Cloud based Infrastructure for Data Intensive e-Science Applications: Requirements and Architecture

Yuri Demchenko, Canh Ngo, Paola Grosso, Cees de Laat
University of Amsterdam
Peter Membrey
Hong Kong Technical University, Hong Kong

ABSTRACT

This chapter discusses the challenges that are imposed by Big Data on the modern and future e-Scientific Data Infrastructure. The chapter discusses a nature and definition of Big Data that include such characteristics as Volume, Velocity, Variety, Value and Veracity. The chapter refers to different scientific communities to define requirements on data management, access control and security. The chapter introduces the Scientific Data Lifecycle Management (SDLM) model that includes all the major stages and reflects specifics in data management in modern e-Science. The chapter proposes the generic SDI architecture model that provides a basis for building interoperable data or project centric SDI using modern technologies and best practices.

The chapter discusses how the proposed models SDLM and SDI can be naturally implemented using modern cloud based infrastructure services, analyses security and trust issues in cloud based infrastructure and summaries requirements to access control and access control infrastructure that should allow secure and trusted operation and use of SDI.

1. Introduction

The emergence of Data Intensive Science is a result of modern science computerization and increasing range of observations, experimental data collected from specialist scientific instruments, sensors, simulation in every field of science. Modern science requires wide and cross-border researchers collaboration. e-Science Data Infrastructure (SDI) need to provide an environment capable both to deal with the ever increasing heterogeneous data production and to provide trusted collaborative environment for distributed groups of researcher and scientists. Additionally SDI needs to provide access to existing scientific information including libraries, journals, datasets and specialist scientific databases, on one hand, and provide linking between experimental data and publications, on the other hand.
Industry is also experiencing wide and deep technologies re-factoring to become data intensive and data powered. Cross-fertilisation between emerging data intensive/driven e-Science and industry will bring new data intensive technologies that will drive new data intensive/powered applications.
Further successful technology development will require the definition of the Scientific Data Infrastructure (SDI) and overall architecture framework of the Data Intensive Science. This will provide a common vocabulary and allow concise technology evaluation and planning for specific application and collaborative projects or groups.
Big Data technologies are becoming a current focus and a new "buzz-word" both in science and in industry. Emergence of Big Data or data centric technologies indicates the beginning of a new form of the continuous technology advancement that is characterized by overlapping technology waves related to different aspects of the human activity from production and consumption to collaboration and general social activity. In this context data intensive science plays key role.

Big Data are becoming related to almost all aspects of human activity from just recording events to research, design, production and digital services or products delivery, to the final consumer. Current technologies such as Cloud Computing and ubiquitous network connectivity provide a platform for automation of all processes in data collection, storing, processing and visualization.

Modern e-Science infrastructures allow targeting new large-scale problems whose solution was not possible before, *e.g.* genome, climate, global warming. e-Science typically produces a huge amount of data that need to be supported by a new type of e-Infrastructure capable to store, distribute, process, preserve, and curate these data [1, 2]: we refer to this new infrastructures as Scientific Data e-Infrastructure (SDI).

In e-Science, the scientific data are complex multifaceted objects with the complex internal relations, they are becoming an infrastructure of their own and need to be supported by corresponding physical or logical infrastructures to store, access, process, visualise and manage these data.
The emerging SDI should allow different groups of researchers to work on the same data sets, build their own (virtual) research and collaborative environments, safely store intermediate results, and later share the discovered results. New data provenance, security and access control mechanisms and tools should allow researchers to link their scientific results with the initial data (sets) and intermediate data to allow future re-use/re-purpose of data, *e.g.* with the improved research technique and tools.

This chapter analyses new challenges imposed to modern e-Science infrastructures by the emerging Big Data technologies; it proposes a general approach and architecture solutions that constitute a new Scientific Data Lifecycle Management (SDLM) model and the generic

SDI architecture model that provides a basis for heterogeneous SDI components interoperability and integration, in particular based on cloud infrastructure technologies.

The chapter is primarily focused on SDI, however provides analysis of the Big Data nature in both e-Science, industry and other domains, analyses their commonalities and difference, discussing also possible cross-fertilisation between two domains.
The chapter refers to the ongoing research on defining the Big Data infrastructure for e-Science initially presented in the papers [3, 4] and significantly extends it with new results and wider scope to investigate relations between Big Data technologies in e-Science and industry. With long tradition of working with constantly increasing volume of data, modern science can offer industry the scientific analysis methods, while industry can bring Big Data technologies and tools to wider public.
The chapter is organised as follows. Section 2 looks into Big Data definition and Big Data nature in science, industry, business, and social networks analysing also the main drivers for the Big Data technology development. Section 3 gives an overview of the main research communities and summarizes requirement to future SDI. Section 4 discusses challenges to data management in Big Data Science, including SDLM discussion. Section 5 introduces the proposed e-SDI architecture model that is intended to answer the future big data challenges and requirements. Section 6 discusses SDI implementation using cloud technologies. Section 6 discusses security and trust related issues in handling data and summarises specific requirements to access control infrastructure for modern and future SDI.

2. Big Data Definition

2.1. Big Data in e-Science, Industry and other domains

Science has been traditionally dealing with challenges to handle large volume of data in complex scientific research experiments. Scientific research typically includes collection of data in passive observation or active experiments which aim to verify one or another scientific hypothesis. Scientific research and discovery methods typically are based on the initial hypothesis and a model which can be refined based on the collected data. The refined model may lead to a new more advanced and precise experiment and/or the previous data re-evaluation. Another distinctive feature of the modern scientific research is that it suggests wide cooperation between researchers to challenge complex problems and run complex scientific instruments.
In industry, private companies will not share data or expertise. When dealing with data, companies will intend always keep control over their information assets. They may use shared third party facilities, like clouds, but special measures need to be taken to ensure data protection, including data sanitization. It might be also a case that companies can use shared facilities only for proof of concept and do production data processing at private facilities. In this respect, we need to accept that science and industry can't be done in the same way, and consequently this will be reflected in a way how they can interact and how the Big Data infrastructure and tools can be built.
With the digital technologies proliferation into all aspects of business activities and emerging Big Data technologies, the industry is entering a new playground when it needs to use scientific methods to benefit from the possibility to collect and mine data for desirable information, such as market prediction, customer behavior predictions, social groups activity predictions, etc.
A number of discussions and blog articles [5, 6, 7] suggest that the Big Data technologies need to adopt scientific discovery methods that include iterative model improvement and

collection of improved data, re-use of collected data with improved model.
We can quote here a blog article by Mike Gualtieri from Forrester [8]:"Firms increasingly realize that [big data] must use predictive and descriptive analytics to find nonobvious information to discover value in the data. Advanced analytics uses advanced statistical, data mining and machine learning algorithms to dig deeper to find patterns that you can't see using traditional BI (*Business Intelligence*) tools, simple queries, or rules."

