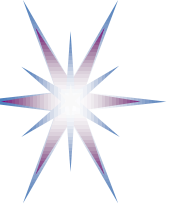


Технологические вызовы науки “больших данных” и Европейские проекты в области e-Science

Big Data Science Challenges and European e-Science
Projects

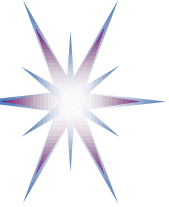
Yuri Demchenko,
SNE Group, University of Amsterdam

TiPVSIT2012
13-20 August 2012, Ulan-Ude, Russia



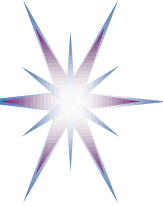
Outline

- Наука “больших данных” (Big Data Science) как следующая технологическая волна
 - Категории данных и жизненный цикл данных в e-Science
(Data categories and Data Lifecycle in modern e-Science)
- Европейская политика в области научной и технологической информации (НТИ) (EC policy on Scientific and Technology Information (STI))
 - Публичная научная информация и Открытый Доступ
(Public Scientific Information and Open Access)
 - Horizon 2020 Consultation meeting in Rome 11-12 April 2012
- Инфраструктура научных данных по отраслям, проекты и инициативы в Европе (Scientific Data Infrastructures (SDI) by scientific communities, projects and initiatives in Europe)
- Архитектура Инфраструктуры научных данных (Defining SDI architecture)
 - Облачные технологии как инфраструктурная база для сложных.больших данных
Clouds as an infrastructure platform for complex/scientific data
 - (Дополнительно) Построение доверительной инфраструктуры для науки
(Defining Trustworthy platform fro SDI)
- Необходимость новой образовательной специальности (Need for new scientific and academic discipline)



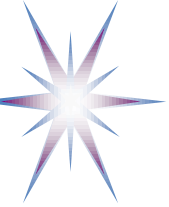
System and Network Engineering (SNE) Group at the University of Amsterdam

- SNE group is primarily a research group but also supports SNE master education
- Main research areas
 - High speed optical networks
 - Recent testbed achieved sub-40Gbps at Amsterdam-CERN link
 - Information modeling for network and infrastructure services description
 - Security and generic AAA Authorisation framework (GAAA-AuthZ)
 - Evolving from client/security model to dynamically provisioned services
- New research areas
 - Cloud and Inter-Cloud Architecture for interoperability and integration
 - Infrastructure and security issues in Big Data Science and Scientific Data Infrastructure
- Long term research cooperation with SURFnet and GigaPort programs in NL
- Recent and current projects participation – DatGrid, NextGrid, EGEE, Phosphorus, GEYSERS, GEANT3, NOVI, AAA Study
- Active contribution to standardisation activity at OGF, IETF, IEEE, TMF, NIST and others
- Maintaining own optical networking testbed and cloud IaaS and PaaS testbeds



Big Data Science as the next technology focus

- Big Data is becoming the next buzz word
- Based on entire information and artifacts digitising
 - Requires also new information and semantic models for information structuring and presentation
 - Requires new research methods using large data sets and data mining
 - Methods to evolve and results to be improved
- Changes the way how the modern research is done (e-Science)
 - Secondary research, data re-focusing, linking data and publications
- Big Data require **infrastructure** both to support distributed data (collection, storage, processing) and metadata/discovery services
 - Demand for trusted/trustworthy infrastructure
 - Clouds provide just right technology for (data supporting) infrastructure virtualisation



Horizon2020 Consultation Meeting (Rome 11-12 April 2012)

- **Vice-President of the European Commission Mme Neelie Kroes**
 - *"Open e-Infrastructures for Open Science" - The Digital Agenda and Access to Scientific Information*
- **Working Group 1: Open Global Data Infrastructure: towards an international framework for collaborative scientific data infrastructure**
 - Several contributing "position papers" produced about the subject of a "Data Web Forum" or a "Data Access and Interoperability Task Force" and reports such as Riding the Wave from the High-Level Group on Scientific Data or the report of the G8+5 working group on Data.
- Working Group 2: Open Scientific Content: e-Infrastructure policies and services to support access, storage, processing and exchange of scientific information
- Working Group 3: Open Research Culture: Engagement of researchers and society with Open Science through collaborative data infrastructure.

Declaration from Rome meeting

“Researchers and practitioners from any discipline are able to find, access and process the data they need in a timely manner. They are confident in their ability to use and understand data, and they can evaluate the degree to which that data can be trusted.

Data are stored, managed, shared, and preserved in a way that optimizes scientific discovery, innovation, and societal benefit. Where appropriate, producers of data benefit from opening it to broad access and routinely deposit their data in reliable repositories. A framework of repositories work to international standards, to ensure they are trustworthy.”

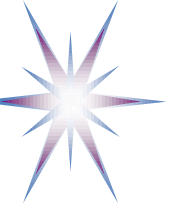
Next EU Framework Program FP 8 Horizon212 2013-2020



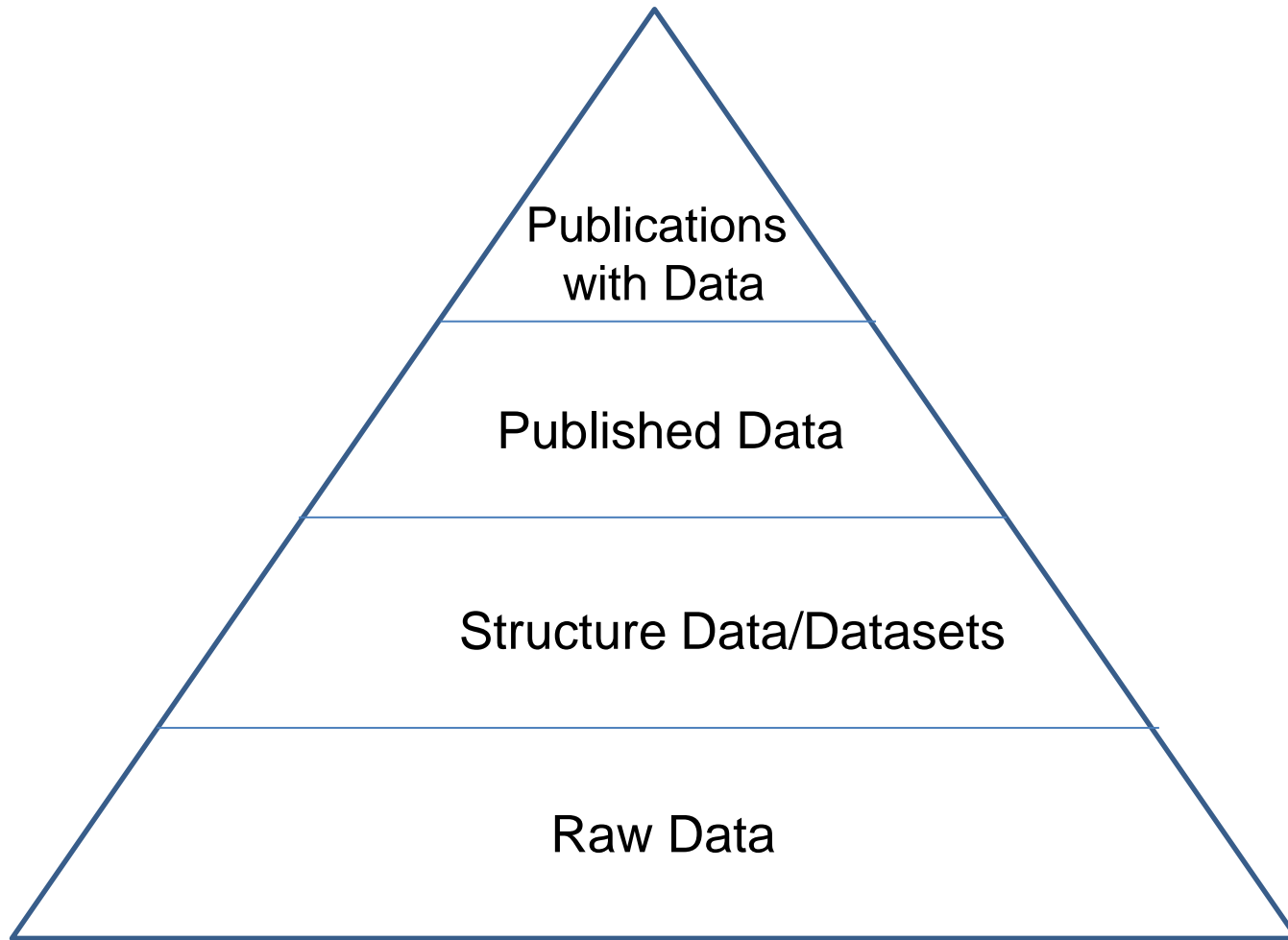
Scientific Research and Scientific Data

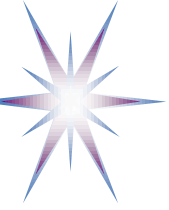
Scientific method is a body of techniques for investigating phenomena, acquiring new knowledge, or correcting and integrating previous knowledge

