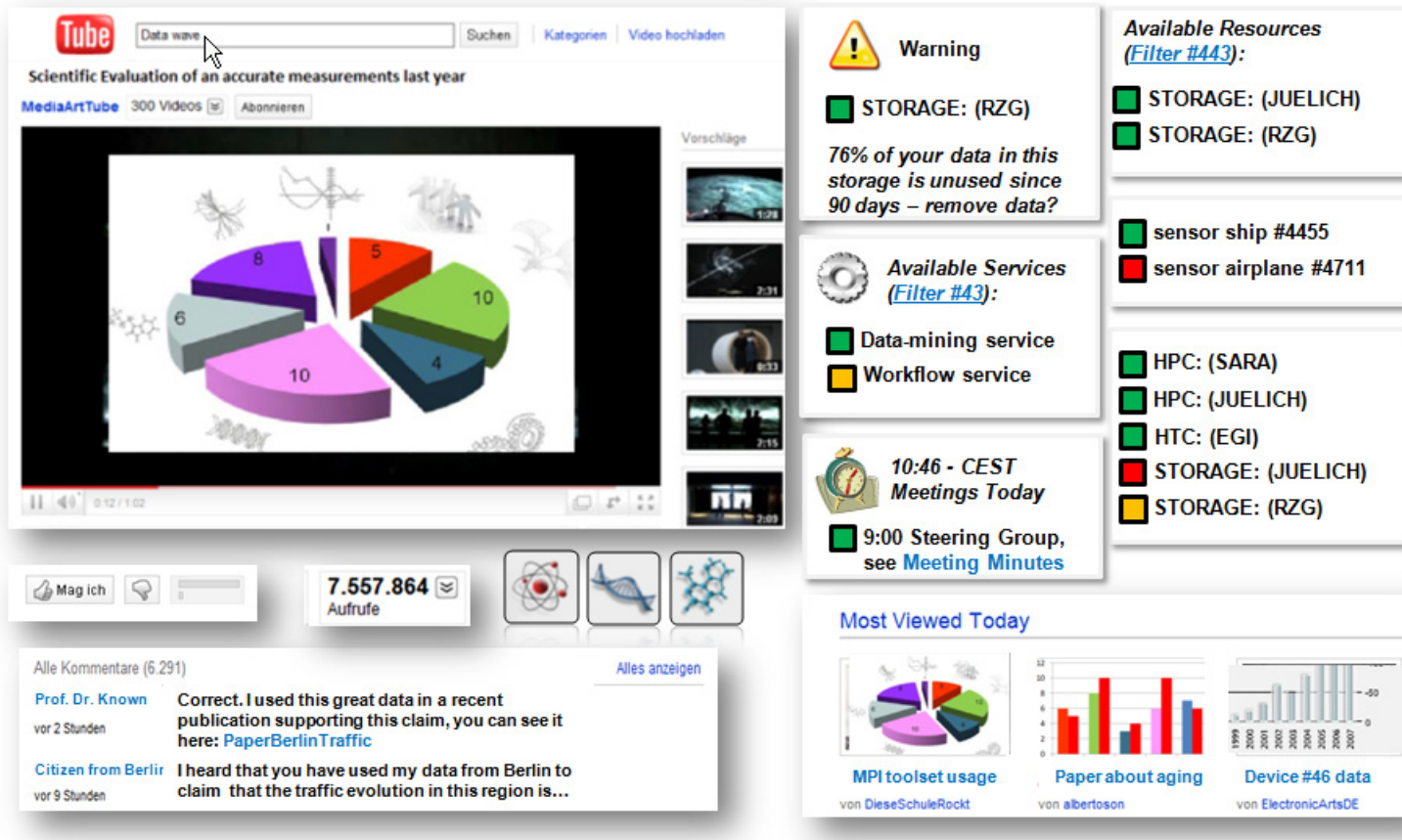
Riedel and Wittenburg et al., 'A data infrastructure reference model with applications: towards realization of a ScienceTube vision with a data replication service', Journal of Internet Services & Applications, Vol. 4, Issue 1, 2013
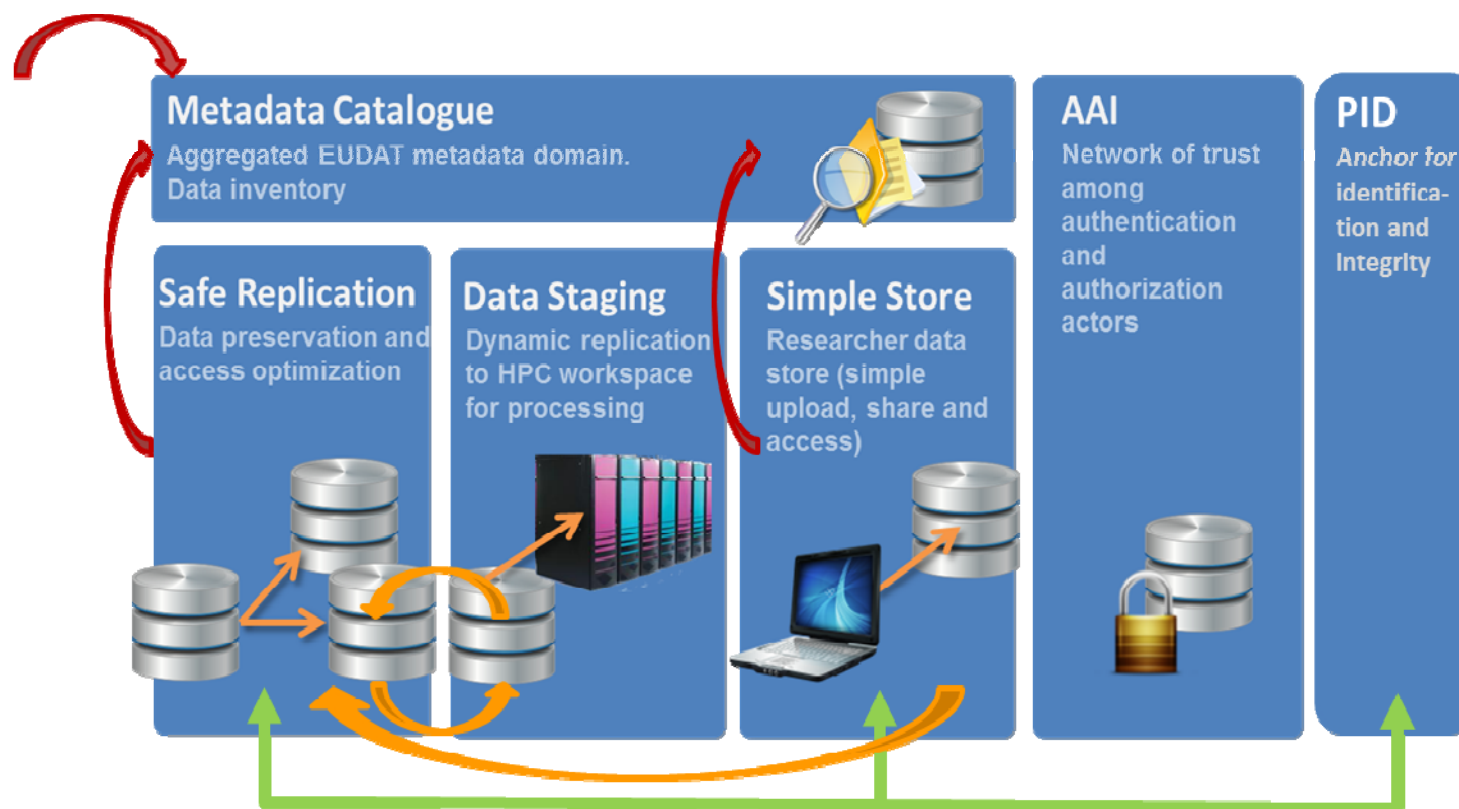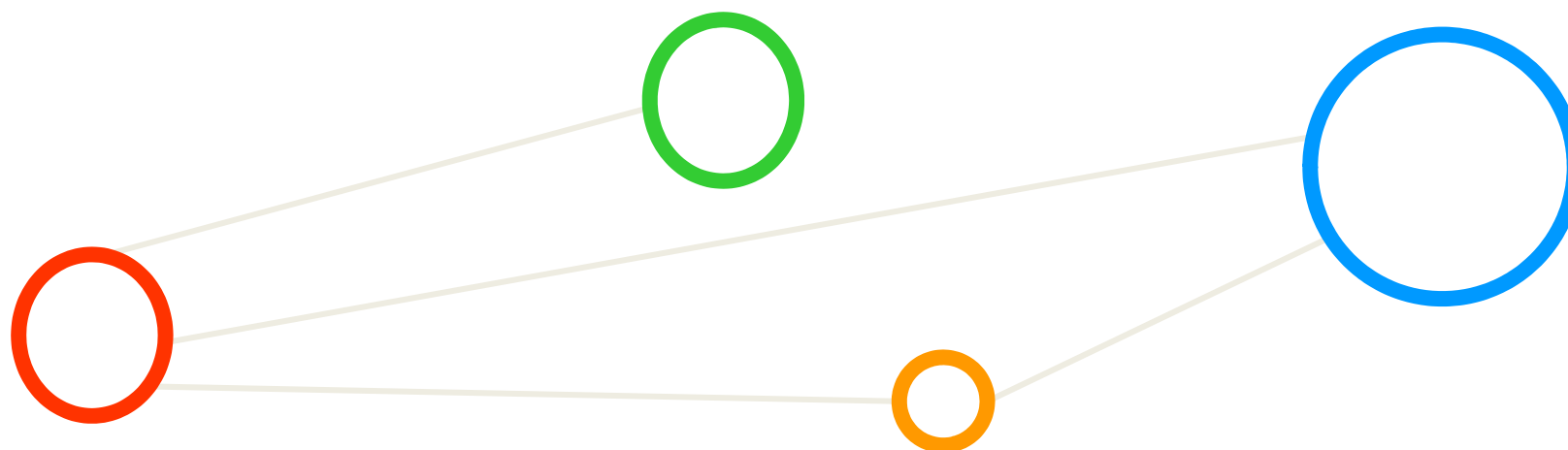
Reference Model Core Building Blocks