2.2. The Big Data Definition

Despite the "Big Data" has become a new buzz-word, there is no consistent definition for Big Data, nor detailed analysis of this new emerging technology. Most discussions until now have been going in blogosphere, where however the most significant Big Data characteristics have been identified and became commonly accepted [8, 9.10]. In this section we will attempt to summarise available definitions and propose a consolidated view on the generic Big Data features that would help us to define requirements to supporting Big Data infrastructure and in particular Scientific Data Infrastructure.
As a starting point, we can refer to the simple definition given in [9]: "Big Data: a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques."
Related definition of the data-intensive science is given in the book "The Fourth Paradigm: Data-Intensive Scientific Discovery" by the computer scientist Jim Gray [10]: "The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science as a new, fourth paradigm for scientific exploration."

2.3. 5 Vs of Big Data

In a number of discussions and articles Big Data are attributed to have such native generic characteristics as Volume, Velocity, and Variety, also referred to as "3 Vs of Big Data". After being stored and entered into the processing stages or workflow, Big Data acquire new properties such as Value and Veracity which together constitute the Big Data as 5 Vs: Volume, Velocity, Variety, Value, and Veracity [4].
Figure 1 below illustrates the features related to 5 Vs which we analyse below.

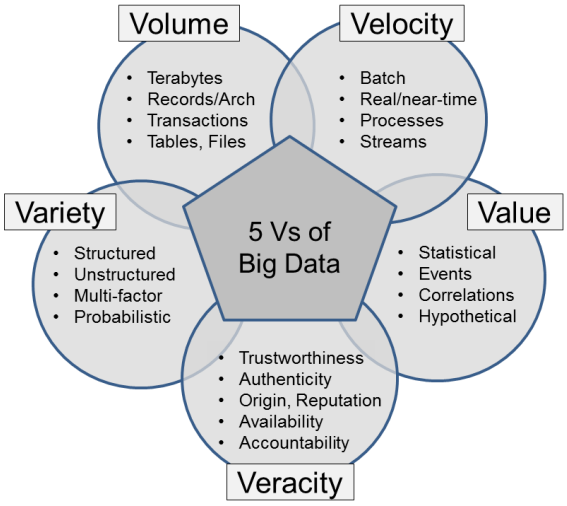


Figure 1. 5 Vs of Big Data

Volume
Volume is the most important and distinctive feature of Big Data which impose additional

and specific requirements to all traditional technologies and tools currently used.
In e-Science, growth of data amount is caused by advancements in both scientific instruments and SDI. In many areas the trend is actually to include data collections from all observed events, activities and sensors what became possible and is important for social activities and social sciences.
Big Data volume includes such features as size, scale, amount, dimension for tera- and exascale data recording either data rich processes, or collected from many transactions and stored in individual files or databases – all needs to be accessible, searchable, processed and manageable.
Two examples from e-Science give also different characters of data and also different processing requirements, such as:

- Large Hadron Collider (LHC) [11, 12] produces in average 5 PB data a month that are generated in a number of short collisions that make them unique events, The collected data are filtered, stored and extensively searched for single events that may confirm a scientific hypothesis.
- LOFAR (Low Frequency Array) [13] is a radio telescope that collects about 5 PB every hour, however the data are processed by correlator and only correlated data are stored.

In industry, global services providers such as Google [14], Facebook [15], Twitter [16] are producing, analyzing and storing data in huge amount as their regular activity/production services. Although some of their tools and processes are proprietary, they actually prove the feasibility of solving Big Data problems at the global scale and significantly push the development of the Open Source Big Data tools.

Velocity
Big Data are often generated at high speed, including also data generated by arrays of sensors or multiple events, and need to be processed in real-time, near real-time or in batch, or as streams (like in case of visualisation).
As an example, LHC ATLAS detector [12] uses about 80 readout channels and collects up to 1PB of unfiltered data in second which are reduced to approx. 100MB per second. This should record up to 40 million collision events per second.
Industry can also provide numerous examples when data registration, processing or visualization impose similar challenges.

Variety
Variety deals with the complexity of big data and information and semantic models behind these data. This is resulted in data collected as structured, unstructured, semi-structured, and a mixed data. Data variety imposes new requirements to data storage and database design which should dynamic adaptation to the data format, in particular scaling up and down.
Biodiversity research [17] provides a good example of the data variety due to the fact that biodiversity involves collecting and processing information from wide range of sources and collected information related to species population, genomic data, climate, satellite information, etc. Another example can be urban environment monitoring (also called “smart cities” [18]) that requires operating, monitoring and evolving of numerous processes, individuals and associations.
Adopting data technologies in traditionally non-computer oriented areas such as psychology and behavior research, history, archeology will generate especially rich data sets.

Value
Value is an important feature of the data which is defined by the added-value that the collected data can bring to the intended process, activity or predictive analysis/hypothesis. Data value will depend on the events or processes they represent such as stochastic, probabilistic, regular or random. Depending on this the requirements may be imposed to

collect all data, store for longer period (for some possible event of interest), etc. In this respect data value is closely related to the data volume and variety. The stock exchange financial data provides a good example of high volume data which have high value for real time market trends monitoring, however decreasing value with time and depending on the market volatility [19].

