e-IRG "Blue Paper" on Data Management

FINAL VERSION

30 October 2012

e-IRG
e-Infrastructure Reflection Group
www.e-irg.org

Table of contents

Foreword

In the emerging era of *Open Science,* research infrastructures, such as the ESFRI projects, play a fundamental role as major data production factories. For science to function effectively, and for society to reap the full benefits from scientific endeavours, scientific data must be made open[1]. Openness and sharing of data allows value creation in terms of new knowledge, products and ideas through re-using, re-purposing or computing of the data. There can of course be some limitations to complete openness induced by personal data protection, security or private investments. Still, the general principle of open science is now being embraced by a growing number of funding agencies worldwide, including the European Commission[2] [EC]

e-Infrastructures provide the tools for creating the added value from data, as well as the technologies for implementing appropriate security and data protection policies. These digital tools include high-speed internet for connecting people and data, powerful computing facilities for analysis and modelling, as well as the technologies that allow the combined use of digitally enabled resources. The purpose of this Blue Paper is to describe some basic data management principles for research infrastructures so as to facilitate the future use of the data produced in the best interest of the general public.

The e-Infrastructure Reflection Group (e-IRG) is working towards a vision of an open e-Infrastructure that enables flexible cooperation and optimal use of all electronically available resources. A first Blue Paper to ESFRI was published in 2010, describing the ways ESFRI projects and their users can engage and exploit common e-Infrastructure services to satisfy their requirements. This Blue Paper focuses on cross-cutting themes for all research infrastructures related to data management. We believe that implementation of the recommendations given in this Blue Paper will significantly contribute to an increased return on public investments in research infrastructures through easy and secure access to the data produced, improved reliability and searchability, as well as interoperability with other elements of a global data infrastructure.

Gudmund Høst

Gudmund Høst, e-IRG Chair

[1] Panton Principles, Principles for open data in science. Murray-Rust, Peter; Neylon, Cameron; Pollock, Rufus; Wilbanks, John; (19 Feb 2010). Retrieved 26.06.2012 from http://pantonprinciples.org/

[2] Neelie Kroes Vice President of the European Commission responsible for the Digital Agenda Opening Science Through e−Infrastructures European Federation of Academies of Sciences and Humanities Annual Meeting - "Open infrastructures for Open Science" Rome, Italy, 11 April 2012, http://europa.eu/rapid/pressReleasesAction.do?reference=SPEECH/12/258

1. Executive summary

The growing importance of data available everywhere and at every stage of research - from the collection of input data, through the intermediate results and their visualisation, to data interpretation and the formation of final results - is causing a rapid increase in the amount of data available and the capacity needed for its storage. The amount of data will increase by a factor of 61 over the next 10 years (source - *The 2011 IDC Digital Universe Study*). The amount of data is estimated to exceed the size of available infrastructure resources on which data can be stored by 60% (*The 2011 IDC Digital Universe Study*). For some areas, such as DNA sequencing, the estimate increase far exceeds this: the volume of public DNA sequence data doubles around every 9 months currently. In addition the importance of data is higher and higher. It has to be carefully analysed what kind of information can be safely rejected.

In 2010 the e-IRG compiled a Blue Paper on e-Infrastructure for ESFRI. This paper and the next one concern data management, which is in the opinion of the contributing groups a crucial problem worldwide. The e-IRG Strategy Document gives an outline of the ambition of e-IRG in the field of policy development, especially for data management.

There are two reasons of the importance of this subject: the enormous growth of data and the missing common infrastructure for long-term data archiving. Data curation is one of the most important problems we must solve in the near future. The data infrastructure for scientific communities was the subject of a call announced by the European Commission in 2010 (Data infrastructures for e-Science), which engaged scientists in several data-related projects.

ESFRI proposed four pilot projects which are cooperating with e-IRG on data-related issues:

- BioMedBridges: Building data bridges between biological and medical infrastructures in Europe
- CRISP - Cluster of Research Infrastructures for Synergies in Physics
- DASISH - Data Service Infrastructure for the Social Sciences and Humanities
- ENVRI - Common Operations of Environmental Research Infrastructures.

e-IRG organised a workshop (October 2011, Poznań, Poland) devoted to data management problems, where the European initiatives presented their requirements and demands (ESFRI pilot projects, DC-NET, ITER, PaNdata). A summarising panel discussion between data owners and infrastructure and service providers highlighted the critical issues, e.g. virtual integration, accessing distributed data and interdisciplinary collections based on visibility, identity, registered syntax and semantics. The outcome of the panel discussion was that we need an **INTEGRATED e-Infrastructure**, i.e. cooperating HPC, grids, clouds and data. However, the integration should start from the bottom-up.

This Blue Paper identifies the most important areas of data management and addresses the following topics:

- **Data e-Infrastructure** by Norbert Meyer
- **Reliability and replications** by Johannes Reetz, John Kennedy and Maciej Brzeźniak
- **Metadata** by Gera Pronk and Daan Broeder
- **Unified access and interoperability** by Angelos Bilas
- **Security** by Steven Newhouse and Sergio Andreozzi

with contributions from the e-IRG Data Management Task Force.

Norbert Meyer
The Editor
PSNC

2. Introduction - scope of the document

This report aims to produce a list of recommendations for data management problems. The e-IRG workshop organised in Poznań (Poland) in 2011 focussed on data services and data infrastructures. The first chapter Grand *challenges and requirements* summarises the current state of development of horizontal and integration projects (ESFRI projects) and other projects dealing with major problems in Europe and worldwide, such as cultural heritage (DC-NET), finding alternative sources of energy (ITER), and Open Data Infrastructure for Sciences with Photon and Neutron Sources (PaNData), giving a good reference for diverse community requirements beyond the ones from ESFRI pilot projects.

This chapter summarises the data parameters relevant to these areas of science presented by the aforementioned projects and determines their relevance from the most minor to critical.

Based on the analysis we have identified important and critical data management issues that describe data access, archiving, searching and movement. These are described in the following sections:

- Data e-Infrastructure (Chapter 4)
- Reliability and replications (Chapter 5)
- Metadata (Chapter 6)
- Unified access and Interoperability (Chapter 7)
- Security (Chapter 8)

This Blue paper should be viewed as an analysis of selected topics that we consider essential for data management. The paper is neither an in-depth analysis (due to the assumption of a limited number of pages), nor a comprehensive analysis addressing the entire list of potential problems. This is illustrated in Table 1 that provides a list of relevant and user-visible problems in accessing data.

In a standardised manner, each of the chapters outlines the issues analysed, describes the key actors, and presents the current state of development, especially in Europe in the context of e-Infrastructure. The last two chapters provide steps that can be taken in the near future to obtain a solution to the problem. They provide a list of proposed recommendations as well as actions to be undertaken towards realising a solution to the objective pursued in terms of the entire ESFRI community, not just the selected project or a single scientific community.

ESFRI wants to elaborate and use an ICT platform that will allow use of the European e-Infrastructure. e-IRG recommends taking certain necessary steps to tackle specific problems in data management. The primarily goal of this report is to give higher level recommendations for big scientific communities interested in accessing large amount of data for long time period.

3. Grand challenges and requirements
by Fotis Karayannis and Norbert Meyer

Data management, including data access and the support of large datasets, is still struggling with several unsolved problems. The e-IRG task force on data management released a special issue on the most urgent data problems selected by ESFRI projects. The requirements mentioned by ESFRI pilot projects were used to create a list of problems awaiting a solution from infrastructure and application providers.

During the last e-IRG workshop in Poznań (Poland, 12-13 October 2011) some special sessions were devoted to grand challenges on data-intensive processing and data management and data infrastructures. There was also a panel discussion that focused on two topics:

- How to integrate the data infrastructure with the existing grid and HPC infrastructures
- Users and infrastructure providers - demands versus offers

The demand side, i.e. demand for the use of data infrastructures and data management services, encompassed two ESFRI cluster projects presented at the e-IRG workshop in the areas of 1) Biomedical and Medical Sciences (BMS), 2) Social Sciences and Humanities (SSH), a pilot project from the Environmental Sciences (chosen as an example for closer cooperation between e-IRG and ESFRI), and three projects working on data infrastructures in the areas of physical sciences and digital cultural heritage.

The ESFRI cluster projects aim to gather all the common challenges of the ESFRI projects belonging to the same thematic area (in ESFRI terminology: Thematic Working Group) in order to develop common and efficient solutions, promote harmonisation and data interoperability, and to deal with common e-Infrastructure requirements. The corresponding projects in the four aforementioned areas are:

- BioMedBridges - developing the shared e-infrastructure—the technical bridges—to allow interoperability between data and services in the biological, medical, translational and clinical domains and thus strengthen biomedical resources in Europe
- DASISH - fostering harmonisation and interoperability between five ESFRI infrastructures within the Social Sciences and Humanities
- ENVRI - providing harmonised solutions and guidelines for the common needs of the environmental ESFRI projects, with a special focus on architecture and data issues
- and CRISP – implementation of common solutions for a cluster of ESFRI infrastructures in the field of Physics, Astronomy and Analytical Facilities.[3]

[3] Not presented directly at the workshop, but completed later

The three other projects on the demand side are:

- PaNdata - Towards an Open Data Infrastructure for Sciences with Photon and Neutron Sources
- DC-NET - A data infrastructure for digital cultural heritage: characteristics, requirements and priority services,
- ITER - the project aims at demonstrating the possibility of producing commercial energy from fusion, currently doing the computation-intensive and data-intensive simulations and data analysis/exploitation.

BioMedBridges is an ESFRI cluster project that aims to build computational and data service bridges between the ESFRI Biological and Medical Sciences research infrastructures (RI), clustering them together and linking basic biomolecular research data obtained to date in the other domains. It also includes major e-Infrastructure stakeholders such as DANTE (GÉANT network project coordinator) and EGI.eu (European Grid Infrastructure operator). Its scope is to provide secure, robust and ethical access to biodata for a wide range of bio-users (around 3 million unique users in 2011). BioMedBridges is coordinated by EMBL, who also coordinates ELIXIR, an ESFRI research infrastructure that can be viewed as a data e-Infrastructure. ELIXIR provides services for ESFRI research infrastructures in medicine, agriculture and the environment.