- Two basic research methods
 - Observation (passive, selective, active)
 - Experiment (on existing object, with object re-creation)
- Scientific instruments and Measurements
 - Measurements as a basic instrumental method and building block of science
 - Science begins when it can measure and record data

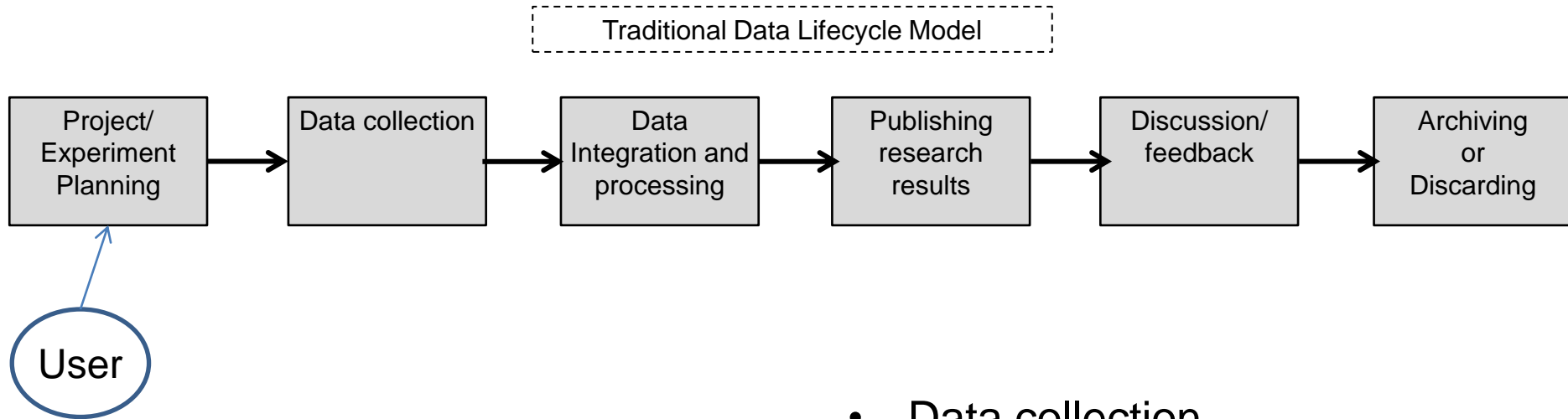


Scientific Data Model/Hierarchy



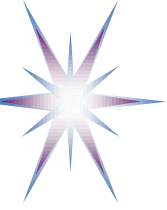


Traditional Data Lifecycle Model



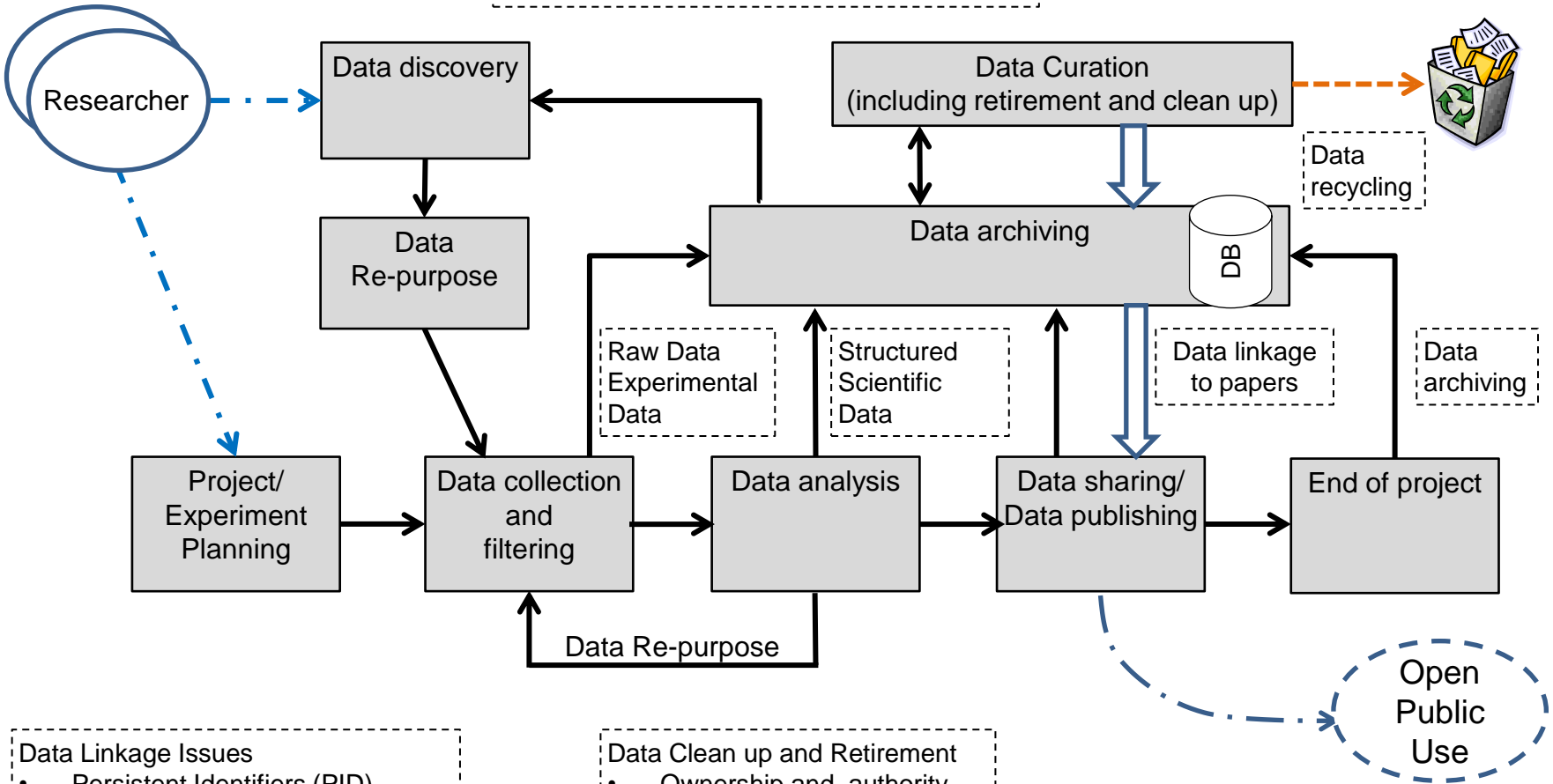
- Data collection
- Data processing
- Publishing research results
- Discussion
- Data and publications archiving

Lack of initial data preservation and data linking to publications



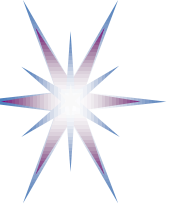
Data Lifecycle Model in e-Science - II

Data Lifecycle Model in e-Science



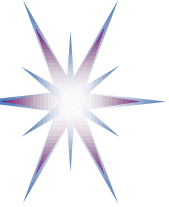
- Data Linkage Issues
- Persistent Identifiers (PID)
 - ORCID (Open Researcher and Contributor ID)
 - Lined Data

- Data Clean up and Retirement
- Ownership and authority
 - Data Detainment



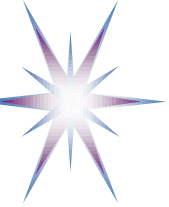
Big Data Science (1)