Veracity
Veracity dimension of Big Data includes two aspects: data consistency (or certainty) what can be defined by their statistical reliability; and data trustworthiness that is defined by a number of factors including data origin, collection and processing methods, including trusted infrastructure and facility.
Big Data veracity ensures that the data used are trusted, authentic and protected from unauthorised access and modification. The data must be secured during the whole their lifecycle from collection from trusted sources to processing on trusted compute facilities and storage on protected and trusted storage facilities.
The following aspects define and need to be addressed to ensure data veracity:

- Integrity of data and linked data (e.g., for complex hierarchical data, distributed data)
- Data authenticity and (trusted) origin
- Identification of both data and source
- Computer and storage platform trustworthiness
- Availability and timeliness
- Accountability and Reputation

Data veracity relies entirely on the security infrastructure deployed and available from the Big Data infrastructure [20].

3. Research Infrastructures and Infrastructure requirements

This section will refer and provide short overview of different scientific communities, in particular as defined by the European Research Area (ERA) [21], to define requirements on infrastructure facility, data processing and management functionalities, user management, access control and security.

3.1. Paradigm change in modern e-Science

Modern e-Science is moving to the Data Intensive technologies that are becoming a new technology driver and require re-thinking a number of infrastructure architecture and operational models, components, solutions and processes to address the following general challenges [2, 4]:

- Exponential growth of data volume produced by different research instruments and/or collected from sensors
- Need to consolidate e-Infrastructures as persistent research platforms to ensure research continuity and cross-disciplinary collaboration, deliver/offer persistent services, with adequate governance model.

The recent advancements in the general computer and big data technologies facilitate the paradigm change in modern e-Science that is characterized by the following 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 allow 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

The future SDI should support the whole data lifecycle and explore the benefit of the data storage/preservation, aggregation and provenance in a large scale and during long/unlimited period of time. Important is that this infrastructure must ensure data security (integrity, confidentiality, availability, and accountability), and data ownership protection. With current needs to process big data that require powerful computation, there should be a possibility to enforce data/dataset policy that they can be processed on trusted systems and/or complying other requirements. Researchers must trust the SDI to process their data on SDI facilities and be ensured that their stored research data are protected from non-authorised access. Privacy issues are also arising from distributed remote character of SDI that can span multiple countries with different local policies. This should be provided by the corresponding Access Control and Accounting Infrastructure (ACAI) which is an important component of SDI [20, 22].

3.2. Research communities and specific SDI requirements

A short overview of some research infrastructures and communities, in particular the ones defined for the Europe Research Area (ERA) [21] allows us to better understand specific requirements for the future SDIs that is capable to address Big Data challenges.
Existing studies of European e-Infrastructures analyze the scientific communities practices and requirements; examples are those undertaken by the SIENA Project [23], EIROforum Federated Identity Management Workshop [24], European Grid Infrastructure (EGI) Strategy Report [25], UK Future Internet Strategy Group Report [26].

The **High Energy Physics** community represents a large number or researchers, unique expensive instruments, huge amount of data that are generated and need to be processed continuously. This community has already the operational Worldwide Large Hadron Collider Grid (WLCG) [11] infrastructure to manage and access data, protect their integrity and support the whole scientific data lifecycle. WLCG development was an important step in the evolution of European e-Infrastructures that currently serves multiple scientific communities in Europe and internationally. The EGI cooperation [27] manages European and worldwide infrastructure for HEP and other communities.

Material science, analytical and low energy physics (proton, neutron, laser facilities) is characterized by short projects, experiments and consequently highly dynamic user community. It requires highly dynamic supporting infrastructure and advanced data management infrastructure to allow wide data access and distributed processing.

Environmental and Earth science community and projects target regional/national and

global problems. They collect huge amount of data from land, sea, air and space and require ever increasing amount of storage and computing power. This SDI requires reliable fine-grained access control to huge data sets, enforcement of regional issues, policy based data filtering (data may contain national security related information), while tracking data use and keeping data integrity.

Biological and Medical Sciences (also defined as Life sciences) have a general focus on health, drug development, new species identification, new instruments development. They generates massive amount of data and new demand for computing power, storage capacity, and network performance for distributed processes, data sharing and collaboration. Biomedical data (healthcare, clinical case data) are privacy sensitive data and must be handled according to the European policy on Personal Data processing [27].

Biodiversity research [17] involves research data and research specialists at least from biology, environmental research and may include data about climate, weather and satellite observation. This primarily present challenges for integrating different sources of information with different data models and processing huge amount of collected information but may also require fast data processing in case of natural disasters. The projects LifeWatch [28] and ENVRI [29] present good example of what research approaches and what kind of data are used here.

Social Science and Humanities communities and projects are characterized by multi-lateral and often global collaborations between researchers from all over the world that need to be engaged into collaborative groups/communities and supported by collaborative infrastructure to share data, discovery/research results and cooperatively evaluate results. The current trend to digitize all currently collected physical artifacts will create in the near future a huge amount of data that must be widely and openly accessible.

3.3. General SDI Requirements

From the overview we just gave we can extract the following general infrastructure requirements to SDI for emerging Big Data Science:

- Support for long running experiments and large volume of heterogeneous data generated at high speed
- On-demand infrastructure provisioning to support data sets and scientific workflows, mobility of data-centric scientific applications
- Provide High Performance Computing facilities to allow complex data analytics with evolving research models
- Support distributed and mobile sensor networks for observation data collection and advance information visualisation
- Support of virtual scientists communities, addressing dynamic user groups creation and management, federated identity management
- Support the whole data lifecycle management, in particular, advanced data provenance, data archiving and consistent data identification
- Multi-tier inter-linked data distribution and replication
- Trusted environment for data storage and processing
- Support for data integrity, confidentiality, accountability
- Policy binding to data to protect privacy, confidentiality and IPR

4. Scientific Data Management

4.1. Scientific Information and Data in modern e-Science

Emergence of computer aided research methods is transforming the way research is done and scientific data are used. The following types of scientific data are defined as illustrates in a form of Scientific Data pyramid (see Figure 2) [22]:

- Raw data collected from observations and from experiments (what actually is done according to an initial research model or hypothesis)
- Structured data and datasets that went through data filtering and processing (supporting some particular formal model which is typically refined from the initial model). These data are already stored in repositories and may be shared to collaborative groups of researchers.
- Published data that supports one or another scientific hypothesis, research result or statement. These data are typically linked to scientific publications as supplemental materials, they may be located on the publisher's platform or authors' institution platform and have open or licensed access.
- Data linked and embedded into publications to support the wide research consolidation, integration, and openness.