The biology domain has some of the most challenging growth patterns in data intensive science. The cost of DNA sequencing has halved every 6 months for the last 3 to 4 years, with considerable future headroom in the fundamental technologies. As a result, the amount of available DNA sequence data (either fully public or under access restrictions consistent with ethical restrictions; see next paragraph) has doubled every 6 to 9 months. This doubling time is substantially more than disk or CPU doubling times. Innovative developments in data-specific compression and algorithmic advances (often using established computer science techniques applied to this problem domain) have partially mitigated the challenge of managing this growth. However, the sheer scale of change has shifted much of the bottleneck in biological discovery to analysis, and much of the infrastructure bottleneck to Data I/O rates, both inside of clusters and between institutions.

Key e-Infrastructure requirements include **data protection,** as access to much of the data in the ESFRI BMS domains has **ethical, legal or societal implications** (ELSI) and includes personally identifiable information (PII). Another key requirement is **interoperability** among the different BMS domains, projects and repositories. It will only be possible to exchange and link data between the different ESFRI BMS domains if they use common identifiers, harmonised content, syntax and semantics. Shared standards will therefore be needed to allow for integration across the BioMedBridges project. Another vital aspect is a **security** framework that will address the ethical, legal and regulatory issues resulting from sharing data and providing access to biological samples in order to ensure that the infrastructure components developed are compliant with national and European regulations, privacy rules and access requirements.

BioMedBridges has now started a broad, productive interaction with the infrastructure components of GÉANT, EGI, PRACE and EUDAT. It is clear that some aspects of the "platform" infrastructure are well catered by existing plans, for example, the projected 100 Gbit/s backbone of GÉANT with the NRENs across Europe is likely to be adequate for the life science community, assuming that there is an appropriate pricing structure for access to this bandwidth. However, in other areas such as data storage, the requirement for effective Data I/O to supply computational techniques is less clearly met. Over 2012 there has been a far better dialogue between these projects and it is important that both the dialogue continues and that practical, workable solutions are proposed that bridge the computational and biological infrastructure providers.

DASISH is a cluster project for Social Sciences and Humanities that brings together five ESFRI projects dealing with social sciences data archives, arts and humanities, health, ageing and retirement and European-wide surveys. Common challenges include:

- **integration and interoperability** beyond the borders of the individual projects
- **preservation** of cultural and scientific memory and **access** to the records of science
- **quality** of data to enable advanced and cross-disciplinary access and enrichment operations
- simplified **access** conditions for researchers
- **trust** establishment of SSH researchers in the infrastructure services

In particular, collaboration with GÉANT/eduGain/TERENA/NRENs to establish a European **trust domain** allowing **single identity** and **single sign-on** mechanisms is deemed important. Finally, data sharing and archiving is also included in the list of involved projects.

ENVRI, the European environment cluster project will contribute to the construction of the ESFRI Environmental Research Infrastructures facilitating their current and future interoperability. ENVRI will not deploy a new single infrastructure for multidisciplinary collaboration, but will instead provide a flexible framework aimed at minimising divergences and seeking long-term convergence and interoperability.

The overall strategy of ENVRI focuses on:

- contributions to the architecture of decentralised infrastructures in the ENVRI cluster
- standards, harmonisation and **interoperability**
- **metadata frameworks**
- support for **access and deposit policies**
- support for a consistent European research infrastructure ecosystem
- strong interaction between ESFRI infrastructures and e-infrastructures

CRISP IT: The objective of the eleven participating Research Infrastructures (RIs) is to build up collaborations and to create long-term synergies to facilitate their implementation and to enhance their efficiency and attractiveness. Information Technology & Data Management is one of the four R&D tasks of CRISP. The importance of experimental data for modern science is constantly growing, and a new approach is required to cope with the resulting “data deluge”. The rapid development and increasing complexity of experimental techniques, instruments and detectors requires developments beyond the current state-of-the-art. To fully justify the huge investments made in scientific instruments, the data produced by these instruments must be **securely and efficiently stored, archived, annotated, queried, and linked**.

A number of concrete applications can be easily identified. A sustainable and interdisciplinary metadata management service consisting of **a federation of data catalogues** across RIs will significantly enhance scientific progress by allowing a joint access to distributed data sets or by reducing the time to search/discover distributed resources. Publications linked to properly curated openly accessible scientific data will help foster a wider public understanding of scientific findings. **A common authentication and authorisation mechanism** will allow and greatly simplify the access to distributed IT resources and remote access to the RIs. These examples show how a common IT platform for the storage, discovery, access, and processing of data can improve research in Europe. The participating ESFRI projects will develop and deploy solutions for common IT and Data Management.

PaNdata (Towards an Open Data Infrastructure for Sciences with Photon and Neutron Sources) brings together eleven major world-class European research infrastructures dealing with photon and neutron sources to create a fully integrated, pan-European, information infrastructure supporting the scientific process. Application areas where such sources are valuable include palaeoanthropology, structural biology & drug design and arts. Other sciences such as anthropology, earth sciences, chemistry and history could also benefit from these sources. Several of those user communities do not have enough experience with ICT tools, so PANdata needs to cope with such inexperienced users. The PaNdata shared data infrastructure aims to provide these user communities with data repositories and data management tools that meet the following requirements:

- deal with **large datasets** and large data rates from the experiments
- enable easy and standardised **metadata**
- allow transparent and secure remote **access** to data
- provide unique and persistent **authentication** to all users
- establish federated data **catalogues**
- allow long-term **preservation** of data
- provide compatible open source data **analysis** software
- allow overall **interoperation** through common standard schemes,
- provide a **common data policy.**

Progress has been made in several of the above areas and prototypes have been developed, such as authentication/authorisation where a shibboleth-based system has been implemented. In addition, a common data policy (including a retention policy and making data openly available after a given period), a common standard data format and a common software repository have been proposed. All of these aspects are steps towards an interdisciplinary data facility. However, many things still need to be developed with the greatest priority being interoperability with other data and computing infrastructures.

In the discussion that followed, the issue of **data sharing**, both the willingness of researchers but also the means to do so, was raised. Up until now it has not been easy for the photon and neutron sources users to share as there was no repository. A repository and a federated catalogue will help a lot, but researchers are still not always willing to share their findings; they want to keep their findings and publish first. A citation system for primary data does not exist to encourage sharing. Sharing data also requires quite a lot of work to explain the data and provide metadata. Finally, there are a lot of practical reasons to get the primary data off the facility such as proprietary formats or the fact that you cannot do another experiment before you delete the primary data!

The **DC-NET (A data infrastructure for digital cultural heritage: characteristics, requirements and priority services)** project is working on cooperation with existing e-infrastructures to define possibilities for cooperation when implementing the defined prioritised services. The amount of digitised material in the European cultural sector is growing rapidly, through national, regional and European programmes to publish the digitised content of museums, libraries, archives, archaeological sites and audiovisual repositories. This generation of data is being accelerated by Europeana that is fostering the digitisation of collections at European cultural institutions [Europeana]

The **needs** of Digital Cultural Heritage (DCH) include **data access and search, long-term preservation of content, trust, availability, reliability and security, as well as sustainability**. Furthermore, **interoperability** among existing repositories is needed. So, what is required is not a new infrastructure, but a new approach based on national and regional interoperable systems using existing resources.

Existing e-Infrastructures need to be used as a channel for digital cultural heritage data (such as connectivity, computing, storage along with related services such as authentication, authorisation and accounting).

The main actions required in the DCH area are improving awareness, promoting trust building, establishing priorities, consulting stakeholders, and promoting international cooperation. The three main projects that are involved in these areas are:

- DC-NET: joint programming for DCH e-Infrastructure implementation
- INDICATE: international cooperation, user case studies, pilots
- LINKED HERITAGE: a best practice network on metadata and standards, linked data and persistent identifiers, multilingual vocabularies, aggregation of content to Europeana.

A **prioritised list of service requirements** has been agreed upon by the digital heritage community, namely **long-term preservation, persistent identifiers, interoperability and aggregation, advanced search, data resources set-up, user authentication and access control, and finally IPR and digital rights management.**

ITER's aim is to demonstrate that it is possible to produce energy from fusion at a commercial level. The project has gathered 3000-4000 remote participants from all over the world. In ITER, the fusion reaction will be achieved in a **tokamak** device that uses magnetic fields to contain and control hot plasma. The fusion between one deuterium and one tritium molecule will produce one helium nucleus, one neutron and energy. The construction of the tokamak has started in Caradache (France). However, the first results are only expected within 20–30 years. Currently most of the R&D is done by simulations, which require large computing resources and **large datasets**. The experiments require a distributed computing infrastructure with very large operational memory, **data access and storage** (distributed exploitation), **data provenance and**

quality assurance, and data integration from multiple sources. The results are available via an international user database that is accessible to participating parties.

The table below presents the most important features of data collection and data services from the perspective of various communities.

Table 1 Most important requirements in terms of accessing and managing data. The meaning of scores: 4 - highly appreciated and critical for the user community, 3 - very important, 2 - important but not critical, 1 - nice to have, 0 - not important

Functionality /feature	Description	User Community Project Name						
		BioMed Bridges	DASISH	ENVRI	DC-NET	PaNdata	CRISP	ITER
Large datasets	Deal with large datasets and large data rates from the experiments	Yes	Yes and no (*)	Yes	Yes	Yes	4	4
Publicly available	The data are made publicly available	Yes and no	Yes and no	Yes	Yes (low resolution images)	Yes (future)	4	3
Restricted access	Restricted access for closed community of users	Yes	Yes and no	Yes	Yes (high resolution images)	Yes	4	3
Metadata structure	Enable easy and standardised metadata Establish federated data catalogues	4	3	4	4 4	4 4	4	4 (**)
Replication	Increase reliability by replicating data sources Increase accessibility by copying source to several places	2	3	2	3 3	1 0	3	4 for both
Federated AAI	Enable single sign-on in multi-	2	3		3	3	4	1 (more of an issue for simulations/not so for

	domain environment							data access)
Accounting	Allow to monitor and check resource usage	2	2		2	2	4	3
Data provenance		3	3	3	3		4	3
Integration	Provide the same set of services understandable between domains	3	2	4	2	2	3	3
Interoperati on	Interoperation through common standard schemes	4	2	4	4	4	4	3
High trust and security of accessing data	Provide unique and persistent authentication to all users	3	2	0	4	3	4	3
Reliability	Increase the QoS and SLA of data infrastructure	3	2	2	3	2	4	3
Access	Broadband data access Allow transparent and secure remote access to data	4	3	3	3 3	4 4	4	4 for both
Advanced search	Provide advanced search functionality	3	3	3	4	2	4	3
Data preservation	Allow long-term availability of data	4	4	3	4	3	4	4
Interpretatio n of data	Provide compatible open source data analysis software	2	2	2	2	3	4	3
Unified access	Overall interoperation	3	2	3	3	2	4	3

	through common standard schemes and finally a common data policy				3	4		
High quality	Quality of data to enable advanced and cross-disciplinary access and enrichment operations	3	3	3	3	2	3	1
Simplified access	Simplified access for user community	3	2	2	4	3	4	1
Searching information	Advanced search functionality	3	2	2	4	3	4	3
Important stakeholders	Please mention the stakeholders of the data	Life Sciences, Medical & Environmental industry, data subjects (patients or healthy volunteers)	Funders, Research Councils	Earth Sciences, Environmental Sciences	Cultural institutions (museums, libraries, archives), audiovisual archives, arts centres, Ministries of Culture	Life Sciences, Materials and Surfaces, Plasma and Quantum Physics, Chemistry, etc. ------- Individual principle investigator for this particular experiment	Communities Groups performing experiments	Plasma physicists, Diagnostic physicians, Engineers

Comments to the table:

(*) The five infrastructures participating in DASISH are very disparate in nature. Some are dealing with highly sensitive data, some are not. This is even the case within several of the individual member organisations that form the ERICs.