- High Energy Physics (HEP)
 - Running experiment LHC and infrastructure WLCG
 - Already producing PBytes of information
 - Worldwide distribution and processing
 - CERN and national HEP centers
 - Projects: WLCG (former EGEE), EGI (European Grid Infrastructure)
- Low Energy Physics and Material Science (photon, proton, laser, spectrometry)
 - Number of research facilities serving international communities
 - Multiple short projects producing TBytes of information
 - Experimental data storage, identification, trusted access to multiple users (including public and private researchers)
 - Projects: ELIXIR and many small projects
- Earth, weather and space observation
 - Climate research and Earth observation
 - With new 4? satellites to be launched starting 2017 to produce PBytes monthly
 - ESA (European Space Agency)
 - Projects: Helix Nebula



Big Data Science (2)

- Life science and biodiversity (Genomic, Biomedical and Healthcare research)
 - Human genome (EMBL-EBI)
 - Currently centralised databases but evolving to distributed
 - ELSI data - Special requirements to data integrity and privacy
 - Numerous local/offline databases to be brought online
 - Living species and biodiversity
 - Mobile/field access, filtering and on-demand computing
 - Public contribution, vocational or citizen researchers
 - Projects: ELIXIR, EGA, LifeWatch
- Humanities (History, languages, human behaviour)
 - Rediscovering research with total information digitising
 - Expected huge amount of data to digitise all human heritage
 - Very spread research community
 - Projects: CLARIN, DARIAH, EUDAT
- Additional: Data collection from sensors and online activities
 - Intelligence, Homeland Security, Log data (+ Facebook data :-)



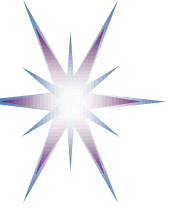
Big Data Challenges and Initiatives

- Peta and Exa scale problems
 - Storage, Computing, Transfer/Network
 - International Exascale Software Project (<http://www.exascale.org/>)
 - A Vision for Global Research Data Infrastructure (<http://www.grdi2020.eu/>)
- G8+O5 Global Research Data Infrastructure, Subgroup on Data, Draft Report, 28 October 2011
- Data Web Forum (DWF)
 - Initiated by NSF (Alan Blatetsky)
 - Launch is planned for autumn 2012 NSF event
- DAITF – Data Access and Interoperability Task Force
 - Initiated by EUDAT project (in association with CLARIN)
 - Primarily not by e-Science/Grid community
 - Concertation meeting – Sept 2011, Lyon
 - Trying to follow IETF format



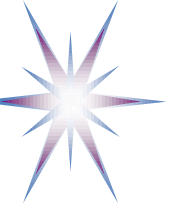
Coordination in European Research Area (ERA)

- European Commission – but not only
- EIROforum – European Intergovernmental Research Organisation
 - Profile committees organised by scientific domain
- ESFRI – European Strategy Forum for Research Infrastructure
 - Coordinates projects and funding for Research Infrastructures (RI) (2002-2010)
- eIRG – e-Infrastructure Reflection Group
 - High level policy development for Europe on e-Infrastructure
- EEF - European e-Infrastructure Forum
 - Principles and practices to create synergies for distributed Infrastructures
- CERN - High Energy Physics and LHC experiment
- TERENA
 - REFEDS – Research and Education Federations
- LIBER – Association of European libraries
 - Growing role of scientific libraries including access to research information



Existing and emerging SDI

- WLCG – Worldwide LHC Grid (CERN, Geneva)
- EGI – European Grid Infrastructure
 - Operational Grid infrastructure serving around 10,000 researches worldwide
 - Published “Seeking new horizons: EGI’s role for 2020”
- HELIX Nebula – The Science Cloud (prospective cloud based SDI)
 - Private partnership project with limited EC support
- Growing RI for different research communities
 - CLARIN, EUDAT, LifeWatch, ELIXIR, etc.
 - Less technology and more subject focused



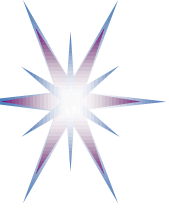
EC Strategy and Policy on STI (Scientific and Technology Information)

- Covers 2 main areas
 - Scientific Information policies defining priorities for STI in Europe and harmonise national policies
 - e-Infrastructure for scientific information
 - Management of scientific data during their whole lifecycle: includes creation, access, (re)use, and preservation
- Main targets
 - New technologies for research and scientific data use
 - Better access to support information exchange and cooperation
 - Preservation of data for future reuse and secondary research
- EC initiative on Open Access scientific publications from publicly funded projects
- G8+O5 Global Research Data Infrastructure, Subgroup on Data, Draft Report, 28 October 2011
 - Requires “reliable infrastructures for persistent identification of data (e.g. digital object identifiers, handle systems), researchers (e.g. digital authors identifiers), and authentication, authorization and accounting systems (AAA)”



Open Access to Scientific Publications

- EC initiative on Open Access scientific publications from publicly funded projects
 - Included into Declaration from Rome meeting
 - Approx 3500 publicly funded ROs and 2000 privately funded ROs
 - Special funding scheme for reimbursing publications
 - *Ongoing consultations with China, India, Russia compliance to OA principles at high governmental level*
- OpenAIRE project exploring models for open access to publications
 - PID (personal researcher identifier), ORCID (Open Researcher and Contributor ID), Linked data
- Community initiative - Panton Principles for Open Data in Science (<http://pantonprinciples.org/>)



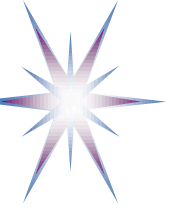
SDI and AAA related Issues for SDI/Big Data

1. Communities and their specific Big Data challenges
2. Peta and Exa scale in Computing, Storage, Communications
3. Data lifecycle management
4. Data provenance
5. Data linking to publications and PID/ORCHID
6. Trusted/trustworthy SDI and trust management
7. Data protection and privacy
8. Data usage accounting
9. Embedded access policy
10. Longterm preservation and related access control issues
11. Federated access control and data centric access control
12. SDI security, trustworthiness and Clouds
13. Future AAI/AAA for SDI and evolution
14. **SDI and AAI architecture, models, mechanisms**