Once the data are published, it is essential to allow other scientists to be able to validate and reproduce the data that they are interested in, and possibly contribute with new results. Capturing information about the processes involved in transformation from raw data up until the generation of published data becomes an important aspect of scientific data management. Scientific data provenance becomes an issue that also needs to be taken into consideration by SDI providers [30].

Another aspect to take into consideration is to guarantee reusability of published data within the scientific community. Understanding semantic of the published data becomes an important issue to allow for reusability, and this had been traditionally been done manually. However, as we anticipate unprecedented scale of published data that will be generated in Big Data Science, attaching clear data semantic becomes a necessary condition for efficient reuse of published data. Learning from best practices in semantic web community on how to provide a reusable published data, will be one of consideration that will be addressed by SDI.

Big data are typically distributed both on the collection side and on the processing/access side: data need to be collected (sometimes in a time sensitive way or with other environmental attributes), distributed and/or replicated. Linking distributed data is one of the problems to be addressed by SDI.

The European Commission's initiative to support Open Access to scientific data from publicly funded projects suggests introduction of the following mechanisms to allow linking publications and data [31]:

- PID - persistent data ID [32]
- ORCID – Open Researcher and Contributor Identifier [23].

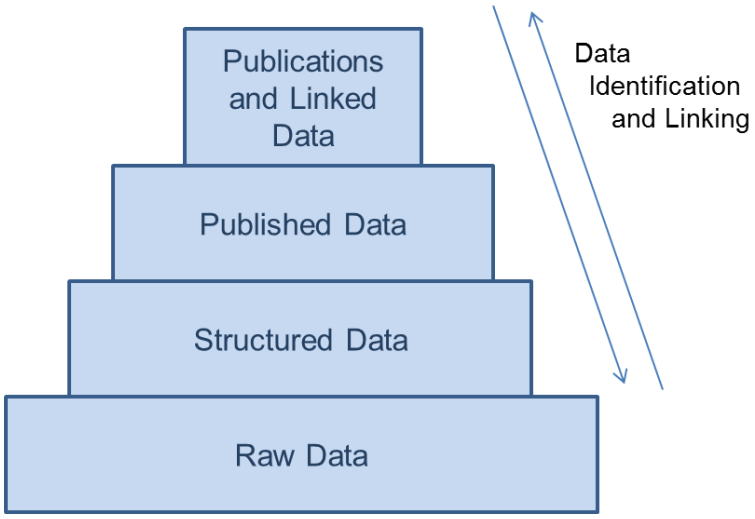


Figure 2: Scientific Data Pyramid

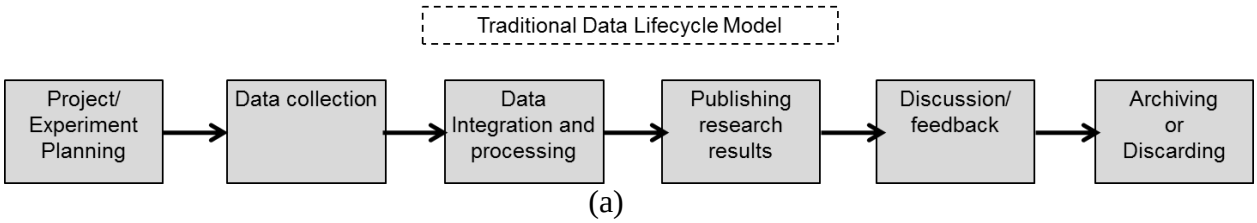
4.2. Data Lifecycle Management in Scientific Research

Computer and IT enabled e-Science allows multipurpose data collection and use and advanced data processing. A possibility to store the initial data sets and all intermediate results will allow for future data use, in particular data re-purposing and secondary research as the technology and scientific methods develops.

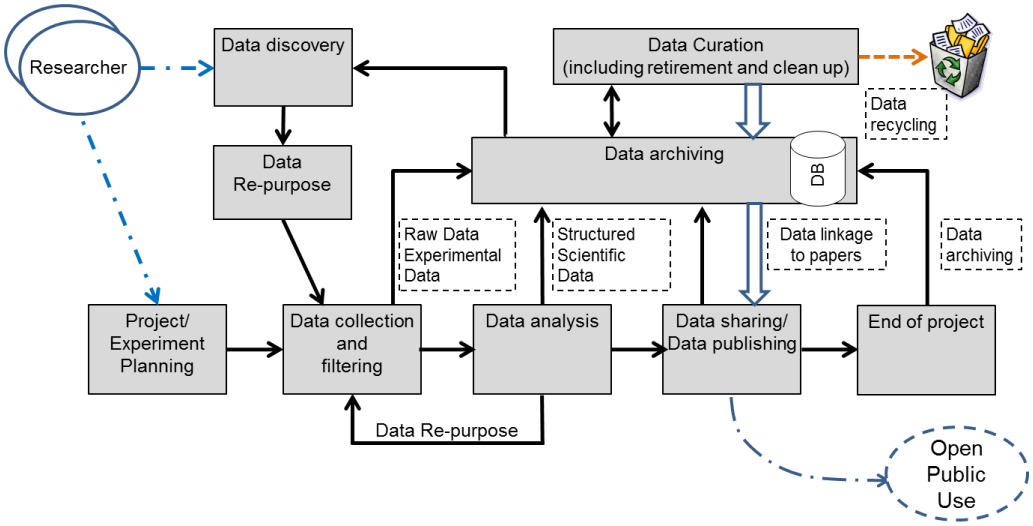
Emergence of computer aided research methods is transforming the way how research are done and scientific data are processed/used. This is also reflected in the changed Scientific Data Lifecycle Management model as shown in Figure 3 and discussed below.

We refer to the extensive study on the SDLM models in [34]. The traditional scientific data lifecycle includes a number of consequent stages (see Fig. 3,a):

- Research project or experiment planning
- Data collection
- Data integration and processing
- Publishing research results
- Discussion, feedback
- Archiving (or discarding)



(a)



(b)

Figure 3. Scientific Data Lifecycle Management in e-Science

The new SDLM model requires data storage and preservation at all stages what should allow data re-use/re-purposing and secondary research on the processed data and published results. However, this is possible only if the full data identification, cross-reference and linkage are implemented in SDI. Data integrity, access control and accountability must be supported during the whole data during lifecycle. Data curation is an important component of the discussed SDLM and must also be done in a secure and trustworthy way.