Datasets vary in size, from very small to large collections containing different media types. This is why I have noted *Yes* and *No* in the first fields of the table. I think most of us would like to make as much data as possible publicly available, but due to legislation and disclosure risks, this is not possible. Therefore we must have secure connections and authentication methods to prevent misuse of data.

(**) Catalogues are important. One idea is to produce one catalogue for each DA and to distribute copies of this to the other DAs.

The outcome of the panel discussion was that an **INTEGRATED e-Infrastructure** is needed, i.e. cooperating HPC, grids, clouds and data. However, the integration should start from the bottom going up.

The research communities need **frameworks that allow users:**

- to virtually integrate and access distributed & interdisciplinary collections (CDI), based on visibility, identity, registered syntax and semantics
- to execute automated workflows on these collections in the data centres
- to quickly and dynamically deploy services close to the data

However, it is not yet known how such frameworks should be realised.

4. Data e-Infrastructure
by Norbert Meyer

Presentation of the problem

The growing tide of data and its rising importance demand a solid e-infrastructure in Europe that can support the entire data workflow process: collection, processing, visualisation, searching and long-term preservation. It concerns all groups of users, those who want to keep 1) the raw data, 2) the visualisation of results and 3) the end results, e.g. publications, digital data or graphics. Differences in requirements may only be defined by specific demands and priorities set for different communities. Examples of this are given in Table 1 (see previous chapter). Activities and initiatives that deal with data processing have already started, such as like PRACE (high-performance computing) and EGI (grid computing). There are also commercial activities providing cloud computing and cloud storage (Amazon S3, Microsoft, EMC Atmos, Google or iCloud) [Google1],[Google2] and scientific approaches like BonFire or Open Nebula [CScape].

The most crucial aspects related to data infrastructures are:

Large datasets	Deal with large datasets and large data rates from experiments
Reliability	Increase the level of QoS and SLA, e.g. increasing reliability by replicating data sources and increasing accessibility by copying source to several places
Accounting	Allow monitoring and checking of resource usage
Integration	Provide the same set of services that are understandable (compatible) between domains
Interoperation	Interoperation through common standard schemes
Access	Broadband data access Allow transparent and secure remote access to data
Data preservation	Allow long-term availability of data
High quality	Quality of data to enable advanced and cross-disciplinary access and enrichment operations
Economic justification	As the scientific community is operating on increasingly larger datasets and want to preserve the information concerned, the infrastructure provided should have a clear roadmap of technology exchange and backwards compatibility.
Access control	Provide the infrastructure to allow for fine-grained access control to the data

A reliable infrastructure is a good and first step towards sustainability of services and long-term data preservation. There is still an unanswered question about how this sustainability can be achieved. Nevertheless, the answers to these questions are crucial for the entire development of the European e-infrastructure. We observe well-organised grid and HPC communities with consortia defined by EGI (European Grid Infrastructure) and PRACE (Partnership for Advanced Computing in Europe, PRACE-AISBL) and activities that started several years ago. From this perspective we note a lack of consolidated activities in Europe and a need to activate the field of data management and common data infrastructure.

Actors in the domain

The necessity of storing data and rendering it accessible for a long time frame is a key issue for all communities, starting from end users up to policy makers.

This report introduces six major groups of stakeholders:

- the end user
- the data owner
- the infrastructure provider
- the service provider
- the computer science researchers (on data and database management)
- the policy maker

The **end user** represents the communities that are interested in accessing the data and information made available by data owners.

Data owners are people or institutions with responsibilities including some form of providing, entering, curating, and authorising access to the data. Data owners may want to use the infrastructure and its services at different levels of abstraction, depending on their needs and ability to build their own services on top of those provided by infrastructure owners. For instance digital libraries, national and international projects may want to build their own, domain- or application-specific services or applications on top of the data services supported in the existing infrastructure. ESFRI projects are definitely good candidates to use and build additional services and applications on top of basic low level services. Good example of international co-operation in the area of distributed data and repository access and exploitation is the p-Medicine project (From data sharing and integration via VPH models to personalized medicine), which builds its own data repositories and data bases on top of the basic services for data management offered by multiple infrastructure providers.

The **infrastructure providers** in scientific communities are data centres equipped with big capacity storage, including resources with very fast access (mainly disk arrays, and SSDs) and slower but more cost-effective tape systems. The data centres are usually located within a computing centre. These data infrastructure providers define their own policies for keeping the infrastructure sustainable, and to provide roadmap technology migration and upgrading of the infrastructure. In parallel there are commercial infrastructure providers who are able deliver virtually any data storage capacity. However in this case, the costs of holding large datasets on their premises are relatively high (per storage unit) assuming that a reasonable level of SLA is considered (price of the service typically relies on the projected availability of the data and/or storage and access performance). Important infrastructure owners that are providing access to data centres and data repositories are the joint providers of networks and network services for research and higher education, like the NRENs in each EU member state, DANTE as the operator of the European GÉANT network backbone and the many institutes and universities operating on campus level. The necessary strategic direction for the networking community has been clearly described in the vision and recommendations of the GÉANT Expert Group as 'GÉANT2020', being the European communications commons and which are widely supported by e-IRG in its reaction of March 2012. As the network capacity and quality is an important part of the effective and reliable data storage and data repository access services, a complete offer can only be defined only by the combination of data centres (data resources owner) and providers at the European level and in each country (as network owners and service providers).

The **service provider** assures, in addition to the infrastructure provider, not only access to basic equipment such as servers, network and storage, but also provides an added value of the data management services by allowing the storage, access, enrichment and searching of data, e.g. the functionality provided by international projects like EUDAT (European Data Infrastructure) and national projects like PLATON with its National Data Storage (NDS) facility. Another example of consolidated services is BioMedBridges, which is aiming to establish research and data links between the ESFRI BMS research infrastructure.

The **computer science researchers** with expertise in the area of data management have already confronted challenges as part of their research that would be valuable for the data e-Infrastructure service provision. Issues that the current data infrastructure providers and data managers face today might have already been faced by computer science researcher in their data management and databases research. These may have included next generation data management systems, mechanisms and approaches for data-intensive scientific analytics, generic approaches to data integration and harmonization, process execution workflow, etc**.** Furthermore, besides their experience, computer scientists have also the theoretical background to better treat some of the challenges and influence architectural decisions. By relying on their knowledge of the foundations, of the innovative systems and solutions, they can suggest innovative solutions, but also initiate new research strand to address the open issues that will emerge in developing interoperable data infrastructures.

A major role is reserved for e-Infrastructure **policy makers** and funders. Their role is particularly important in the area of long-term data preservation, where the existence, sustainability and pricing policy of infrastructure and services should be guaranteed for at least 20-30 years (or more in case of some applications such as biomedicine and/or digital protection of cultural heritage, where data owners want to or are obliged to store the data for dozens of years or even centuries). Therefore the policies defined by the ministries of science and research of Member States and by the European Commission should take into account the perspectives that go far beyond the Horizon 2020 programme.

Current situation

Heterogeneous solutions

A challenging characteristic of the Collaborative Data Infrastructure (CDI) is that it will have to include a very diverse, heterogeneous ecosystem of data infrastructures. Scientific disciplines have different cultures and histories and work in very different contexts. They are not equally prepared to deal with the management, curation and preservation of large data volumes. Furthermore, they are al at different stages in realising the importance of preserving and re-using the data they produce. Some have already built an interdisciplinary or disciplinary integration and interoperability framework, others are only beginning to address the problem of data conservation for future re-use, and most are in between. The CDI will have to preserve what is already working and interface with it without requiring major adaptations of specific community solutions. The CDI should allow for the enhancement of capabilities, in particular for interdisciplinary use. It should also propose a framework for newcomers wishing to enter the data infrastructure. The success of the CDI therefore depends on a generic architecture that facilitates the integration of pre-existing data solutions from participating communities and data centres that support common data services. The development of such an architecture presents a significant challenge, which requires active collaboration between all actors, especially between the communities involved in designing specific services and the data centres that are willing to provide generic solutions. Irrespective of which technology is used to implement such an abstract architecture, no one single holistic system will fit. Instead, the core of such an architecture will contain independent components and registries with the potential of a broad global acceptance.

Global collaboration

The creation of an integrated and interoperable data domain - data as an infrastructure covering several layers - must also be achieved at a global level. In countries such as Australia, Canada, China, Japan and the USA, the topic is being actively discussed. In the USA, there is a similar impetus to create sustainable data infrastructures supporting world-class research and multiple disciplines[4], and several initiatives recently received NSF funds to develop cross-disciplinary infrastructures for data preservation and re-use for research (e.g., DataONE), or to deploy a prototype cross-disciplinary national data management infrastructure addressing some of the key data challenges faced by scientific researchers (e.g., the DataNet Federation Consortium initiative). The European Commission has recently published a call (INFRA 2012-3.2) to establish a platform between the EU and the US aimed at achieving full interoperability between the infrastructures of the two continents. [DataONE], [DataNet]

Cloud storage

Cloud storage refers to saving data to an off-site storage system maintained by a third party. Instead of storing information to the computer's hard drive or other local storage device, it is saved in any remote location. The Internet provides the connection between the user’s computer and the remote location and is the key enabling technology for cloud computing. .