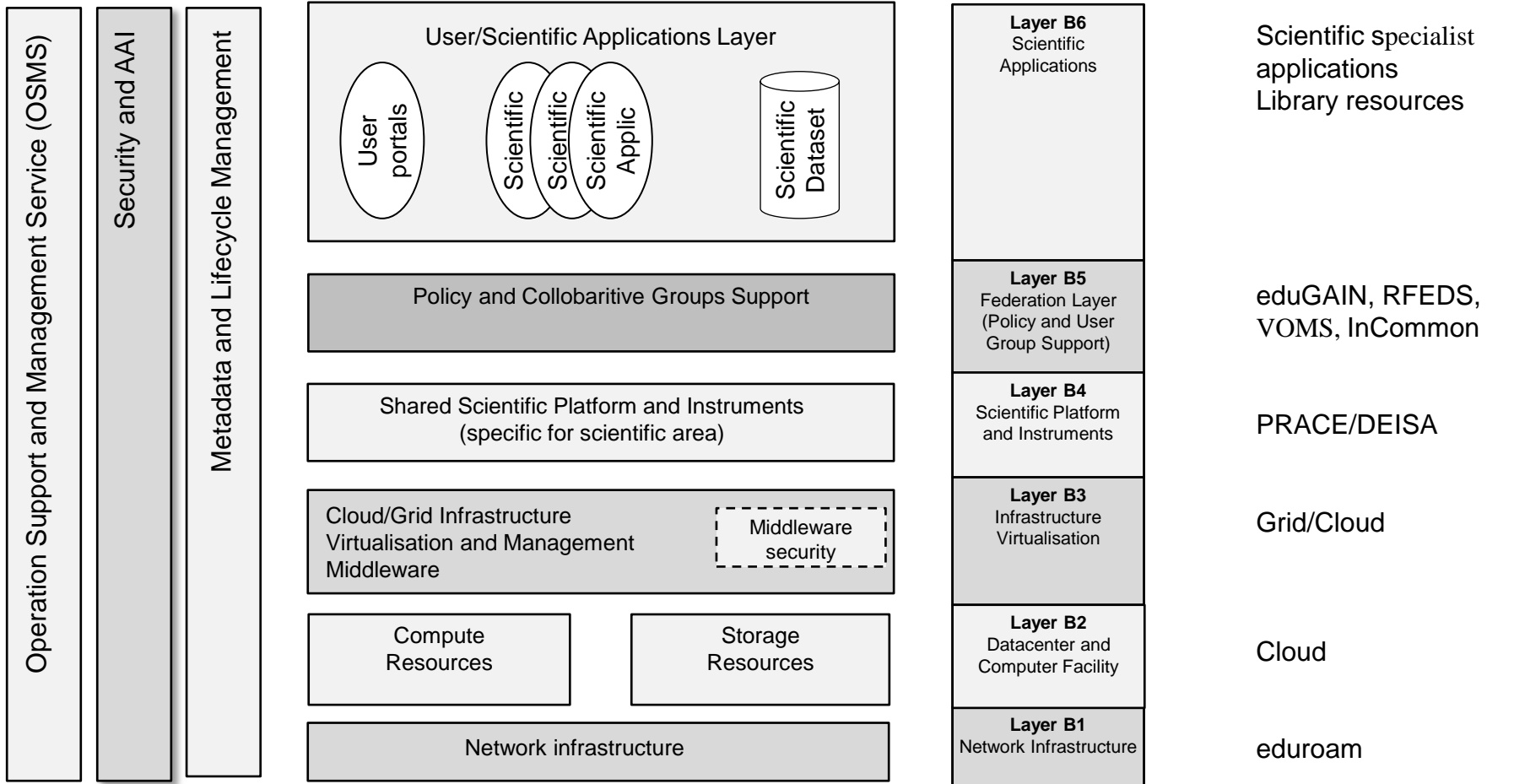


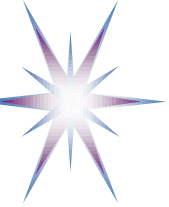
Architecture framework for SDI and ACAI/AAI

- e-SDI multi-layer architecture model
- Data lifecycle model
- RORA model to define relationship between resources and actors
 - Potentially new actor in SDI – Subject of data (e.g. patient, or scientific object/paper)
- Access Control and Accounting Infrastructure
 - Authentication, Authorisation, Accounting
 - Supported by logging service
 - Extended to support data access control and operations on data



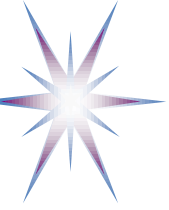
SDI Architecture Model



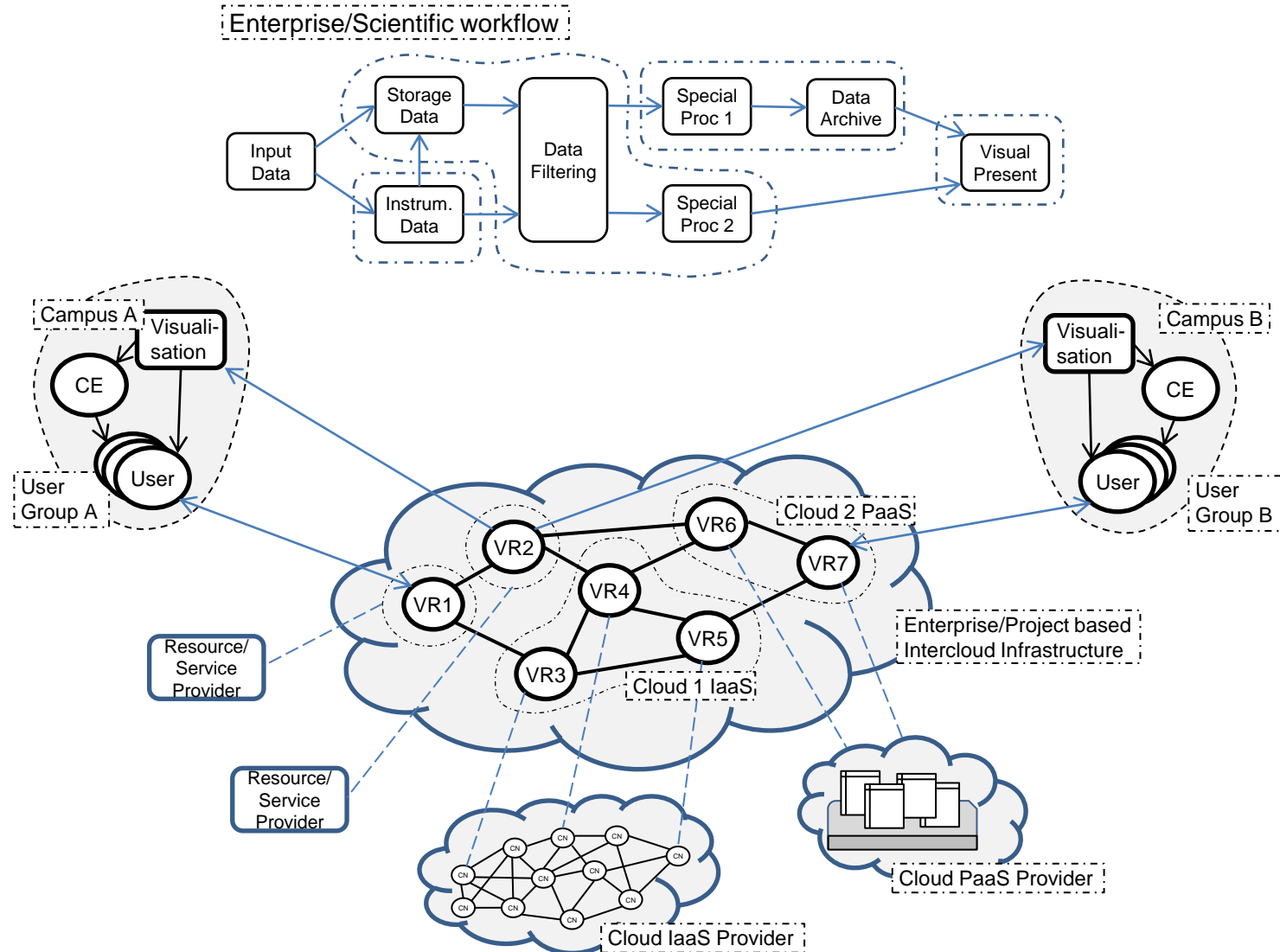


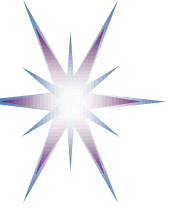
SDI move to Clouds

- Cloud technologies allow for infrastructure virtualisation and its profiling for Data structures
 - Clouds provide just right technology for infrastructure virtualisation to support data sets
- Cloud can provide infrastructure on-demand to support scientific workflows
 - Similar to Grid but with benefits of the full infrastructure provisioning on-demand
 - Including support for dynamically provisioned Virtual Organisations or associations