Support data security and access control to scientific data during their lifecycle: data acquisition (experimental data), initial data filtering, specialist processing; research data storage and secondary data mining, data and research information archiving.

5. Scientific Data Infrastructure Architecture Model

The proposed generic SDI architecture model provides a basis for building interoperable data or project centric SDI using modern technologies and best practices. Figure 4 shows the multilayer SDI architecture for e-Science (e-SDI) that contains the following layers:
Layer D1: Network infrastructure layer represented by either the general purpose Internet infrastructure or 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 and Delivery Layer that represent the general Federated Access and Delivery (FADI) that includes infrastructure components for interconnecting, integrating and operating complex scientific infrastructure to support project oriented collaborative groups of researchers.

Layer D6: Scientific applications, subject specific databases and user portals/clients.

Note: “D” prefix denotes relation to data infrastructure.

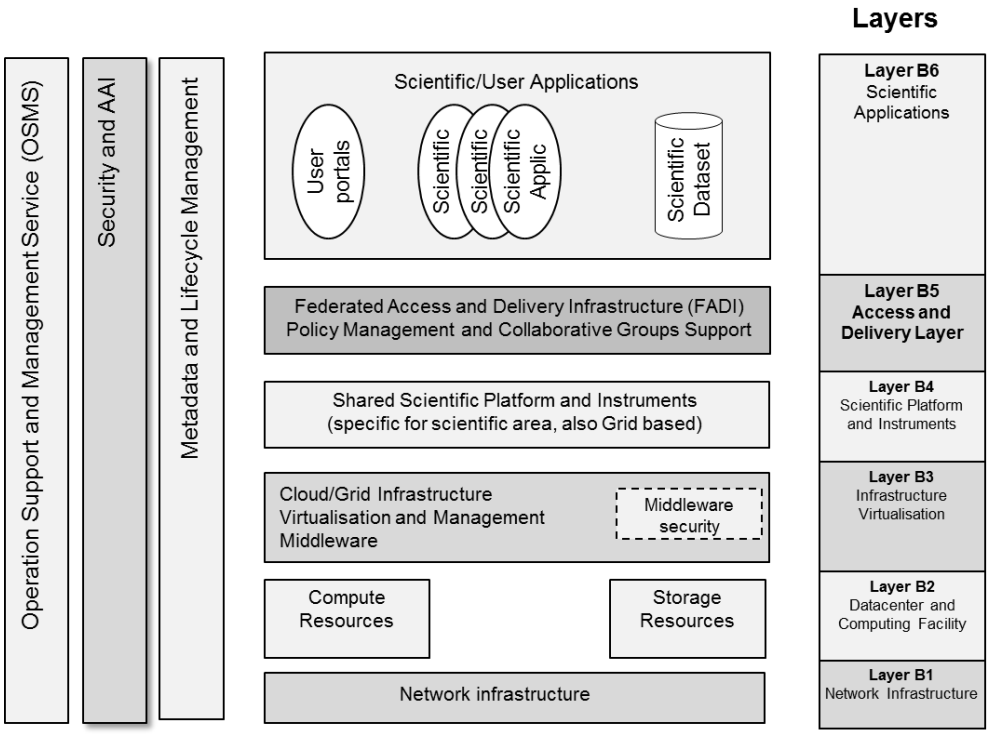


Figure 4. The proposed SDI architecture model

We define also the three cross-layer planes: Operational Support and Management System, Security plane, and Metadata and Lifecycle Management. •

The dynamic character of SDI and its support of distributed multi-faceted communities are guaranteed by the dedicated layers: D3 – Infrastructure Virtualisation layer that typically uses modern cloud technologies; and D5 – Federated Access and Delivery Infrastructure layer that incorporates related federated infrastructure management and access technologies [21, 35, 36]. Introducing the FADI layer reflects current practices in building and managing complex SDIs (and also enterprise infrastructures) and allows independently managed infrastructures to share resources and support the inter-organisational cooperation.

Network infrastructure is presented as a separate lower layer in e-SDI but dedicated network infrastructure provisioning is also relevant to the FADI layer. Network aspects in Big Data are becoming even more important than it was e.g. with Computer Grids and clouds. We can identify two main challenges that Big Data transport will impose on the underlying layer of the SDI:

- Timely delivery, in order to bring all data where required with the smallest possible latency.
- Cost reduction, in order to optimize the amount of network equipment required (either via purchasing it or on a pay-per-use) without scarifying the Quality of Service (QoS).

For many SDIs the basic best-effort Internet is the only available network transport architecture. In these cases, given the constraints imposed by this shared medium, it will be difficult to fully provide the low latency and guaranteed delivery required for Big Data processing. Performance may be lower but it will be manageable.
A smaller number of SDI will rely on circuit-based networks where the timely delivery of

data will be guaranteed but the costs for operating or using the network path will be significantly higher.
We see a third possibility for dealing with Big Data at the lowest layer of the SDI. Emerging protocols for network programmability, we can think for example of OpenFlow and in general of Software Defined Networks, provide interesting solutions. By fully controlling the network equipment both time and costs can be optimized.
Although the dilemma of moving data to computing facilities or vice versa moving computing to data location can be solved in some particular cases, processing highly distributed data on MPP (Massively Parallel Processing) infrastructures will require a special design of the internal MPP network infrastructure.

6. Cloud Based Infrastructure Services for SDI

Figure 5 illustrates the typical e-Science or enterprise collaborative infrastructure that is created on demand and includes enterprise proprietary and cloud based computing and storage resources, instruments, control and monitoring system, visualization system, and users represented by user clients and typically residing in real or virtual campuses.

The main goal of the enterprise or scientific infrastructure is to support the enterprise or scientific workflow and operational procedures related to processes monitoring and data processing. Cloud technologies simplify the building of such infrastructure and provision it on-demand. Figure 3 illustrates how an example enterprise or scientific workflow can be mapped to cloud based services and later on deployed and operated as an instant inter-cloud infrastructure. It contains cloud infrastructure segments IaaS (VR3-VR5) and PaaS (VR6, VR7), separate virtualised resources or services (VR1, VR2), two interacting campuses A and B, and interconnecting them network infrastructure that in many cases may need to use dedicated network links for guaranteed performance.