[4] NSF Taskforce report on data and visualisation, March 2010

Beside the Internet, Network as a Service (NaaS) might need to be developed to provide a seamless integration of the "Everything as a Service" concepts as seen in IaaS (Amazon) or PaaS (VMWare).

The development of a cloud computing strategy is one of the action points from the Digital Agenda for Europe, which has the aim of making Europe cloud-friendly and cloud-active in terms of both provision and consumption [DAE].

We should consider two dimensions of cloud services:

1. the relational dimension: it is about outsourcing on a contractual base to a service provider, based on functional requirements, including performance. This is primarily an organisational, legal and financial problem. It is not new, but it is crucial for the successful use of clouds.
2. the technological and functional dimension: the services use particular technologies, offering a specific functionality for service customers and at the same time releasing them of their proprietary facilities and enabling them to profit from economies of scale and scope.

The cloud approach can be limited to a set of functionalities provided by one service provider, e.g. the commercial examples of Google (Google Cloud Platform), Microsoft (Windows Azure) or Apple (iCloud). Another solution allows the support of hybrid cloud scenarios, where an organisation has neither all of its resources in-house (e.g. a private cloud), nor all of these externalised or outsourced [BonFire], [CScape].

OpenNebula was designed to address the requirements of business use cases from leading companies and across multiple industries, such as hosting, telecom, e-government, utility computing, and grid computing. Users of OpenNebula are companies, research centres, and universities interested in a flexible, open and scalable technology solution to construct a production-ready cloud infrastructure for their internal operations, to support new IT, scientific or business cloud services, to provide an embeddable virtualisation orchestration to enhance their cloud computing platforms and to provide solutions or an open platform for innovation in enterprise cloud computing management [OpenNebula].

Clouds are definitely promising technologies in the future, especially for multi-domains (heterogeneous) and hybrid solutions.

Implementation

"A fundamental characteristic of our age is the rising tide of data - global, diverse, valuable and complex. In the realm of science, this is both an opportunity and a challenge" [HLEG]. This excerpt from the 2010 report of the High Level Expert Group on Scientific Data succinctly captures a major trend affecting almost all scientific disciplines, and the transformations required if we are to optimally manage and exploit this explosion of data. The report not only speaks of the challenges of data life cycle management to keep our data accessible and to overcome barriers, it also expresses the great opportunities for tackling the grand challenges we are faced with and for increasing researchers' efficiency. The creation of an integrated and interoperable data domain - data as an infrastructure covering several actors - must be achieved at a global level since many of the captured phenomena and the research communities are acting globally.

The challenges ahead are not only technical, but also social and organisational. Successful collaboration must be built on trust between service providers and users, but also between the researchers and disciplines themselves. Trust is also essential in the robustness and high availability of the infrastructure, the integrity and authenticity of the data collected and deposited, and the continued existence and persistence of the infrastructure and its components. The creation of an open and inclusive infrastructure requires the definition of adequate partnership rules, policies, governance structures, control mechanisms, and business and funding models. With respect to future funding models, we can refer to the High Level Expert Group on Scientific Data's document which states that "an infrastructure for scientific data has a public dimension; it should also have appropriate public funding". Needless to say, the cost of the total data infrastructure ecosystem will be critical and will require a lean approach at the technological and organisational levels.

The ESFRI projects will play an important role in information delivery for the entire global scientific community. How the access to data will be achieved is therefore vitally important. The access to data requires consolidation and integration into a single or federated infrastructure, which will allow access to relevant information quickly, without unacceptable delays, in a comprehensive and easy manner. This will make it impossible to continue the current approach in which scientific groups implement their own incompatible approaches. A common approach should be agreed in cooperation with all stakeholders, as the final result depends on all groups. It is, however, important to take into account the diversity of the user requirements, the data organisation of different groups and the actual expectations of data management services. The federated data infrastructure should therefore offer a variety of data services, based on the common set of basic data management operations and processes.

A good quality of service requires further specialisation at all levels. The data centres will provide reliable infrastructure and basic services. In addition to these services, information access solutions such as portals, database systems or applications may be realised. The infrastructure should provide a certain predefined SLA, which is a subject of negotiation between the infrastructure provider and data owner. Outsourcing may provide opportunities to realise a better quality of data access with parameters important from the end user's perspective. The data centres may also guarantee a long-term archiving of data and handle the process of migration between changing data storage and management technologies. The role of the funders (policy makers) is especially important for the preservation aspect and long-term support, while the experience of data management researchers is valuable for the design and implementation of the infrastructure and the related services. Without any long-term support infrastructures will not work and neither will these be trusted by users.

How should a vision for a global ecosystem of interoperable research data infrastructures be defined?

In the immediate future, an ecosystem of interoperable, federated research data infrastructures composed of regional, disciplinary and interdisciplinary components such as repositories, archives, library portals and data centres needs to be established to offer data services for both primary datasets and publications. This ecosystem will support data-intensive science and research, and stimulate interaction among all its elements, thus promoting multidisciplinary and interdisciplinary science.

Recommendations

- Define business cases and requirements for ESFRI projects related to data access and data infrastructure
- Define cross-related requirements generalised for all ESFRI projects
- Detail the current state of the data infrastructure within ESFRI
- Ask the service providers to provide a sustainability policy and to state what they can offer ESFRI projects
- Provide a redundant data infrastructure where strong data centres form a backbone for data access and preservation
- Define a role of each ESFRI project. For the near future, we have to assume a high degree of specialisation, where the roles of service/application provider, infrastructure provider and data owner/producer will be separated. Only really big organisations will still be able to consolidate all of these roles.

5. Reliability and replications

by Johannes Reetz, John Kennedy, Maciej Brzeźniak

Presentation of the problem

Data replication is an important function of data infrastructures for at least one of four reasons: the *reliability* of the data infrastructure, the *availability* of service, ensuring data *persistency* and the *performance* of the services.

Reliability is a prerequisite of *availability* and *persistency* but can also be the sole motivation for data replication. In the case of *reliability*, and in addition to controller-based and server-based RAID techniques, the data is replicated to different[3] preferentially distant storage facilities. If the primary data store fails, a secondary or tertiary one is available for correction. The use of heterogeneous storage technologies can further enhance the overall reliability of the data infrastructure since a technology-specific failure of the primary data store might not happen in the case of a different technology. The reliability of a data infrastructure also depends on policies, i.e. properly defined rules which are governing[5] the replication, and the quality assurance measures.

The *availability* of data services depends on the reliability of their underlying data infrastructure. In addition, designing fail-safe data infrastructures by means of rule-based data replication makes it possible to ensure the overall data integrity and therefore a higher level of service availability. This engenders a great deal of trust from the end users.

Keeping data *persistent* relies on the reliability of the data infrastructure. By replicating data to different storage facilities, preferentially geographically distributed, important measures are taken for guaranteeing safe, reliable and production-level archiving on timescales of decades. Data curation is another requirement that needs to be properly addressed, for instance by community-specific rules.

For *performance* it is possible to gain on two fronts. Firstly, multiple data locations mean that the load on the storage system is shared over multiple data servers and sites, and by avoiding bottlenecks the solution becomes more scalable. Secondly, by intelligently placing the data closer to the adequate compute capacity and the consumers it is possible to speed up access to the data.

Unfortunately, there is a price to be paid for data replication in the form of:

- increased complexity, since the consistency of replicas needs to be maintained
- higher demand for storage space (multiple data redundancy)
- higher demand for network capability, also regarding network capacity on demand

[5] E.g.: How many copies? Into which repositories shall replicas be deposited? Which kind of storage (online, nearline, offline)?

The key to achieving an optimal replication solution, mitigating the price to be paid, lies in the formation of a collaborative partnership between the user communities and the service providers of the data infrastructure.

Actors in the domain

One of the major infrastructure and service providers with experiences in the domain of data replication is the WLCG project, which supports the particle physics experiments located at the CERN laboratory near Geneva, Switzerland. The challenge faced by the WLCG was how to distribute and manage the estimated 15 PB of data to be produced annually at the LHC. The chosen solution was a tiered hierarchy of data centres that were federated using grid technologies. Data originating at CERN ("Tier 0") is distributed to 11 regional ("Tier 1") centres from which it is further distributed to the corresponding regional Tier 2 centres for analysis by the physicists. The WLCG is currently being re-evaluated. Lessons learned during the deployment and initial running of the WLCG are being used to redefine the operational model and new technologies are being considered which could lead to improved data deployment and retrieval. The CERN project includes data owners and service and infrastructure providers. Cooperating projects and scientific communities worldwide play the role of end users.

One project which aims to consider the big picture, incorporating all aspects of data management, and that makes an effort to address these aspects by providing relevant service is EUDAT. EUDAT will provide a multidisciplinary collaborative data infrastructure, the design and development of which is driven by the needs of various user communities. The infrastructure of data services is implemented by recognised data centres and service providers who can practically ensure the long-term future of their data storage resources and services[6]. An interesting challenge in data storage infrastructure compared to compute infrastructure is that one cannot as easily "time share" data storage – it is either stored or not – and so the costs of storage are not so easily shared between projects. In addition, requirements of the storage, in particular the Data I/O aspect can be quite different.

Further to the above-mentioned projects, the DASISH, BioMedBridges, ENVRI and CRISP projects will provide common solutions for ESFRI clusters in the fields of Social Sciences and Humanities, Life Sciences, Environmental Sciences and Physics, Astronomy, and Analytical Facilities, respectively. Each of these projects will have varied requirements for data replication and harmonising solutions, and allowing for data sharing would provide a great benefit to researchers across Europe.

[6] EUDAT is building a safe replication infrastructure based on iRODS [**iRODS**] in conjunction with persistent identifiers based on the Handle System [HS]. The persistent identifiers ensure that the data replication needs of the science communities are met in a reliable and scalable manner. The (community-specific) replication policies are to be embedded directly into the EUDAT infrastructure and the management overhead for the communities will be accordingly reduced.

Current situation

One of the first policy aspects to consider is “where” to store the data, i.e. *storage placement* and *content placement* within the e-infrastructure. *Storage placement* refers to the physical location of data storages on an e-infrastructure. *Content placement* refers to the question of which data storage server should actually be chosen for a given data object.

The *storage placement* is typically addressed in the early phase of a data project. Aspects to be considered are the proximity to compute resources for data analysis and which data centres are appropriate in terms of capacity, capability, organisational affiliation and trust. The choices of the *content placement* are rather dynamic and may change during the course of a project, also depending on how the user communities evolve.