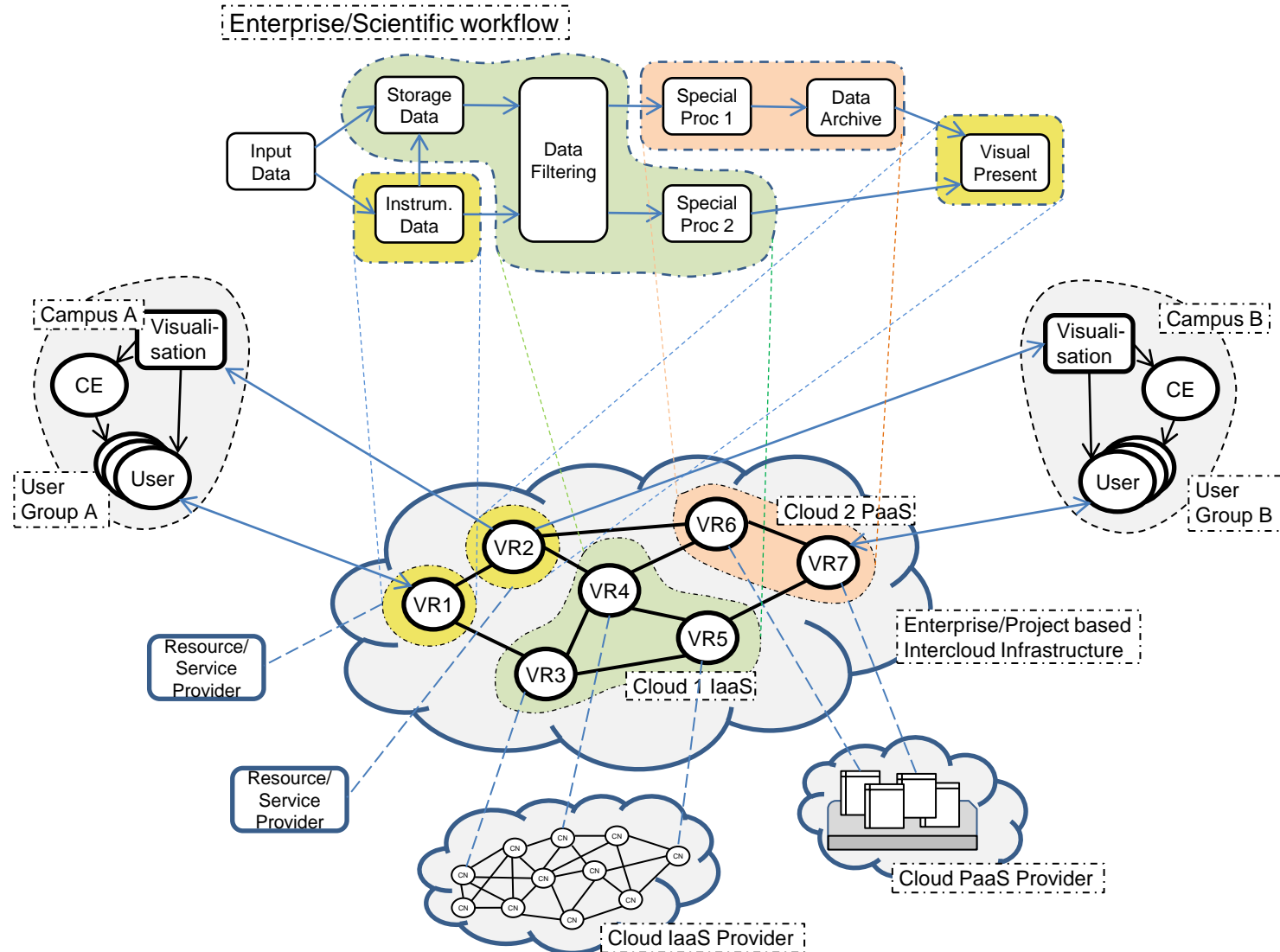


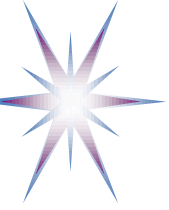
General use case for infrastructure provisioning: Workflow => Logical (Cloud) Infrastructure



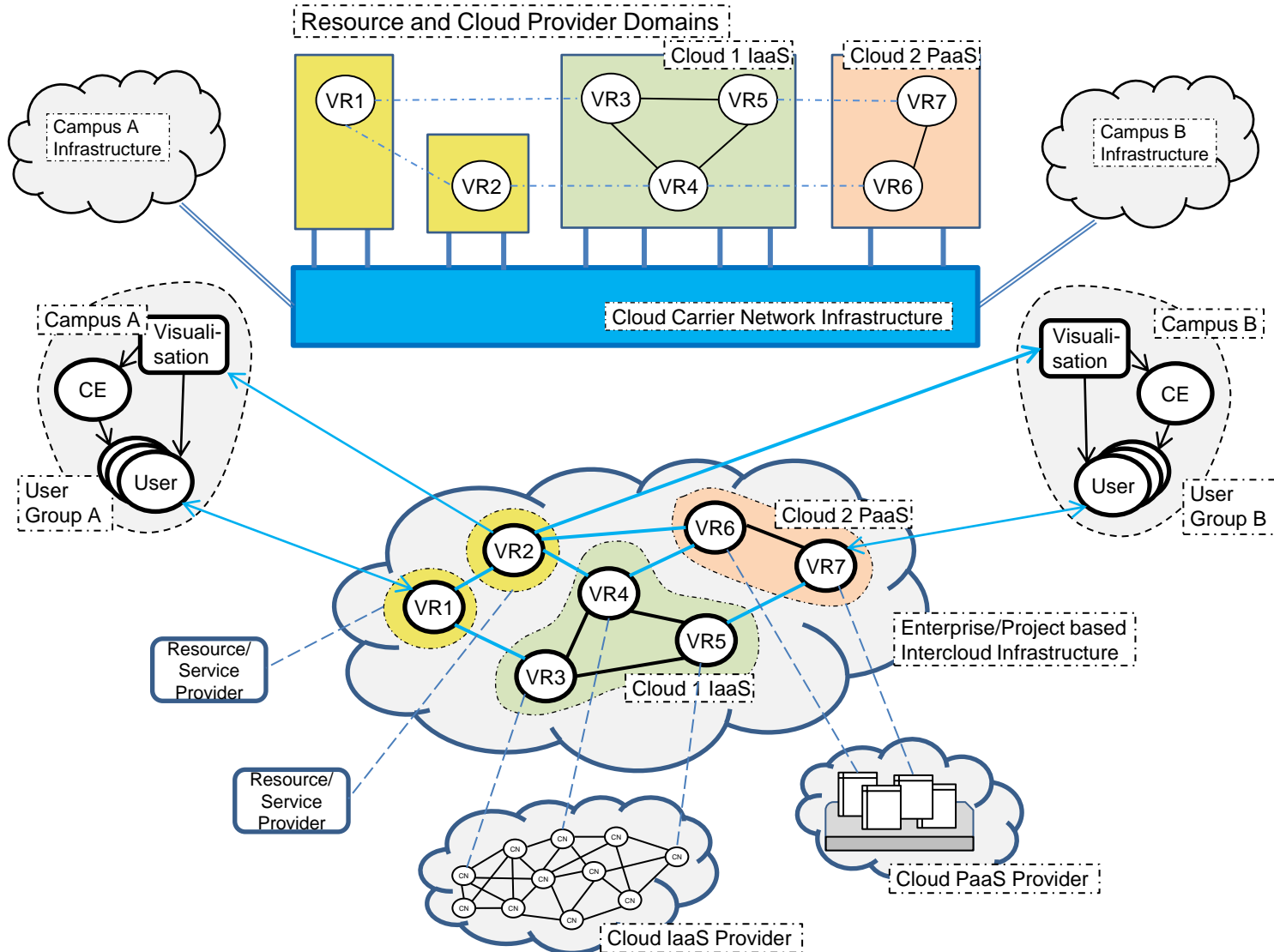


General use case for infrastructure provisioning: Workflow => Logical (Cloud) Infrastructure





General use case for infrastructure provisioning: Logical Infrastructure => Network Infrastructure (1)





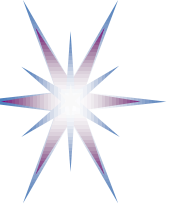
New Scientific and Academic Discipline

- New Scientific and Academic Discipline is needed for Big Data Science
 - Included into declaration of the Rome meeting
 - One of the priority item for the next FP8 program Horizon2012 (2013-2020)
- To address
 - ICT infrastructure building, operation and optimisation
 - Big data and distributed computing
 - Metadata and semantics
 - Security and trustworthiness Infrastructure and Data
- Need to educate/train new specialists in Big Data
 - Primarily for not ICT savvy community



Trustworthy SDI – Concept and Foundation

- Trust model for SDI and ICT
- Components of trustworthy ICT
- Security in virtualised environment



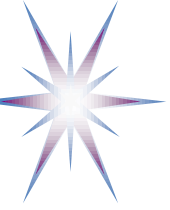
Trust model – Transposition from real world (human practice) to ICT domain

- Trust is two or multi-party relational concept
- Common trust definition and presented in [1] primarily taken from life

A trusts B to do something (typically Action, Operation, or Assertion)

