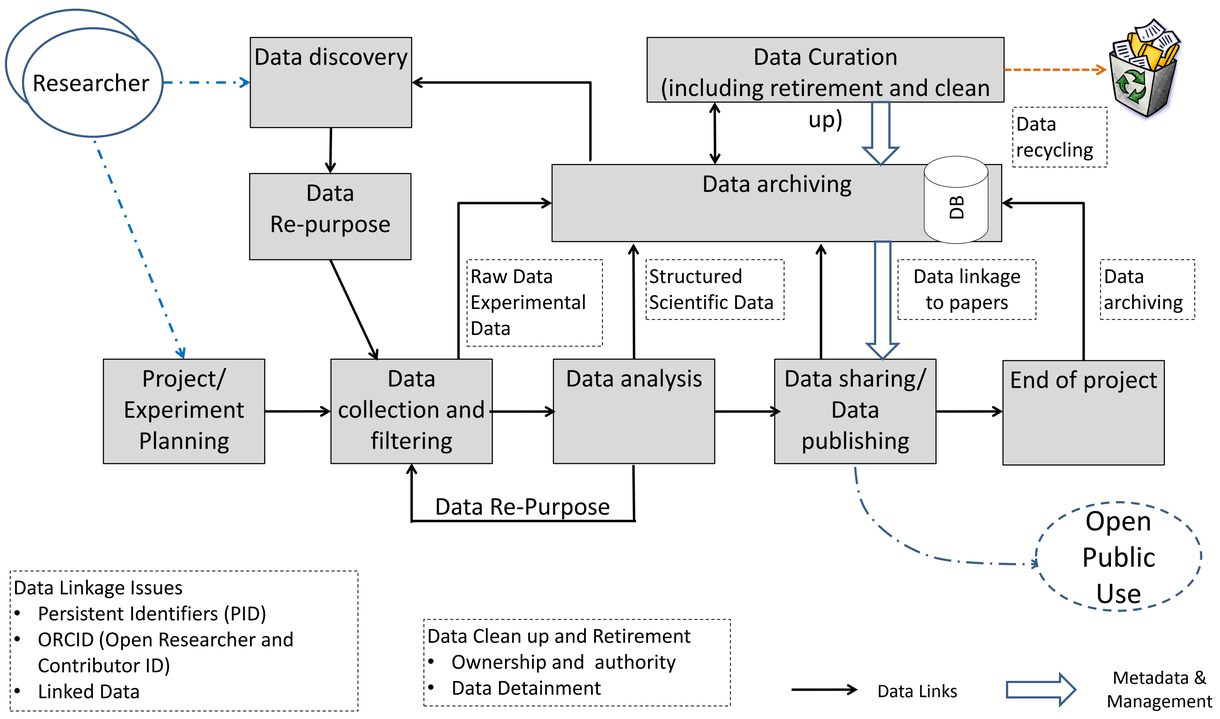
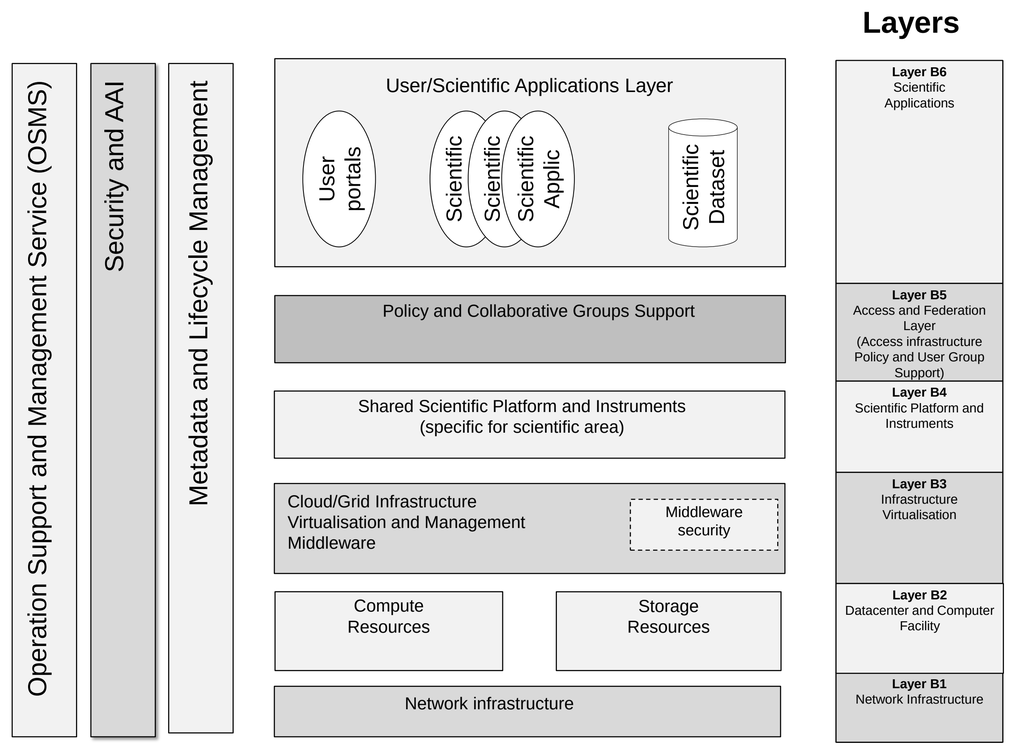
Addressing Big Data Challenges for e-Science Infrastructure