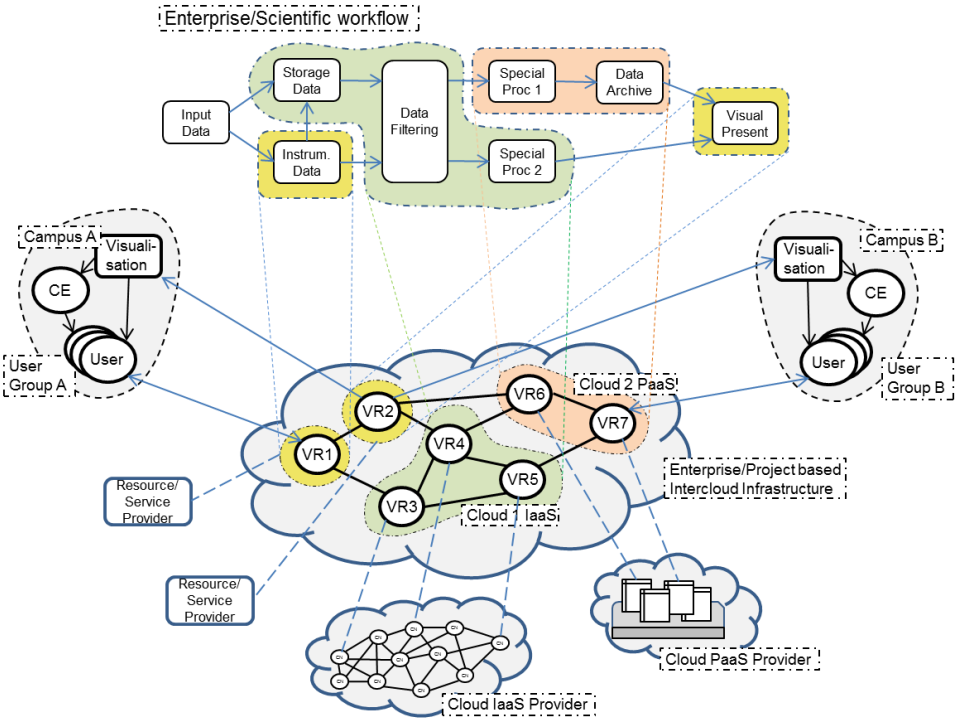


Figure 4. From scientific workflow to cloud based infrastructure.

Efficient operation of such infrastructure will require both overall infrastructure management and individual services and infrastructure segments to interact between themselves. This task is typically out of scope of the existing cloud service provider models but will be required to support perceived benefits of the future e-SDI. These topics are a subject of another research by the authors on the InterCloud Architecture Framework [37, 38, 39].

Besides the general cloud base infrastructure services (storage, compute, infrastructure/VM management) the following specific applications and services will be required to support Big Data and other data centric applications [40]:

- Cluster services
- Hadoop related services and tools
- Specialist data analytics tools (logs, events, data mining, etc.)
- Databases/Servers SQL, NoSQL
- MPP (Massively Parallel Processing) databases
- Big Data Management tools
- Registries, indexing/search, semantics, namespaces
- Security infrastructure (access control, policy enforcement, confidentiality, trust, availability, privacy)
- Collaborative environment (groups management)

Big Data analytics tools are currently offered by the major cloud services providers such as: Amazon Elastic MapReduce and Dynamo [41], Microsoft Azure HDInsight [42], IBM Big Data Analytics [43]. HPCC Systems by LexisNexis [44], Scalable Hadoop and data analytics tools services are offered by few companies that position themselves as Big Data companies such as Cloudera, [45] and few others [46].

7. Security Infrastructure for Big Data

7.1 Security and Trust in Cloud based Infrastructure

Ensuring data veracity in Big Data infrastructure and applications requires deeper analysis of all factors affecting data security and trustworthiness during their whole lifecycle. Figure 5 illustrates the main actors and their relations when processing data on remote system. User/customer and service provider are the two actors concerned with their own data/content security and each other system/platform trustworthiness: user wants to be sure that their data are secure when processed or stored on the remote system.

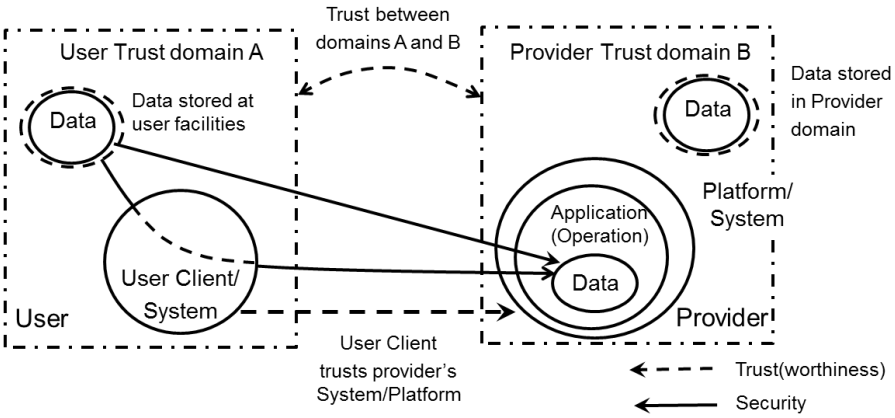


Figure 5. Security and Trust in Data Services and Infrastructure.

Figure 5 illustrates the complexity of trust and security relations even in a simple usecase of the direct user/provider interaction. In clouds data security and trust model needs to be extended to distributed, multi-domain and multi-provider environment.

In the general case of multi-provider and multi-tenant e-Science cooperative environment, the e-SDI security infrastructure should support on-demand created and dynamically configured user groups and associations, potentially re-using existing experience in managing Virtual Organisations (VO) and VO-based access control in Computer Grids [47, 48].

Data centric security models when used in generically distributed and also multi-provider e-SDI environment will require policy binding to data and fine grained data access policy that should allow flexible policy definition based on the semantic data model. Based on the authors' experience, the XACML (eXtensible Access Control Mark-up Language) policy language can provide a good basis for such functionality [49, 50]. However support of the data lifecycle and related provenance information will require additional research in policy definition and underlying trust management models.