- This definition is suitable for simple trust models like in PKI, authentication, authorisation or delegation
- However transposition of this definition to more complex cases as intended Trustworthy infrastructure require more factors to be included
 - Two or more parties involved into trust relations, typically Users (or user client/application) and Provider (or server or service/application)
 - Data with their ownership, privacy, confidentiality level, and applied/attached handling policy
 - Computing platform and storage (which however can associated with the general platform)
 - Underlying communication infrastructure (which however can associated with the general platform)
 - Possible interdomain issues related to different policies and assurance/provisions

User A trusts service B to do an operation O (or make Assertion) with data D on platform P



Trust model – Transposition to ICT/SDI domain

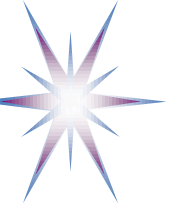
Classic Trust definition (“humanised”)

*A trusts B to do something
(typically Action, Operation, or Assertion)*

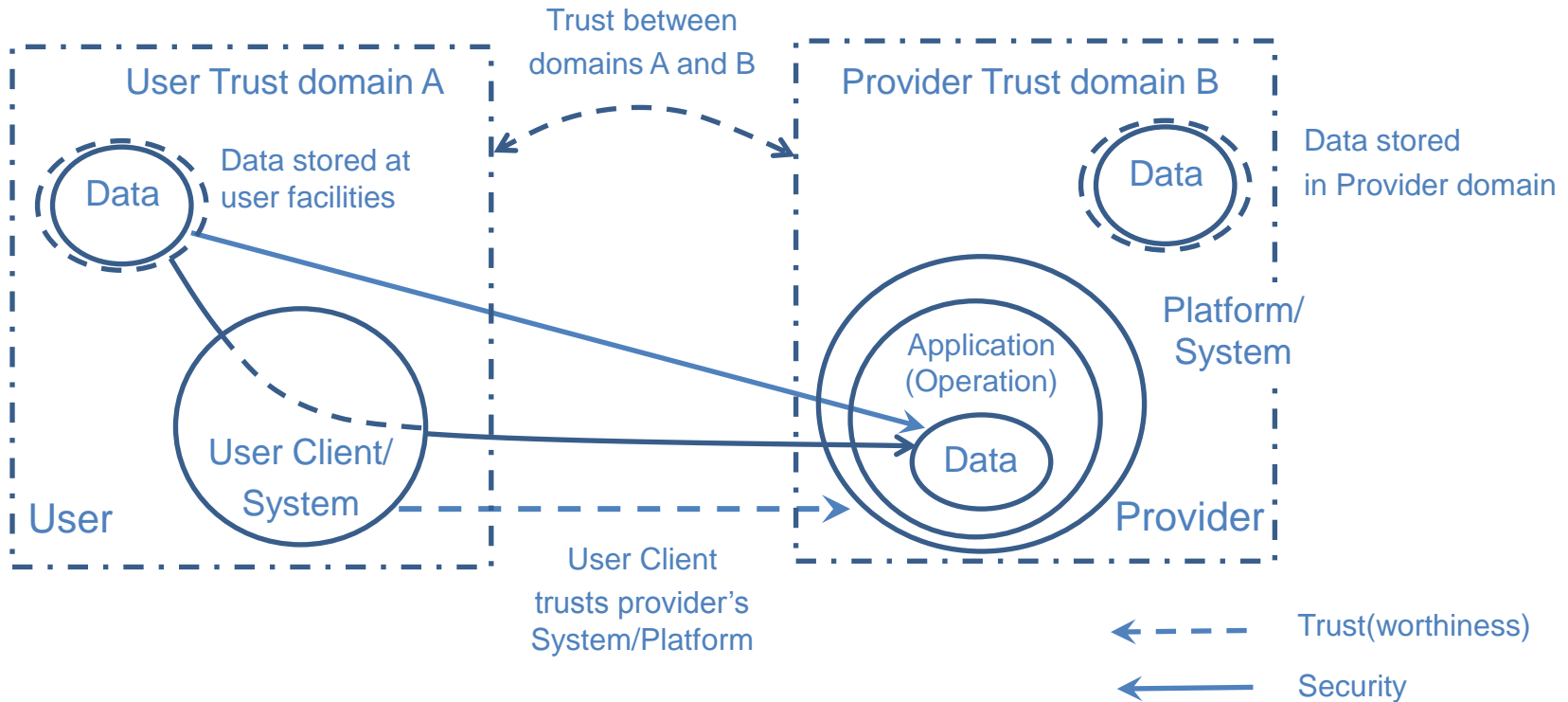
Trust definition transposed to Trustworthy ICT

*User A trusts service B to do an operation OP
with data D on platform P*

- Data are associated with the handling policy
- Platform is multi-domain and multi-provider
- Platform includes Computing, Storage and Network
- Data may be stored by 3rd party and belong to domain different to user



Component of the Trusted/Trustworthy ICT



User and Service Provider – two actors concerned with own Data/Content security and each other System/Platform trustworthiness

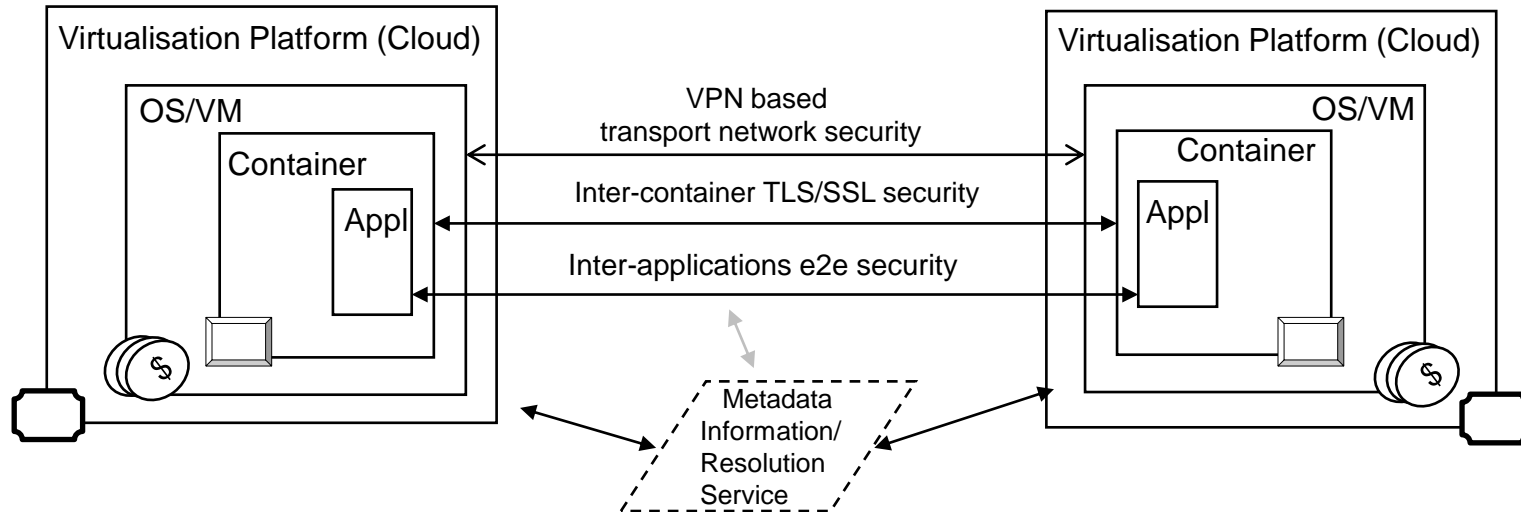
Data may be stored separately but typically belong to user trust domain

- To be extended to multidomain and distributed environment

Security in Virtualised Environment: Platform, OS/VM, Application and Network

Platform – OS/VM – Container – Application
Multilayer security

Platform – OS/VM – Container – Application
Multilayer security



Session Credentials/Context

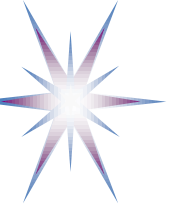


User/System Credentials, secrets



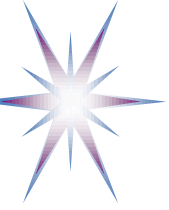
Platform trust anchor (TPM, hardware secret key)