Another aspect to consider is "how" to move the data. Technologies are needed to quickly and reliably transfer large amounts of data between dozens of data storage sites. Solutions such as gLite FTS (File Transfer Service) and Globus RTF (Reliable File Transfer) have paved the way, showing not only the benefits of such reliable end-to-end file transfer services but also their limitations, particularly with respect to scalability. The next generation of Web interfaces, such as Globus Online, aims to enable an easy access to file transfer services and to minimise the management overhead. In the domain of databases there are advanced data replication technologies [LSDM]. Thanks to their simplified data model, NoSQL [NoSQL] databases offer a high scalability, using large amounts of nodes and thus the ability to process large volumes of data with a heterogeneous structure - structural, semi-structural and non-structural. When deployed on a distributed infrastructure NoSQL databases can be a foundation for a highly available, scalable data platform for new applications and services. Such platforms can serve for Web applications, legacy systems, structured event logs, mobile applications or simply as document containers.

The management of large amounts of multiple replicas distributed across different sites (administrative domains) raises data consistency issues. Strategies (policies) need to be defined to treat data as they are ingested, updated and deleted. Data queries must return consistent results so that the reproducibility of any research query, one of the cornerstones in science, is guaranteed.

To make beneficial use of replicated data the data infrastructure needs to be designed so that not just *physical replication*, moving of data objects, is provided but also a full *logical replication*, where the underlying system is aware that multiple copies of every data object exist and can present them to the data consumer with appropriate efficiency.

Logical data replication can be viewed more as data placement than data transfer. Once a digital object is transferred, global or domain-specific data location services, such as persistent identifier services or file catalogues, need to be updated to ensure that the entire data stock remains consistent. Furthermore, besides replicating and registering the digital object, the access rights for the original data objects need to be correctly associated with the replicated object.

During the past few years, much valuable experience has been gained in large e-infrastructure projects such as EGEE/EGI and DEISA/PRACE as well as in the discipline-oriented infrastructure projects of many science communities (e.g. WLCG). Problems associated with data replication have been clearly identified. One of the most valuable lessons learned was the need to ensure that the physical and logical data placement is kept consistent. Inconsistencies easily lead to non-referenced "Dark Data", where the physical replication was successful but the ensuing registration in the location service failed, and "empty references" where the underlying physical data has been deleted and yet the reference to it still exists were produced.

The latest data replication technologies rely heavily on the experiences gained in the past and provide production-ready solutions for current data replication challenges. One increasingly deployed technology is iRODS. For replication services this allows the definition of policies that reflect the data management needs of the communities. Rule-based replication allows to build a large long-term persistent data infrastructure.

Implementation

There is no single approach to address all potential, domain-specific needs of the various user communities (data owners) related to data replication as well as to help optimising the costs and efficiency of providing data storage resources and services that is necessary to establish a sustainable collaborative infrastructure. However, some basic, preferably rule-based data replication functionalities should be provided at the pan-European scale while assuring adequate level of SLA. These functionalities lie at the heart of current and future data management systems, repositories and data infrastructures. The establishment of a data replication infrastructure that provides the basis for advanced, domain-specific and user community-specific systems and services would be an important cornerstone and accelerator for the process of building advanced distributed data storage, access and handling services for e-science.

Some scientific communities worldwide have already invested in and developed tried- and-tested solutions for their data management issues making use of data replication technologies. Their experiences should be analysed and documented for other user communities.

At the policy level, conditions should be created that encourage and facilitate the technology transfer and support for common initiatives such as infrastructure building and common services development. At the organisational level, a critical mass of both interested users and infrastructure owners should be brought together and a means for effective collaboration between them should be delivered. A relevant taskforce should possibly coordinate activities in this area. At the funding level, the EU should support the initiatives and projects aimed at building cross-domain data management infrastructures at a pan-European level and scale that could both benefit from and offer benefits to multiple data centres, storage systems, technologies and storage locations that are spread internationally and involve multiple, differently profiled user communities.

Recommendations

Data replication is an important function of a data infrastructure. However, from a technological viewpoint there is no single "best solution". The best ways to replicate data are highly dependent on the application, workflows and usage pattern of the user communities. The optimum use of data replication requires the needs of the communities to be specified when designing and implementing data infrastructures. Experience has shown that data infrastructures which aim to deliver scalable and configurable data replication solutions need to provide:

- reliable end-to-end data transfer tools - ‘fire and forget’ data transfer
- true logical data replication
- efficient scalable Persistent Identifier services
- replication policy and rule integration

With these core components in place, communities and service providers can jointly use data replication to realise more reliable, faster and safer data infrastructures.

However, this will only happen if an environment is created that enables sustainable collaboration among user communities and general data service providers across multiple science and administrative domains. A critical element of the process is to enable the usage and transfer of technologies and know-how developed and collected by these scientific communities with optional topical data centres and the general data centres.

5. Metadata

by Gera Pronk and Daan Broeder

Presentation of the problem

Metadata is regarded by the user community as one of the highest valued requirements in data collection and services, as can be seen in Table 1 (*Most important requirements in data access and management – see Chapter 3)*. In this chapter we highlight just a few of the current questions and issues in the metadata community.

Two important issues we would like to discuss concern **cross-disciplinary metadata** and **metadata quality versus costs**.

The most crucial aspects related to metadata:

Cross-disciplinary metadata	
Organisation	Stimulate the joining up of disparate, young and rapidly evolving centres from single disciplines e.g. biosciences
Implementation	Consolidate mature, stable centres, and enable cross-disciplinary searching
Infrastructure	Provide an overarching infrastructure that can link to the strongly established communities, without getting in their way, to support emerging disciplines, and enable sharing of data across many disciplines
Metadata quality versus costs	
Requirements	Define what metadata is minimally needed to support good data management Define metadata quality
Costs	Give advice between a metadata strategy favouring the 'archiving' perspective (store as much metadata as is available) or a consumer perspective (only store metadata that is useful and practical) so we can create metadata for many resources Examine new technologies in greater detail (e.g. exploitation of automatic metadata extracting)

Another problem that is considered beyond the scope of this document is the role of the librarians: Originally, librarians were very important in the metadata quality process but their territory is decreasing fast. How can we still benefit from their knowledge?

It is also important to recognise that, for some disciplines, there can be more than one set of annotations, in different metadata languages, for the same set of data. For example, some crystallographic data might be annotated once in CERIF to associate it with a larger research project that uses a variety of techniques, and again with a specific crystallography metadata schema. So the storage of metadata should be only loosely coupled to the storage of data. In such cases a data repository should store, for each data resource, a list of links to relevant metadata.

Actors in the domain

The C4D (CERIF for Datasets) project delivered a document in which different actors and users of dataset metadata and usage scenarios are described [C4D]:

- researchers who want to access datasets (end users)
- research organisations that are jointly responsible for datasets (data owners)
- funding bodies who fund projects where datasets are generated / used (policy makers).

Beneath we give an overview of some of the actors that are using or working on metadata.

Infrastructure and service providers

Examples of organisations that provide metadata services on behalf of the communities are EUDAT, DASISH or CRISP and INSPIRE. In the future other e-infrastructure organisations may step in, e.g. EGI.

We propose that these metadata service providers play a role in joining up 'newcomers', e.g. the **EUDAT** project aims to contribute to the production of a Collaborative Data Infrastructure. The project's target is to provide a pan-European solution to the challenge of data proliferation in Europe's scientific and research communities.

Another initiative, **OpenAIRE** has established an open access publication infrastructure to raise visibility of cross-domain European published research. Elaborate support of open science is taken care by its current project, OpenAIREplus.

Other initiatives: ENGAGE, EuroRIS-Net+.

Computer science researchers in data management

The experience of computer science researchers in the area (such as the ones that have worked in the CERIF model and metadata schema) is valuable for the introduction of related metadata schemas and services. They interact with the infrastructure and service provider people to better understand the challenges and work together in solving them.

Data owners

Data owners or information providers[8]can be split into at least four separate categories:

- information providers holding active data
- information providers curating inactive data (e.g. CLARIN)
- immature communities
- ad-hoc users

The active and immature information providers are the largest unknown factor. It is difficult to quantify their potential data holding since the impending 'data tsunami' is based on an increase in the volume of data stored, the number of distinct objects stored and the number of communities providing digital archives.

Policy makers

Both European and national funding organisations can influence metadata developments. This can be realised through both financial incentives and regulations (e.g. by asking results to be provided with certain metadata).

End users

Several user communities exist: monodisciplinary and cross-disciplinary research communities, communities based on the shared use of e-science infrastructure (instruments, grids, HPC, advanced networks) and communities gathered around (ESFRI) projects.

In addition to data owners, end users can be broadly categorised into at least four types:

- Production consumers (often related to e-research or e-science)
- Research consumers (cross disciplinary)
- So-called citizen scientists
- Policy consumers (e.g. IPCC and IPF)

It is anticipated that the heaviest users will be the production consumers who will possibly require metadata querying for the identification of candidate datasets for use in analysis and querying to obtain DOIs during processing runs. Citizen scientists could potentially introduce scalability issues depending on the number of such users. . For research infrastructures, there is a partial de-facto overlap between data owners and end users (see sect. 4). This has e.g. the consequence, that end users have influence on the definition and generation of metadata.

Current situation

The current situation can be described from two perspectives: the user community and the technology (or methodology). From both perspectives:

- the closer collaboration between different researcher communities,
- the alignment and set-up of supportive research infrastructures and
- the struggle between security and open access

are the main catalysts in the current situation.

The landscape is very varied

- from *young disciplines* (bioinformatics, climate communities) with emerging standards, rapidly multiplying file formats, evolving (and unstable) ontologies, but a strong need to link datasets together to t*he federation of related disciplines* such as Open Data Infrastructure for Photon and Neutron Sciences,

- to the more *mature disciplines* such as astronomy, with stable, well-understood ontologies, and the ability to share a common datasets across the wide astronomy community,
- and the spaces in between.

Some important initiatives in the current landscape are now described.

Metadata service managers

Metadata Service Managers are organisations that provide or will provide metadata services on behalf of the research communities. Examples are EUDAT, DASISH, BioMedBridges, CRISP and INSPIRE. In the future other e-infrastructure organisations may step in.

We recommend allowing these metadata service managers to play a role in joining up 'newcomers'.

W3C Semantic Activities - Linked Data

The current Linked Data work concentrates on publishing data for read-only usage. In the future an easy way to read and *write* data will be needed. Semantic Web technologies (RDF(S), vocabularies, SPARQL, etc.) can play a major role in publishing and using data on the Web. W3C has plans for a Linked Data Platform WG.