7.2. General Requirements to Access Control Infrastructure

To support secure data processing, the future SDI should be supported by a corresponding Access Control and Accounting Infrastructure (ACAI) that would ensure normal infrastructure operation, assets and information protection, and allow user authentication and policy enforcement in the distributed multi-organisations environment.
Moving to Open Access [31] may require partial change of business practices of currently existing scientific information repositories and libraries, and consequently the future ACAI should allow such transition and fine grained access control and flexible policy definition and control.
Taking into account that future SDI should support the whole data lifecycle and explore the benefit of the data storage/preservation, aggregation and provenance in a large scale and during long/unlimited period of time, the future ACAI should also support all stages of the data lifecycle, including policy attachment to data to ensure persistency of the data policy enforcement during continuous online and offline processes.
The required ACAI should support the following features of the future SDI:

- Empower researchers (and make them trust) to do their data processing on shared facilities of large datacentres with guaranteed data and information security
- Motivate/ensure researchers to share/open their research environment to other researchers by providing tools for instantiation of customised pre-configured infrastructures to allow other researchers to work with existing or own data sets.

- Protect data policy, ownership, linkage (with other data sets and newly produced scientific/research data), when providing (long term) data archiving. (Data preservation technologies should themselves ensure data readability and accessibility with the changing technologies).

8. Summary and Future Development

The presented in this chapter research provides a snapshot of the fast developing Big Data and Data Analytics technologies that merge modern e-Science research methods and experience of dealing with the large scale problems, on one hand, and modern industry speed of the technology development and global scale of implementation and services availability. At this stage we tried to summarise and re-think some widely used definitions related to Big Data, further research will require more formal approach and taxonomy of the general Big Data use cases both in science and industry.

As a part of the general infrastructure research we will continue research on the infrastructure issues in Big Data targeting more detailed and technology oriented definition of SDI and related security infrastructure definition. Special attention will be given to defining the whole cycle of the provisioning SDI services on-demand, specifically tailored to support instant scientific workflows using cloud IaaS and PaaS platforms. This research will be also supported by development of the corresponding Cloud and InterCloud architecture framework to support the Big Data e-Science processes and infrastructure operation.

Although currently proposed SDLM definition have been accepted as the European Commission Study recommendation [21], the further definition of the related metadata, procedures and protocols as well as SDLM extension to the general Big Data lifecycle is required.

The presented research is planned to be contributed to the two standardisation bodies related to the emerging Big Data technology where authors are actively involved: the Research Data Alliance (RDA) [51] and recently established the NIST Big Data WG (NBD-WG) [52].

References

[1] Global Research Data Infrastructures: Towards a 10-year vision for global research data infrastructures. Final Roadmap, March 2012. [online] http://www.grdi2020.eu/Repository/FileScaricati/6bdc07fb-b21d-4b90-81d4-d909fdb96b87.pdf

[2] Riding the wave: How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data. October 2010. [online] Available at http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

[3] Demchenko, Y., Z.Zhao, P.Grosso, A.Wibisono, C. de Laat, Addressing Big Data Challenges for Scientific Data Infrastructure. The 4th IEEE Conf. on Cloud Computing Technologies and Science (CloudCom2012), 3 - 6 December 2012, Taipei, Taiwan. ISBN: 978-1-4673-4509-5

[4] Demchenko, Y., P.Membrey, P.Grosso, C. de Laat, Addressing Big Data Issues in Scientific Data Infrastructure. First International Symposium on Big Data and Data Analytics in Collaboration (BDDAC 2013). Part of The 2013 Int. Conf. on Collaboration Technologies and Systems (CTS 2013), May 20-24, 2013, San Diego, California, USA.

[5] Reflections on Big Data, Data Science and Related Subjects. Blog by Irving Wladawsky-Berger. [online] http://blog.irvingwb.com/blog/2013/01/reflections-on-big-data-data-science-and-related-subjects.html

[6] Extracting Value from Chaos, By John Gantz and David Reinsel, IDC IVIEW, June 2011. [online] http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

[7] The Forrester Wave: Big Data Predictive Analytics Solutions, Q1 2013. Mike Gualtieri, January 13, 2013. [online] http://www.forrester.com/pimages/rws/reprints/document/85601/oid/1-LTEQDI

[8] Dumbill, E., What is big data? An introduction to the big data landscape. [online] http://strata.oreilly.com/2012/01/what-is-big-data.html

[9] The Big Data Long Tail. Blog post by Jason Bloomberg on Jan 17, 2013. [online] http://www.devx.com/blog/the-big-data-long-tail.html

[10] The Fourth Paradigm: Data-Intensive Scientific Discovery. Edited by Tony Hey, Stewart Tansley, and Kristin Tolle. Microsoft Corporation, October 2009. ISBN 978-0-9825442-0-4 [online] http://research.microsoft.com/en-us/collaboration/fourthparadigm/

[11] Worldwide Large Hadron Collider Grid (WLCG) [online] http://wlcg.web.cern.ch/

[12] ATLAS Experiment [online] http://atlas.ch/

[13] Low Frequency Array (LOFAR) [online] http://www.lofar.org/

[14] Google BigQuesry [online] https://cloud.google.com/products/big-query

[15] Perry, T.S., The Making of Facebook's Graph Search, Posted 6 Aug 2013 [online] http://spectrum.ieee.org/telecom/internet/the-making-of-facebooks-graph-search

[16] Cole, J., How Twitter Stores 250 Million Tweets A Day Using MySQL, December 19, 2011 [online] http://highscalability.com/blog/2011/12/19/how-twitter-stores-250-million-tweets-a-day-using-mysql.html

[17] Biodiversity ;online] http://www.globalissues.org/issue/169/biodiversity

[18] Key to Innovation Integrated Solution Enabling seamless multimodality for end users. European Innovation Partnership for SmartCities and Communities" [online] http://www.eu-smartcities.eu/sites/all/files/SMP%20KI%20-%20Enabling%20seamless%20multimodality%20for%20end%20users.pdf

[19] Membrey, Peter, Keith C.C. Chan, Yuri Demchenko, A Disk Based Stream Oriented Approach For Storing Big Data. First International Symposium on Big Data and Data Analytics in Collaboration (BDDAC 2013). Part of The 2013 International Conference on Collaboration Technologies and Systems (CTS 2013), May 20-24, 2013, San Diego, California, USA.