Yuri Demchenko, Paola Grosso, Cees de Laat (UvA)

Scientific Data Lifecycle Management Model (SDLM)



SDI Architecture Multilayer Model



Scientific Data Types

- **Raw data** collected from observation and from experiment (according to an initial research model)
- **Structured data** and datasets that went through data filtering and processing (supporting some particular formal model)
- **Published data** that supports one or another scientific hypothesis, research result or statement
- **Data linked to publications** to support the wide research consolidation, integration, and openness.

e-Science features

- Automation of all e-Science processes including data collection, storing, classification, indexing and other components of the general data curation and provenance.
- Transformation of all processes, events and products into digital form by means of multi-dimensional multi-faceted measurements, monitoring and control; digitising existing artifacts and other content.
- Possibility to re-use the initial and published research data with possible data re-purposing for secondary research.
- Global data availability and access over the network for cooperative group of researchers, including wide public access to scientific data.
- Existence of necessary infrastructure components and management tools that allows fast infrastructures and services composition, adaptation and provisioning on demand for specific research projects and tasks.
- Advanced security and access control technologies that ensure secure operation of the complex research infrastructures and scientific instruments and allow creating trusted secure environment for cooperating groups and individual researchers.

General requirements to SDI for emerging Big Data Science

- Support for long running experiments and large data volumes generated at high speed
- Multi-tier inter-linked data distribution and replication
- On-demand infrastructure provisioning to support data sets and scientific workflows, mobility of data-centric scientific applications
- Support of virtual scientists communities, addressing dynamic user groups creation and management, federated identity management
- Support for the whole data lifecycle including metadata and data source linkage
- Trusted environment for data storage and processing
- Support for data integrity, confidentiality, accountability
- Policy binding to data to protect privacy, confidentiality and IPR

Contributing projects

AAA Study: Study on AAA Platforms For Scientific data/information Resources in Europe - https://confluence.terena.org/display/aaastudy/AAA+Study+Home+Page

GEYSERS – Generalised Architecture for Infrastructure services - http://www.geysers.eu/

GEANT3 JRA3 Task 3 – Composable services (GEMBus) - http://www.geant.net/

Scientific Data Infrastructure Model Layers

Layer D1: Network infrastructure layer represented by the general purpose Internet infrastructure and dedicated network infrastructure