We recommend a more detailed study of the significance of these new technologies for the research communities.

Implementation

Three suggestions for improvements are:

- enhance the cross-disciplinary collaboration by enlarging the visibility of the metadata **community**. Make it clear who can be approached for metadata expertise (this is a combined effort of all involved)
- provide **best practices** that consider metadata value and quality and interoperability and,
- evaluate new **techniques.**

Interoperability and visibility of the metadata community

Make it clear who can best be approached for metadata questions and advice.

For example, ESFRI projects could name a metadata **contact person** and provide information about the metadata **strategy**, especially how metadata requirements and quality are treated in the project.

The metadata service managers can provide guidance, information and expertise and could help to convince researchers to provide high-quality metadata and to publish their resources.

Best Practices (metadata value and quality)

Best practices for realising the alignment or **interoperability** of metadata in cross-disciplinary collaborations should be collected and disseminated. Monodisciplinary best practices regarding metadata implementations might still be missing and these will need to be established first. One important aspect of quality metadata is "structured metadata" or contextual metadata.

Metadata catalogues

Best practices could also help in setting up metadata catalogues:

- Suggestions for metadata requirements that can be described in best practices are: minimal metadata needs, metadata structure, storage issues, archiving issues, virtual collections, Web and RDF.
- Suggestions for aspects of metadata quality that can be described in best practices are: measurement of quality, granularity, structured versus flat, specific issues concerning interdomain use, data-type checking, vocabulary services, and extraction versus manual input.

Sort out new techniques

Explicit adding of metadata is often labour intensive. It is worthwhile investigating whether new techniques can help to reduce the costs.

Another suggestion is to carefully follow the W3C semantic activities and assess their significance for several research disciplines.

User friendliness and sorting out new techniques

High-quality metadata over the whole chain from raw-data production and acquisition to the final publication are essential. Although the procedures should be automated as much as possible, input has to be provided and inserted by the researcher. The active effort to be invested must be minimized and counterbalanced with the advantages both, in respect to the amount of data and the way of provisioning.

Explicit adding of metadata is often labour-intensive. It is worthwhile to investigate if automatic and new techniques can help reduce needed resources. This, again, may decide if a metadata system should be implemented or not.

Another suggestion is to pay attention to the W3C (and NEXUS) semantic activities and their meaning for several research disciplines.

Recommendations

Establish a Taskforce

- The Taskforce should favour a practical approach, where we see the availability and quality of existing metadata as the guiding force rather than attempts to come to a new all encompassing metadata set.

Communities and interoperability

- Give metadata service managers a greater role in supporting 'newcomers' with their specific requirements to join, and set up cross-disciplinary metadata search functionality
- State metadata contact persons in ESFRI projects
- Provide recommendations for distributing responsibilities with respect to
  - metadata quality
  - discipline-specific metadata scheme and best practices
  - providing guidance for interoperability (across disciplines)
  - providing metadata services (across disciplines)

Best practices

- Both monodisciplinary and cross-disciplinary best practices will be valuable for implementing the following three recommendations:
  - enabling easy and standardised metadata
  - establishing federated data catalogues

    - paying attention to aspects of granularity. The need to describe sets of resources (datasets) at different levels of granularity calls for different metadata schemes.

Evaluate new techniques

    - Automatic metadata extraction. What is possible and impossible?
    - Check what impact will the W3C Linked Data and semantic activities have on ESFRI projects

- Check the RDF export format

6. Unified access and interoperability
by Angelos Bilas

Presentation of the problem

Data management is emerging as one of the main problems in modern research infrastructures and beyond. In particular, unified access and interoperability have attracted a lot of attention due to the data-centric nature of many applications: Most applications and services today are somehow or other fundamentally dependent on data access in a way that surpasses the capabilities of existing infrastructures. The main parameters and requirements that make data access an important and challenging problem are:

Large datasets: Datasets tend to quickly grow in size as a result of our effort to collect more data, analyse this, and gain a deeper understanding of problems. At large infrastructures there has been a tremendous push by recent developments. Novel accelerators and detectors are planned and coming into operation offering unprecedented research possibilities but with the price of producing huge datasets. Overall, datasets have been growing at faster rates than any other technology improves, leading to a data deluge. This increase in data volume places significant stress on our ability to build storage infrastructures as well as to move data between locations.

Diverse metadata: Innovation and new knowledge rely to a large extent not only on the data itself but also on the metadata that have been generated. In addition, they create new metadata and in forms that are not easy to predict. This diversity in the type and form of information that is created introduces additional complexity to our effort to organise data in an efficient and convenient manner.

Many locations: Sources that generate data and users or applications that make use of data are by necessity remotely located. Federated resources are an important trend and are projected to become the norm in the near future for both latency and cost reasons.

User Requirements: Application and service requirements for data and metadata span a broad range from performance to longevity, from protection to ease of access, and from user to management operations. All of these need to be taken into account to better serve the needs of each community.

Data access: In highly-competitive research environments data need to be protected during the embargo period (covering typically the time span between data taking and publication). Data are often scattered over several infrastructures and home institutions. Members of an experimental team need an efficient, secure and user-friendly access to physically distributed data sets.

Open infrastructures: Traditionally, storage and data management infrastructures incur significant costs and cannot be easily replaced from generation to generation. This results in a tendency to use multi-vendor hardware and software components and thus to diversify the approaches used for dealing with data access and management requirements.

In this landscape, new technologies and policies for unified access and interoperability are an important element that will allow applications, services, and users to better achieve their goals. It is easy to imagine the holy grail of unified access and interoperability, where data is generated, accessed, and managed in an infrastructure-agnostic and location-agnostic manner and without compromising data quality, consistency, protection, and performance. However historically, achieving this goal has been challenging with only partial solutions emerging so far.

Actors in the domain

The main actors for unified access and interoperability include user communities, infrastructure and service providers, platform (software and hardware) providers and vendors, research organisations and computer scientists with expertise in data and database management.

End users and data owners play an important role since they will specify requirements for each application domain. The expectation is that users will provide feedback towards forming different SLA-type specifications for unified access and interoperability for each application domain. These specifications will depend, for example, on whether data generation can be collocated with applications, if there is need for access control, and whether replication is required for performance or availability purposes. End users and data owner include scientific communities using specific infrastructures and datasets, existing virtual organisations for collaboration purposes, and domain-specific standardisation working groups.

Service providers will need to extend existing systems with features that cater for unified access and interoperability since to date most existing components, and in particular storage access and management systems, do not adequately support diverse requirements, especially from federated systems. Such organisations include open software providers for infrastructure and data access at different layers, standards organisations and bodies (OGF, ISO), vendors of domain-specific platforms (DBMS, geospatial), community-specific and commercial cloud systems, virtual organisation and virtual lab platform consortia, and dataset and content management organisations.

Infrastructure providers will play a key role since they need to balance user requirements and platforms when building and offering both convenient and viable infrastructures. Such organisations include NRENs, large research infrastructures, data centre and HPC centre hosting organisations (including PRACE and EGI organisations), coordinating bodies such as TERENA, standards organisations (OGF, ETSI, NIST, ISO, IETF, ITU), and network infrastructure organisations such as GÉANT.

Computer science researchers with expertise in data and database management have already faced access and interoperability challenges as part of their research in many occasions and for a long period of time. Such background knowledge is valuable for the infrastructure and service provision and computer science researchers can better grasp the related domain scientists' requirements and suggest innovative solutions in the design and implementation phase. Furthermore, new research strand can be proposed to address the open issues that emerge in developing interoperable data infrastructures.

Policy makers need to ensure adequate flexibility and at the same time sufficient protection for remote storage and access to data and related services. An environment that supports seamless access to remote facilities and allows interoperability of infrastructures provides substantially broader capabilities and blurs the boundaries between the responsibilities and rights of various actors.

Current situation

Unified access to existing research infrastructures has been the subject of extensive previous work in the context of grid systems and infrastructures, such as EGI and EGEE. In addition, work has focused on offering the ability to access infrastructures not just for basic services but also for running complex workflows as well [EEF]. Recently, there is a trend to divide emerging approaches for dealing with unified access and interoperability in three categories:

- Infrastructure as a Service (IaaS): In cases where requirements lead to offering the storage infrastructure to users and applications, there are available mostly low-level solutions with little functionality. Infrastructures tend to offer custom APIs. Applications and users have to essentially make a firm choice upfront for the type of infrastructure and API, which is not an easy task.

- Platform as a Service (PaaS): In these cases infrastructures offer a software stack that can be used to develop applications. Although this approach provides higher-level APIs and services to application developers, it is even less portable and provides less flexibility compared to IaaS approaches.

- Software as a Service (SaaS), where a full service is offered to users, e.g. a portal that includes all functional aspects required and allows users to interact with data only via pre-specified functions and operations.

In all cases, there is currently little or no provision for seamless federated resource access, except for cases where the provider manages multiple locations. In these cases, federated infrastructures are typically located within a single administrative domain (the provider) and so the user can exert little control. Given the current state-of-the-art in unified access and interoperability times, users and applications often operate in the mode "copy-out, use, copy-in" and by explicitly collocating data and computation. Data is fetched in the application/user location, where it is processed and if data or metadata is updated then at the end of the session it is written back, if the data access policy permits this. This approach has a number of limitations: it is cumbersome for users, it may create consistency issues, it does not necessarily cater for performance or longevity, and it cannot scale, as it requires extensive management operations from users.

Instead, it is important to move towards approaches where research infrastructures offer true, location-agnostic and infrastructure-agnostic access at all levels of access: infrastructure, system, and application levels.

Implementation

For the implementation of recommendations in this direction, a taskforce needs to function as the centre for gathering information from actors, such as requirements from user groups, studies on the limitations of current solutions as well as technology trends, and proposals for architectural directions and paradigms that will best address unified access and interoperability concerns. The taskforce should aim to produce specific reports on requirements, APIs and semantics, metadata, and policies. The reports should function as a consolidation of current information and legacy approaches as well as draft designs for further discussion and improvements. Inclusion of computer scientists in the taskforce that will bring their related knowledge and skills will be beneficial for all the above areas.

APIs and semantics: The current situation should be examined and recommendations be made about priority areas for the needs of the research community. For instance, there is a need for access to and management of APIs for each level of access (NaaS, IaaS, PaaS, SaaS) and a process to review suggestions and recommendations in a timely manner. For example, VRE (virtual research environments) needs to be available to give researchers access to integrated data and interoperability mechanisms.