[20] Demchenko, Y., P.Membrey, C.Ngo, C. de Laat, D.Gordijenko, Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure, Proc. Secure Data Management (SDM'13) Workshop. Part of VLDB2013 conference, 26-30 August 213, Trento, Italy.

[21] European Research Area [onlie] http://ec.europa.eu/research/era/index_en.htm

[22] European Union. A Study on Authentication and Authorisation Platforms For Scientific Resources in Europe. Brussels : European Commission, 2012. Final Report. Contributing author. Internal identification SMART-Nr 2011/0056. [online] Available at http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/aaa-study-final-report.pdf

[23] Federated Identity Management for Research Collaborations. Final version. Reference CERN-OPEN-2012-006. [online] https://cdsweb.cern.ch/record/1442597

[24] SIENA European Roadmap on Grid and Cloud Standards for e-Science and Beyond. SIENA Project report. [online] http://www.sienainitiative.eu/Repository/Filescaricati/ 8ee3587a-f255-4e5c-aed4-9c2dc7b626f6.pdf

[25] Seeking new horizons: EGI's role for 2020. [online] http://www.egi.eu/blog/2012/03/09/seeking_new_horizons_egis_role_for_2020.html

[26] Future Internet Report. UK Future Internet Strategy Group. May 2011. [online] https://connect.innovateuk.org/c/document_library/get_file?folderId=861750&name=DLFE-33761.pdf

[27] European Data Protection Directive. [online] http://ec.europa.eu/justice/data-protection/index_en.htm

[28] LifeWatch – E-Scienc European Infrastructure for Biodiversity and Ecosystem Research. [online] http://www.lifewatch.eu/

[29] ENVRI, Common Operations of Environmental Research Infrastructure [onlie] http://envri.eu/

[30] Koopa, David, et al, A Provenance-Based Infrastructure to Support the Life Cycle of Executable Papers, International Conference on Computational Science, ICCS 2011. [online] http://vgc.poly.edu/~juliana/pub/vistrails-executable-paper.pdf

[31] Open Access: Opportunities and Challenges. European Commission for UNESCO. [online] http://ec.europa.eu/research/science-society/document_library/pdf_06/open-access-handbook_en.pdf

[32] OpenAIR – Open Access Infrastructure for Research in Europe. [online] http://www.openaire.eu/

[33] Open Researcher and Contributor ID. [online] http://about.orcid.org/

[34] Data Lifecycle Models and Concepts. [online] http://wgiss.ceos.org/dsig/whitepapers/Data%20Lifecycle%20Models%20and%20Concepts%20v8.docx

[35] EGI federated cloud task force. [online] http://www.egi.eu/infrastructure/cloud/cloudtaskforce.html

[36] eduGAIN - Federated access to network services and applications. [online] http://www.edugain.org

[37] Demchenko, Y., M. Makkes, R.Strijkers, C.Ngo, C. de Laat, Intercloud Architecture Framework for Heterogeneous Multi-Provider Cloud based Infrastructure Services Provisioning, The International Journal of Next-Generation Computing (IJNGC), Volume 4, Issue 2, July 2013

[38] Makkes, M., C.Ngo, Y.Demchenko, R.Strijkers, R.Meijer, C. de Laat, Defining Intercloud Federation Framework for Multi-provider Cloud Services Integration, The Fourth International Conference on Cloud Computing, GRIDs, and Virtualization (CLOUD COMPUTING 2013), May 27 - June 1, 2013,Valencia, Spain.

[39] Cloud Reference Framework. Internet-Draft, version 0.5, July 3, 2013. [online] http://www.ietf.org/id/draft-khasnabish-cloud-reference-framework-05.txt

[40] A chart of the big data ecosystem, take 2. By Matt Turk [online] http://mattturck.com/2012/10/15/a-chart-of-the-big-data-ecosystem-take-2/

[41] Amazon Big Data. [online] http://aws.amazon.com/big-data/

[42] Microsoft Azure Big Data. [online] http://www.windowsazure.com/en-us/home/scenarios/big-data/

[43] IBM Big Data Analytics. [online] http://www-01.ibm.com/software/data/infosphere/bigdata-analytics.html

[44] HPCC Systems: Introduction to HPCC (High Performance Computer Cluster), Author: A.M. Middleton, LexisNexis Risk Solutions, Date: May 24, 2011 [online] http://cdn.hpccsystems.com/whitepapers/wp_introduction_HPCC.pdf

[45] Cloudera Impala Big Data Platform http://www.cloudera.com/content/cloudera/en/home.html

[46] 10 hot big data startups to watch in 2013, 10 January 2013 [online] http://beautifuldata.net/2013/01/10-hot-big-data-startups-to-watch-in-2013/

[47] Demchenko, Y., L.Gommans, C, de Laat, M.Steenbakkers, V.Ciaschini, V.Venturi, VO-based Dynamic Security Associations in Collaborative Grid Environment, Proc. The 2007 International Symposium on Collaborative Technologies and Systems (CTS 2006), 14-17 May, 2006 LasVegas.

[48] Demchenko, Y., C. de Laat, O. Koeroo, D. Groep, Re-thinking Grid Security Architecture. Proceedings of IEEE Fourth eScience 2008 Conference, December 7–12, 2008, Indianapolis, USA. Pp. 79-86. IEEE Computer Society Publishing. ISBN 978-0-7695-3535-7

[49] Demchenko Y., L. Gommans, C. de Laat. "Using SAML and XACML for Complex Resource Provisioning in Grid based Applications". In Proc. IEEE Workshop on Policies for Distributed Systems and Networks (POLICY 2007), Bologna, Italy, 13-15 June 2007.

[50] Demchenko, Y., C. M. Cristea, de Laat, XACML Policy profile for multidomain Network Resource Provisioning and supporting Authorisation Infrastructure, IEEE International Symposium on Policies for Distributed Systems and Networks (POLICY 2009), July 20-22, 2009, London, UK.

[51] Research Data Alliance (RDA). [online] http://rd-alliance.org/

[52] NIST Big Data Working Group (NBD-WG). [online] http://bigdatawg.nist.gov/home.php