Layer D2: Datacenters and computing resources/facilities

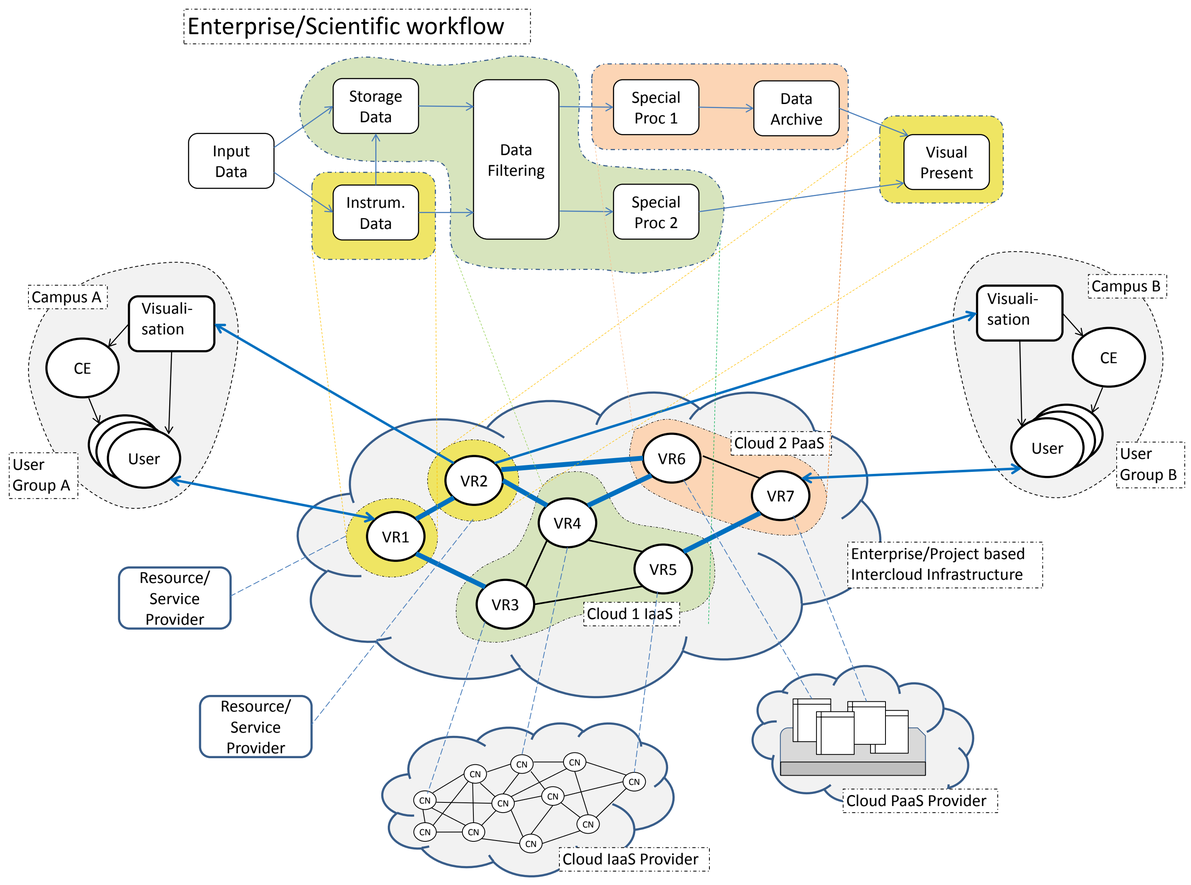
Layer D3: Infrastructure virtualisation layer that is represented by the Cloud/Grid infrastructure services and middleware supporting specialised scientific platforms deployment and operation

Layer D4: (Shared) Scientific platforms and instruments specific for different research areas

Layer D5: Access Infrastructure and Federation Layer that includes federation infrastructure components, policy and collaborative user groups support functionality

Layer D6: Scientific applications and user portals/clients

Cloud based Scientific Data Infrastructure Provisioning



Virtual Infrastructure Components

- Cloud IaaS segment (VR3-VR5) and Cloud PaaS segment (VR6, VR7)
  - Typically provisioned by the Cloud Service Providers
- Virtualised resources or services (VR1, VR2), e.g. instruments or visualisation (streaming) devices
- Interacting campuses A and B containing their own resources, network infrastructure and campus based infrastructure services
- Dedicated network infrastructure interconnecting virtualised resources, cloud domains and campuses

Credits: Yuri Demchenko, Paola Grosso, Cees de Laat
Contact: Yuri Demchenko <y.demchenko@uva.nl>

GEYSERS

UvA Universiteit van Amsterdam