- Explanation about multilayer security and trust
- TODO: Add Data related components: metadata, encapsulated policies
- TODO: Add policies at each layer



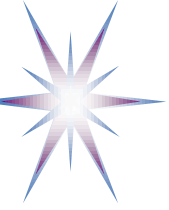
Discussion - Обсуждение

Вопросы - Questions



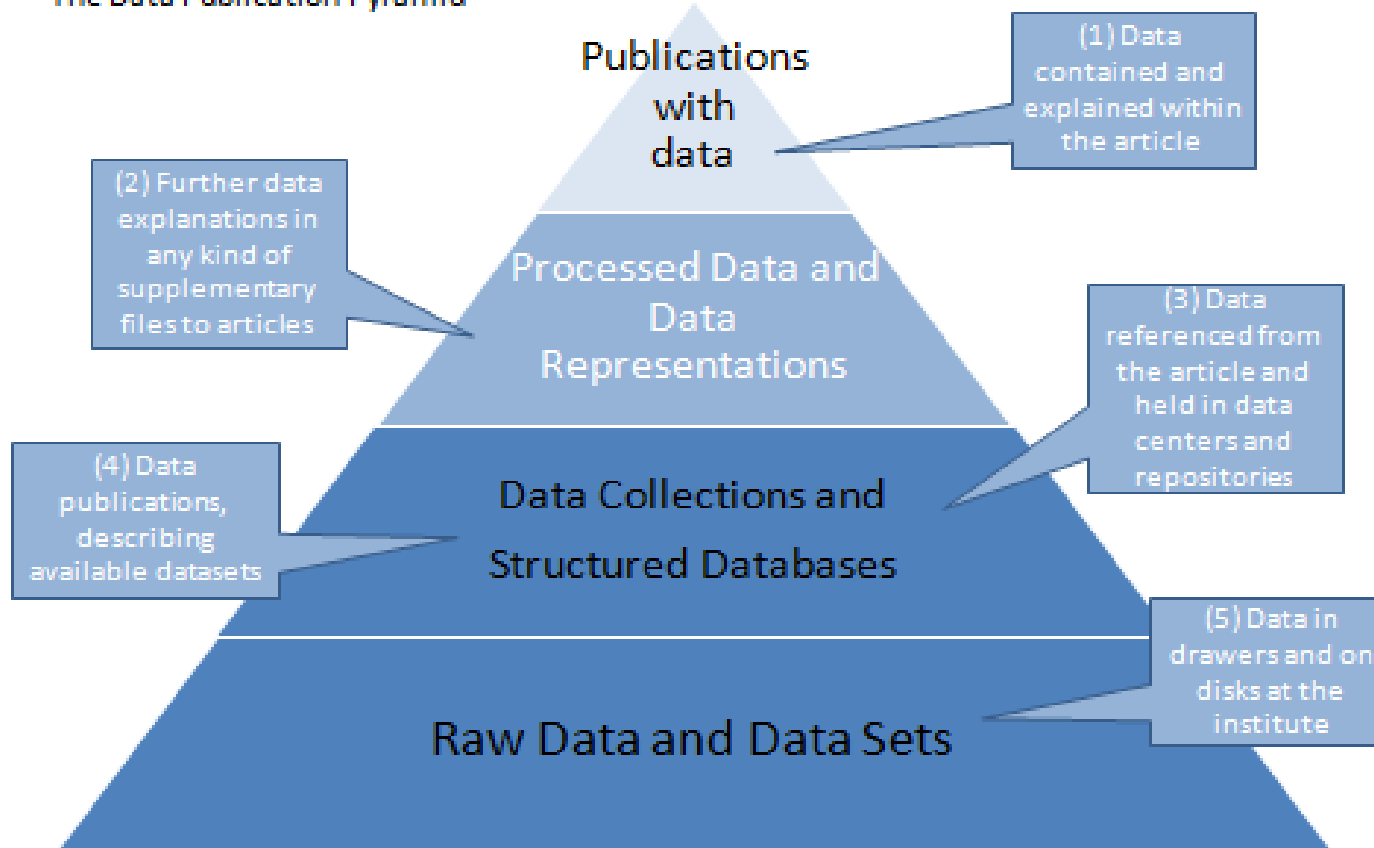
Additional and background materials

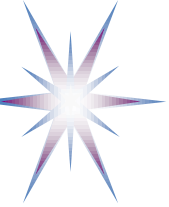
- LIBER Publications Pyramid
- DAITF model
- Distributed Computing Fallacy



Data Publication Pyramid (Credits LIBER)

The Data Publication Pyramid





Distributed Computing Fallacy

Classical definition

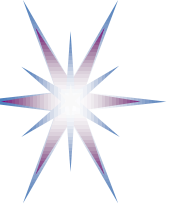
1. The network is reliable
2. Latency is zero
3. Bandwidth is infinite
4. Transport cost is zero
5. The network is secure
6. Topology doesn't change
7. There is one administrator
8. The network is homogeneous

New aspects brought by clouds and virtualisation

9. Network connection setup time is zero
- 10.

<http://www.infoq.com/news/2009/05/fallacies-distributed-computing>

http://en.wikipedia.org/wiki/Fallacies_of_Distributed_Computing



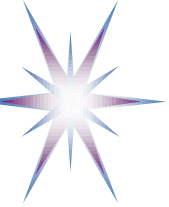
AAI/AAA for SDI and Big Data

- AAI for SDI requirement
- Traditional AAI
- AAI architecture model for AAI

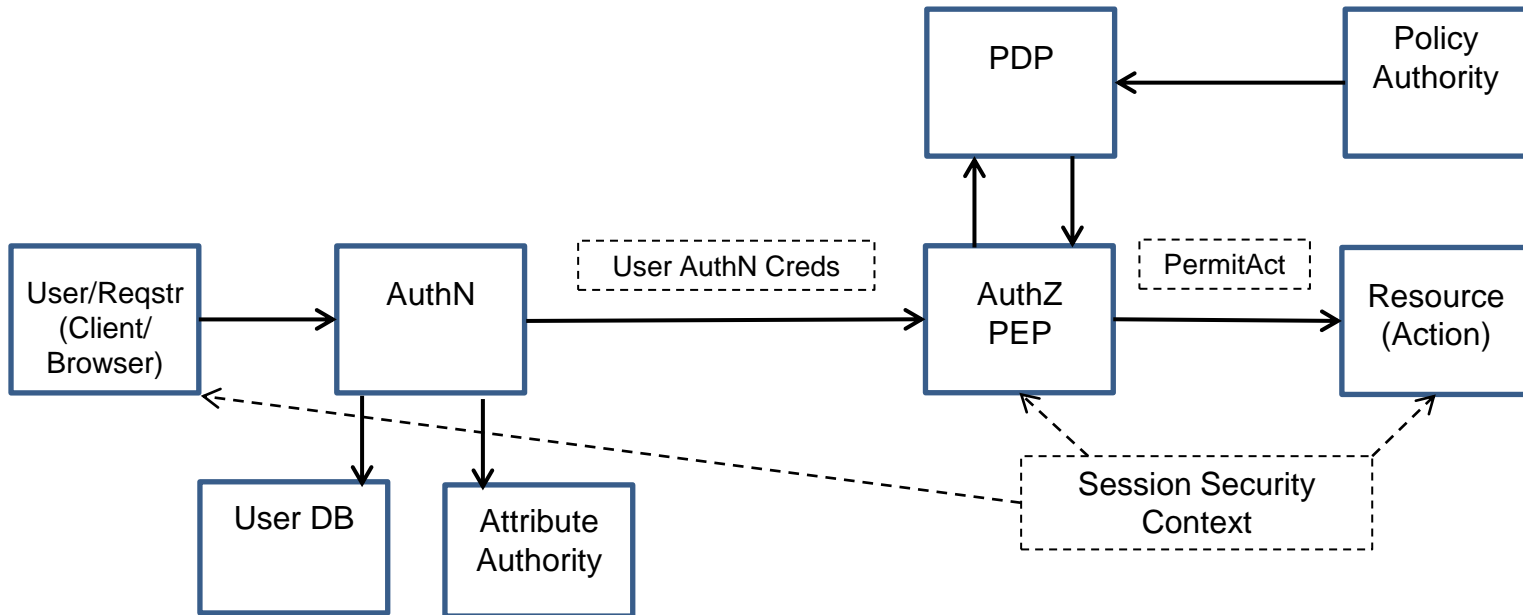


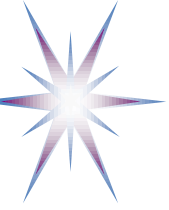
AAI Architecture model for e-SDI - Requirements