Metadata requirements: The ESFRI projects should take into account issues such as cross-application and cross-domain interactions, large and small data, and human-generated and machine-generated data. For instance, given the size of future datasets it is important to ensure efficient methods for moving large datasets across systems, either among federated infrastructures or between infrastructures and user sites. ESFRI should also examine the aggregation of metadata to build a variety of (domain-specific) portals and how metadata should use registered schemes and concepts as well as standard access protocols. A working group within ESFRI should also examine induced requirements to network infrastructure and communication protocols and analyse and project limitations and requirements.

User requirements for placement and access: The end user and data owner should examine how we can inherently support federation in research infrastructures and allow for dynamic policies that will optimise the use of infrastructures to serve higher loads, improve adherence to SLAs, or reduce operational costs, based on user access patterns. This analysis should also cover an analysis of mechanisms and support for diverse SLAs, including performance, availability, reliability, longevity, online and offline protection, and access control across infrastructures and platforms. It should also explore high-level approaches to both specifying SLAs from the user perspective and mapping these to operational requirements for providers.

User requirements for replication: New approaches to distributed replication of data to mitigate the impact of latency and improve data availability without compromising consistency. Such mechanisms need to take into account hot and cold data, application-specific knowledge about data use patterns, opportunities and limitations when partitioning data, and infrastructure characteristics in different locations.

Research organisations need to explore solutions to novel mechanisms for distributed data access and management that do not compromise application requirements, e.g. on consistency or performance, as well as explore policies that serve both user communities and management organisations. Research organisations in areas such as storage systems, networking, data management and scalable infrastructures, play an important exploratory role, especially before specific approaches and solutions start to be deployed.

These taskforce results, reports, and research should eventually lead to:

- The formation of a pan-European inter-IP (infrastructure provider) architecture based on a federated approach. Brokers can play the role of intermediate entities that facilitate the implementation of the federated approach allowing an end user/data owner to seamlessly access data stored on multiple (local and/or remote) sites.
- Convergence being achieved between network and compute/data infrastructure providers for seamless and uniform connectivity to resources. NRENs have experience with both types of resource access in a variety of research environments and can play an important role here.
- Financial aspects of a federated (public/private) architecture being addressed. In a market where infrastructure providers offer a variety of service offerings and cost models, a broker should perform online optimisation of the cost of federated services, and aim to match the provider’s financial model to that expected by the user community.

Recommendations

Future research infrastructures, platforms, and services need to provide fundamental support for unified access and interoperability in collaboration with user and application communities. In particular, there is a need for following actions:

- **APIs and semantics:** Form an ESFRI taskforce on common interfaces across infrastructures to examine both client-to-infrastructure and infrastructure-to-infrastructure APIs and semantics, taking into account current efforts in both industry and research initiatives.

- **Metadata requirements:** Form an ESFRI taskforce to discuss data and metadata structure and implications for unified access.

- **User requirements for placement and access:** Form an ESFRI taskforce to examine how we can inherently support federation in research infrastructures and allow for dynamic policies that will optimise the use of infrastructures to serve higher loads, improve adherence to SLAs, or reduce operational costs, based on user access patterns.

- **User requirements for data accessibility:** Users should define their requirements for latency and improved data availability in future applications.

- **Research in unified data access:** Explore new mechanisms and policies for distributed data access and management via applied exploratory projects among stakeholders and research organisations.

7. Security

by Steven Newhouse and Sergio Andreozzi

Presentation of the problem

e-Infrastructures are continuously growing in capacity and a number of connected organisations are sharing their ICT assets. Research communities are evolving their *modus operandi* and are progressively relying on e-infrastructures to perform their day-to-day research activities. The amount of digital data generated is growing at a speed that outperforms the capacity of a single research team in a single location to perform all the validation, analysis, visualisation, storing and curation tasks.

The paradigm shift from managing data in a local and dedicated infrastructure to a distributed and shared infrastructure built using resources in both public and private organisations spanning different countries worldwide is raising new challenges in the area of security at different levels: technological, operational and regulatory.

The first challenge to address concerns provenance information: how to judge the reliability and authenticity of data that is stored somewhere in a shared infrastructure? Who generated the data, how and when? How has the data been transformed? Data provenance information is essential in order to evaluate the quality of data based on its initial source and derivations, track back sources of errors, and provide attribution of data sources. Provenance is also relevant at the business level, e.g., to track the creation of intellectual property or to provide an audit trail for regulatory purposes.

Identity and access management therefore becomes a critical issue in a shared distributed infrastructure. Defining digital credentials, evaluating the degree of confidence that this is associated to the related entity (also known as level of assurance) and making credentials portable across heterogeneous systems are key aspects for many entities such as data, dataset, users, groups or organisations.

Another challenge that needs to be addressed is the issue of controlling who can access the data and what kind of operations can be performed. In recent decades, a variety of access control models have been proposed, each designed to address different aspects of the problem. Regardless of the approach, they all rely on the following four components: identification, authentication, authorisation and access decision. The authentication is usually done in the user's domain, while the access decision is taken in the service's local domain at the time of the request.

Actors in the domain

Data owners: Researchers all over the world, either through local activity or as part of global collaborations, are producing a vast amount of data from instruments or simulations, and which needs to be stored and made available for later analysis and exploitation. The policy around data access will vary from community to community. Some data may be released immediately for unrestricted access whereas other data may have restricted access for a specified time period before being available for open access.

Infrastructure providers own the physical infrastructure used to host the datasets. Either as generic or community-specific data centres they are tasked with running the required data access services for the data they host and reliably hosting the data they hold on trust for the communities they support.

End users are researchers who wish to discover and use data generated by their own or other research communities to further their own research goal. They will need to be able to search metadata catalogues to find the data they require, and then prove (if needed) that they have rights to access the data under the published data access policy.

Service providers have a vital role in establishing the trust fabric between the data owner, the infrastructure provider that hosts the data, and the individual user that consumes the data. The service providers will need to be able to assert the entities identity (authentication) and manage the allocation of any attributes that they may possess (e.g. institutional role, project membership, etc.) so that authorisation policy can be enforced when a data service is accessed.

From the perspective of EU regulations we also identify:

- **data controller**: the entity that determines purposes and means of the processing of personal data and are responsible for compliance with data protection law
- **data processor:** the entity that processes personal data on behalf of the controller

Current situation

The data producers and the resource providers tasked with hosting the generated data are clearly a coupled problem. In the BMS community, EBI relies on a two-track approach: curated warehousing of the data by EBI, and locally stored datasets with federated central metadata searching through DAS. Generally, these datasets are open to all researchers with only a small subset of data containing personal information being subject to access control restrictions. The WLCG community uses a tiered model of data storage to provide geographical redundancy and to distribute access across multiple sites with all data being open to the entire community. The SSH community has a very distributed model of storage and ownership of its dataset. Consequently any access control model needs to be strongly rooted around the institutions that either own or wish to establish access.

Identity in the WLCG community is driven by a certificate-based model with the trust anchors being provided through the IGTF. Access to restricted datasets in EBI is provided through a conventional user name and password model. The SSH community has been exploring the use of federated identity model to provide authentication that is closely aligned to institutional model.

Implementation

The distributed model of all research communities in the European Research Area requires that any potential approach be built upon a federated model. No institution can hold the identity credentials and access control attributes for the entire research community that spans Europe or the world. Institutions are a source of validated attributes that are linked to an individual through their place of employment or study. Initiatives such as the European Citizen Card (ECC) may provide a route through which a government-issued electronic identity token can be associated with institutional or community attributes to drive authorisation decisions.

For the foreseeable future different identity tokens and attributes will be used across Europe. Establishing mechanisms to bridge between different attributes, to harmonise attributes and to transform identity tokens must be a priority. Much work has already been done to establish demonstrators for particular communities or service providers. Now there is a need to establish a federated identity model that can satisfy the requirements of the European Research Area by being ubiquitously adopted throughout all research organisations and their service providers. We must move from being ready for a federated identity to being active in a federated identity!

Many of the concerns that need to be addressed in this area are being tackled through a number of different initiatives both within and outside the scientific domain. To facilitate the rapid establishment of a sustainable mechanism for ubiquitous deployment, we propose setting up a taskforce with key representatives drawn from the important stakeholders in the relevant areas. These are mainly experts from the scientific communities (e.g., ESFRI cluster projects representing the needs of the ESFRI community, EIROs, Photon and Neutron RIs) whose community would be using their federated credentials to consume these services, service providers from the e-infrastructure community (e.g., EGI, EUDAT, PRACE, GÉANT & NRENs) and elsewhere, and policy makers (e.g., officers of the EC expert in the area). Based on the usage cases previously provided (e.g., Federated Identity for Research Collaborations [FIM]) and the existing demonstrator projects (e.g. REFEDS [REFEDS], CRISP, PaNdata) the taskforce can start to define a technology implementation and roll-out plan that can be effected during the early years of Horizon 2020. The taskforce should organise open workshops at key community events for the stakeholders (e.g., EGI forums, PRACE events, ESFRI workshops addressing technical aspects of security) to disseminate the work in progress and to gather feedback from the community.

Data protection is a crucial aspect for ESFRI projects. The movement and storage of personal data across national borders within Europe and beyond is becoming a concern for many research communities and their resource providers. Clear advice is needed to the community on the impact of national and European legislation and the use of data encryption to protect the confidentiality of data stored remotely. Furthermore, the EU Data Protection directive currently under review should address the following aspects in order to build trust in the online environment, which is good for both individuals and businesses:

a. Clear identification of “data controller” and “data processor”

b. “Right to be forgotten”: if an individual no longer wants his/her personal data to be processed, and there is no legitimate reason for an organisation to keep it, it must be removed from their system
c. “Right to data portability”: the right to obtain a copy of their data from one Internet company and to transmit it to another one without hindrance from the first company
d. Ensuring a single set of rules applicable across the EU
e. Clear rules on when EU law applies to data controllers outside the EU
f. Increased responsibility and accountability for those processing personal data
g. Whenever consent is required for data processing, it will have to be given explicitly, rather than be assumed.
h. Data owners should have control over or knowledge about the physical location where their data are maintained (country level) and the related level of data protection and privacy granted by regulations (data location).