Riedel and Wittenburg et al., 'A data infrastructure reference model with applications: towards realization of a ScienceTube vision with a data replication service', Journal of Internet Services & Applications, Vol. 4, Issue 1, 2013

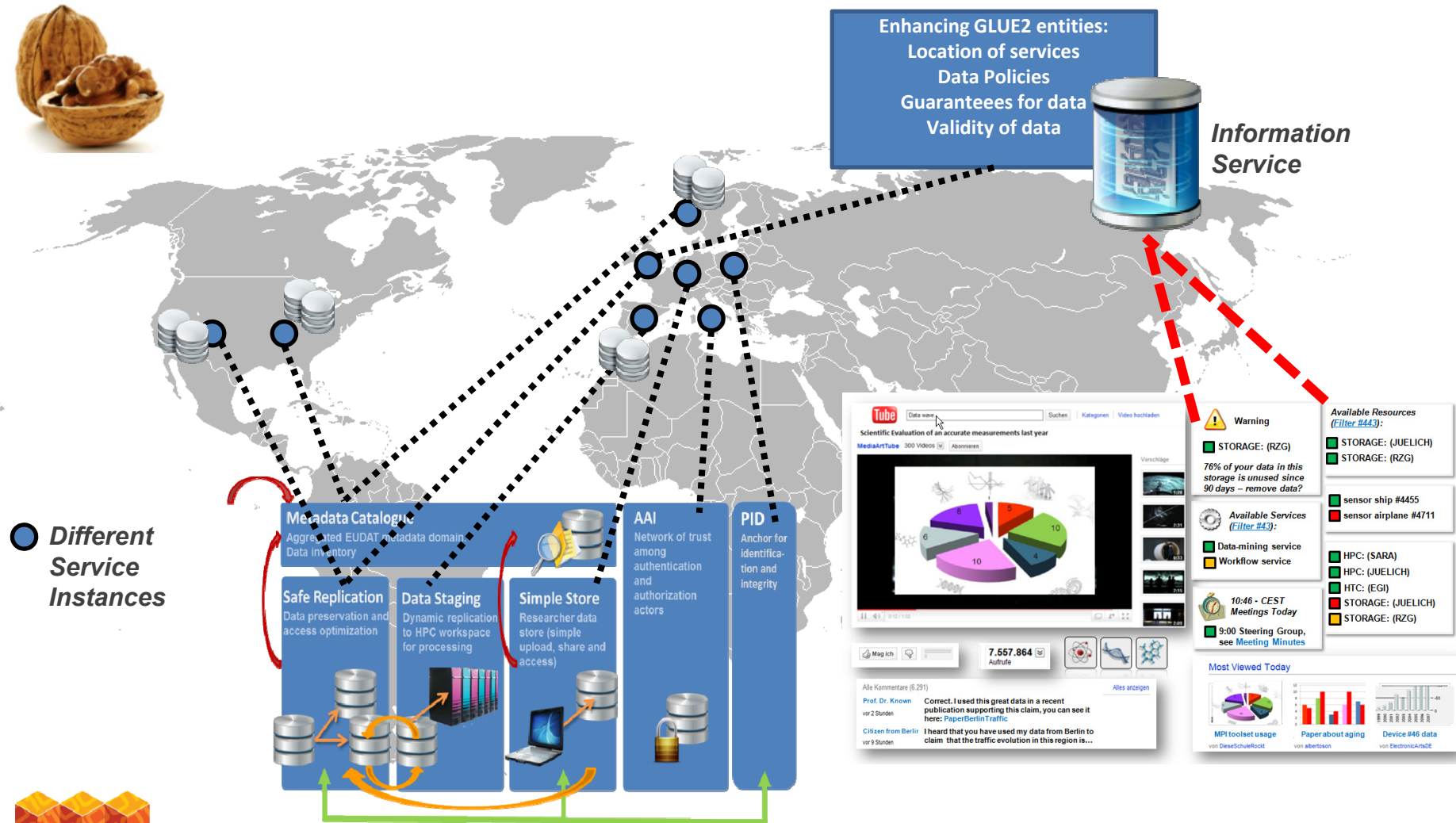# Initial Set of EUDAT Service Types

Thoughts about using GLUE2

GLUE2 to describe EUDAT services?

Need for new GLUE2 Entities/Attributes?

- Compute and Storage Services are good to be used already
- Several more data-driven services might be good to add
  - Commonly shared data and community-specific services (e.g. CLARIN/WebLicht)
- Data Replication Service
  - [Note it is not about metadata of scientific data, but about the services itself]
  - E.g. guarantees: data is preserved for 50 years guaranteed by Max Planck Society
  - E.g. policy: data is automatically replicated to JSC, EPCC, and SARA
  - E.g. storage space: XYZ PB community storage space available
  - E.g. location and access: URI of service, security levels required (e.g. read-only)?
  - E.g. Ingested scientific data automatically receives Persistent Identifier type 'handle'
  - E.g. Hints on tertiary storage impacts on restoring (many small files can take ages)
- Similar attributes/entities for other services alongside existing computing
  - Metadata Service, Data-Staging Service, Simple Store Services,
  - Persistent Identifier Services
  - [Different? AAI Services maybe in the future]

Thanks