- Federated access and identity management
 - Different access devices and media
- Multi-level access control to distributed information resources
 - Distributed inter-linked data and supporting services/information
- Controllable access to datasets and data
 - Data lifecycle stages
 - Long term preservation and linking policy to data
 - Linking papers and data
- Capable to protect data at Peta- and Exascale
- Data use accounting
 - At least Who/Subject accessed and What/Action did?
- Trusted platform and virtualised/Cloud environment

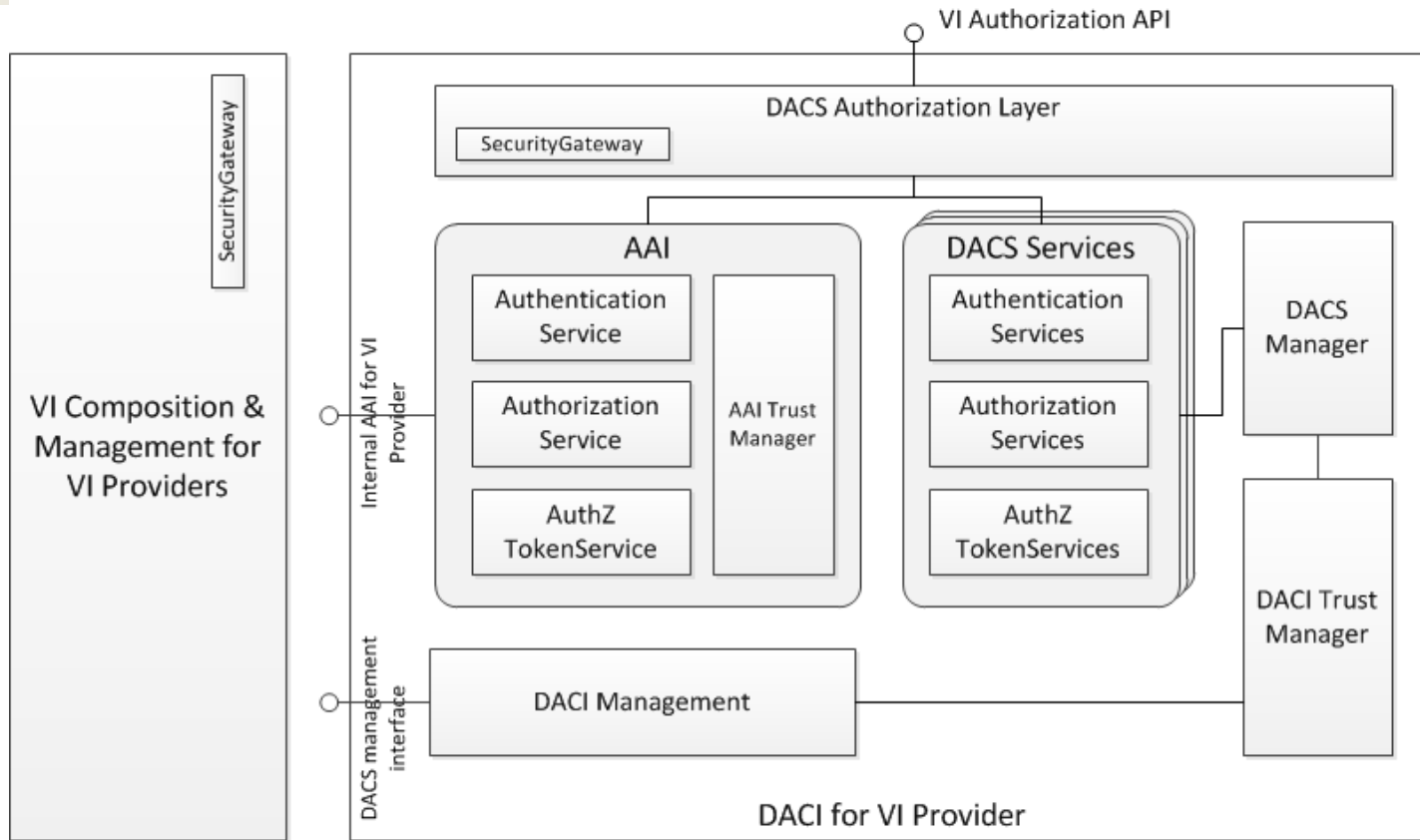


AAI Architecture Model - Traditional

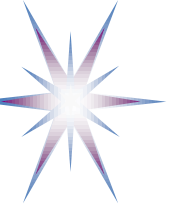




AAI Architecture Model for Data Centric SDI

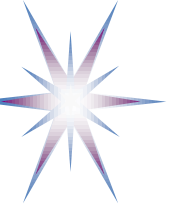


- AAI for e-SDI incorporates a number of functionalities to support infrastructure virtualisation and on-demand provisioning
 - Addressed by DACI/DACS and Trust management components



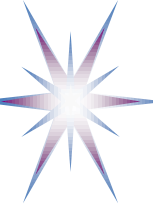
IETF standardisation on AAI

- IETF abfab-wg (Application Bridging for Federated Access Beyond web)
 - Initiative and contribution from Moonshot Project (JANET and GEANT3)
 - Extends web-based federated access (authentication and authorisation) to inter-applications and inter-layers
- OAuth2.0
 - Initially proposed and used by Facebook, now community contribution
- OpenID - Enables an End-user to communicate with a Relying party (RP) to verify the end-user's identifier
 - Widely supported by NREN and research community
- IETF kitten-wg – Kerberos and SASL update and extension



OGF Standardisation on AAI in Grids and Clouds

- Grid AAI related standards
 - Basic Authentication profile
 - GIN-WG (Grid Interoperability Now) AAI related standards
 - Grid Authorisation and VOMS X.509 Attribute Certificate
 - XACML-Grid profile (draft version)
- Cloud interoperability and federation
 - Federated access control to Cloud resources



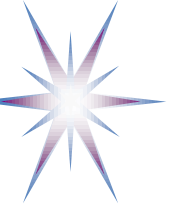
User Access and Data Security in Clouds

- Security challenges in clouds
- Cloud environment and problems to be addressed
- Emerging cloud Security models



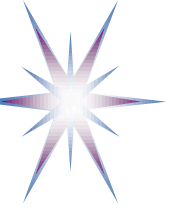
Cloud Computing Security – Challenges

- **Fundamental security challenges and main user concerns in Clouds**
 - Data security: Where are my data? Are they protected? What control has Cloud provider over data security and location?
 - Identity management and access control: Who has access to my data?
- Two main tasks in making Cloud secure and trustworthy
 - Secure operation of Cloud (provider) infrastructure
 - Provide user controlled access control (security) infrastructure
 - Provide sufficient amount of security controls for user
- Cloud security infrastructure should provide a framework for dynamically provisioned Cloud security services and infrastructure



Cloud Environment and Problems to be addressed

- Virtualised services
- On-demand/dynamic provisioning
- Multi-tenant/multi-user
- Multi-domain
- Uncontrolled execution and data storage environment
 - Data protection
 - Trusted Computing Platform Architecture (TCPA)
 - Promising homomorphic/elastic encryption (to be researched)
- *Integration with customer legacy security services/infrastructure*
- *Integration with the providers business workflow*



Emerging Cloud Security Models

- Former (legacy): Provider - User/Customer
- New Cloud oriented security provisioning models
 - Provider - Customer - User
 - Enterprise as a Customer, and employees as Users
 - Enterprise/campus infrastructure and legacy services
 - Provider – Operator (Broker) - Customer – User
 - Application area IT/telecom company serves as an Operator for application services infrastructure created for customer company
- Security issues/problems in new security provisioning models
 - Integration of the customer and provider security services
 - Identity Management and Single Sign On (SSO)
 - Identity provisioning for dynamically created Cloud based infrastructure or applications