Recommendations

- Investigate the security requirements of all ESFRI projects.
- Check if the community is ready to use a federated authentication process and the cost of the transition phase
  Federated authentication process: it should be user friendly, simple and intuitive. The user should be able to handle the authentication process with a user experience comparable to the most common Web applications, and leverage existing institutional or national electronic identities (such as the European Citizen Card) to access institutional or community attribute servers to gain access to distributed data and services. The process should also handle different agreed levels of authentication assurance and the consequences of authentication errors or the misuse of credentials should be clearly identified. A user should be given full visibility and control over the attributes that are needed/going to be delivered to a service. The release of some attributes may be mandatory. In the event that the user does not accept this then he/she will be denied to access the service.
- **Implement data encryption**: data-centric, file-level encryption that is portable across all computing platforms and operating systems should be available to users as a means of increasing data protection, confidentiality and integrity in transit and at rest.
- Influence the EU Data Protection directive under revision. In particular, this should address the aspects which are important from the data owner and end user perspectives.

8. Summary

This Blue Paper focuses on issues related to data, probably the most important topic for the scientific community at present. However, they are not unique in this respect. Rapidly increasing amounts of data cannot be accommodated in the infrastructure capacity currently available. The amount of generated data has already exceeded the size of available infrastructure resources on which data can be stored by 60% (source - *The 2011 IDC Digital Universe Study*), and in some fields there is a very rapid growth of data volumes above the growth of data storage technologies.

A review of ESFRI projects led to the following issues being distinguished as relevant in the first step of the analysis. The analysis of these issues is extremely important in the context of data management:

- access and management of data infrastructures
- reliability of services
- metadata
- unified access and interoperability of data structures
- security

The requirements collected from end users show that important is data curation - the access to infrastructure, which guarantees its stability over the next 20 - 30 years of archiving it in a predictable long-term business model. This applies to the scientific communities' experimental data, raw data and final results as collected, for example, by ITER, PaNdata, CRISP, and BioMedBridges. The key problem is the durability of archived data,

The data infrastructure is critical for applications, and its durability and reliability are vital for the quality of services provided on it.

The key question for ESFRI is who will provide the data infrastructure:

1) the ESFRI project, which will act as an infrastructure provider, service provider and data owner?

or

2) an advanced operator that provides a specialised infrastructure, such as a scientific or commercial data centre?

The infrastructure provider will have to provide reliable access to its production infrastructure with a certain level of SLA and a defined sustainability of at least 20 - 30 years. Replication functionality will have to be provided for reliability to be achieved.

The near future will see even more specialised roles for stakeholders and greater demands in relation to each of the actors. Stakeholders will therefore have to consider whether there will be more advantages from the added value and economic justification of outsourcing data infrastructure and basic services, and focus on the role of the data owner that complements their database. To help stakeholders take this decision a prepared set of recommendations has been included in Chapters 4 and 5

The data provided in the network should be accessible to a wide range of end users. We know that data infrastructure will not be homogeneous. Interoperability of services in a global context must therefore be ensured so that all infrastructures can be integrated and the ‘islands of data’ eliminated. Unified data access is also needed. Only such prepared data structures and services with a common base can be fully utilised by users (Chapter 6).

Providing raw data is no longer sufficient in the era of data integration. Additional information about the data is also required. Although the expression "data about data" is often used this can be misleading. Structural metadata, the design and specification of data structures, cannot be about data, as in the design phase the application contains no data. In this case the correct description would be "data about the containers of data."

The harmonisation of metadata is a problem for which a solution has not yet been found. This hinders global data access and searching. As data owners, **ESFRI projects should be engaged in the debate about the definition of metadata. Recommendations related to these actions are given in Chapter 5.**

Data security includes the security of stored and transmitted data as well as data access security (authentication). The federated authentication process is popular in networking communities, especially in GÉANT and NRENs and could be widely ported to the e-Infrastructure, allowing end users to own only one certificate for almost all services. This approach is included in one of the recommendations. Another recommendation proposes the use of data-centric, file-level encryption that is portable across all computing platforms and operating systems and should be available to users as a means of increasing data protection, confidentiality and integrity (Chapter 7).

APPENDIX A: Bibliography

[AAA] Study on AAA (authentication, authorisation and accounting) Platforms For Scientific data/information Resources in Europe https://confluence.terena.org/display/aaastudy/AAA+Study+Home+Page

[BonFire] http://www.bonfire-project.eu

[CScape] Cloudscape IV, Advances in Interoperability and Cloud Computing Standards http://www.trust-itservices.com/uploads/Publications/CloudscapeIV_PositionPapers_Feb%202012.pdf

[C4D] http://cerif4datasets.wordpress.com/2012/03/23/c4d-metadata-ontology

[DAE] Digital Agenda for Europe, Annual Progress Report 2011 http://ec.europa.eu/information_society/digital-agenda/documents/dae_annual_report_2011.pdf

[DataNet] http://www.renci.org/news/releases/nsf-datanet

[DataONE] http://www.dataone.org

[EC] http://ec.europa.eu/research/science-society/document_library/pdf_06/era-communication-towards-better-access-to-scientific-information_en.pdf

[EEF] European e-Infrastructures Forum http://www.einfrastructure-forum.eu

[Europeana] http://www.europeana.eu/

[FIM] Federated Identity Management for Research Collaborations, Research Paper, CERN-OPEN-2012-006, https://cdsweb.cern.ch/record/1442597

[GN2020] "Knowledge without Borders GÉANT 2020 as the European Communications Commons", Report of the GEANT Expert Group, Oct. 2011, http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/geg-report.pdf

[Google1] https://developers.google.com/storage/

[Google2] How the Apple iCloud compares to Google's cloud http://www.computerworld.com/s/article/9217438/How_the_Apple_iCloud_compares_to_Google_s_cloud

[HLEG] High Level Expert Group on Scientific Data. 2010. *Riding the wave. How Europe can gain from the rising tide of data.* Final report submitted to the European Commission, European Union http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

[HS] Handle System http://www.handle.net

[LSDM] A Survey of Large Scale Data Management Approaches in Cloud

Environments, Sherif Sakr, Anna Liu, Daniel M. Batista and Mohammad Alomari

[NoSQL] C. Strauch. NoSQL Databases. Feb. 2011. Accessed 25 January 2012 http://www.christof-strauch.de/nosqldbs.pdf

[OGF] Open Grid Forum http://www.gridforum.org

[OpenNebula] http://www.OpenNebula.org

[REFEDS] REFEDS https://refeds.org/

[W3C] “What is Happening in the W3C Semantic Web Activity?”, Presentation by Ivan Herman (Semantic Web Activity Lead) http://www.w3.org/2012/Talks/0315-Luxembourg-IH/

APPENDIX B: Glossary and abbreviations

AAI - Authentication and Authorisation Infrastructure refers to the systems used to identify and authorise users of shared resources. Authentication is the process of verifying or disproving a claimed electronic identity; authorisation is the process of deciding if a request to perform an action on a resource shall be granted or not. AAI includes authentication and authorisation services, components for identity and privilege management, and the entities responsible for these services.

APIs - Application Programming Interfaces

BioMedBridges - Biological and Medical Sciences cluster project coordinated by ELIXIR, building data bridges between biological and medical infrastructures in Europe

BMS - Biomedical and Medical Sciences

Capability computing refers to serving at one single moment in time a coarse number of specialised computing tasks requiring an extremely powerful and tightly integrated computing system. Capability computing can be also referred to as high-performance computing (HPC).

Capacity computing refers to serving an extremely large number of parallel tasks on a large-scale computing infrastructure. Capacity computing can be also referred to as high-throughput computing (HTC) or grid computing.

CDI - Collaborative Data Infrastructure

CERIF - The Common European Research Information Format

Cloud computing (or simply 'Cloud') is an on-demand service offering a large pool of easily usable and accessible virtualised resources (such as hardware, development platforms and/or services) in a pay-per-use model. Clouds are usually offered commercially and currently use proprietary interfaces.

CRISP - Cluster of Research Infrastructures for Synergies in Physics. CRISP is a partnership within FP7 which builds collaborations and creates long-term synergies between research infrastructures on the ESFRI (European Strategy Forum on Research Infrastructure) roadmap in the field of physics, astronomy and analytical facilities to facilitate their implementation and enhance their efficiency and attractiveness.

C4D - CERIF for Datasets

DASISH - Data Service Infrastructure for the Social Sciences and Humanities

DCH - Digital Cultural Heritage

DC-NET - A data infrastructure for digital cultural heritage: characteristics, requirements and priority services

DEISA - Distributed European Infrastructure for Supercomputing Applications is a series of FP7 EC co-funded projects interconnecting major high-performance computers around Europe.

DOI - Digital Object Identifier

EBI - European Bioinformatics Institute

EC - European Commission

ECC - The European Citizen Card

eduGAIN aims to provide the means for achieving interoperation between different authentication and authorisation infrastructures. It enables the sharing of identify data between different federations over existing organisations and policies. It therefore plays the role of a confederation: a federation of federations (see also Federation).

EEF - The European e-Infrastructure Forum is a forum for the discussion of principles and practices to create synergies for distributed Infrastructures. The initial membership included GÉANT, TERENA (research networking), EGEE, EGI (grid computing), DEISA and PRACE (high-performance computing).

EGEE - Enabling Grids for E-sciencE was a series of FP7 EC-co-funded projects interconnecting more than 100,000 computers in Europe and beyond. EGEE serves e-Science. When EGEE-III ended in April 2010, EGI took over the current infrastructure (supported by the EC-co-funded EGI-InSPIRE project).

EGI - European Grid Initiative is the next phase in the implementation of capacity computing in Europe. EGI unites the resources of the NGIs, guaranteeing transnational access to data and services.

EGI.eu is the legal body that hosts the EGI headquarters. It includes personnel with central responsibility, as well as the management structure. EGI.eu is located in Amsterdam (after a bidding process) and the EGI.eu team is currently being recruited.

e-Infrastructure covers ICT-related infrastructure and encompasses, among others, networking, computing, data and software components. e-Infrastructure by default refers to research, as the term was introduced by the EC, and can be also described as e-RI (in ESFRI terminology).

EIRO - The European Industrial Relations Observatory

ELSI - Ethical, Legal or Societal Implications

EMBL - European Molecular Biology Laboratory is a major research centre coordinating molecular biology research.

ENVRI - Common Operations of Environmental Research Infrastructures

EPIC - The European Persistent Identifier Consortium provides a Service for the European research community.

e-Science is the invention and application of ICT-enabled methods to achieve better, faster or more efficient research, innovation, decision support and/or diagnosis in any discipline. It draws on advances in computing science, computation and digital communications.

ETSI - The European Telecommunications Standards Institute

EUDAT - European Data Infrastructure

EUGRIDPMA - European Grid Policy Management Authority is the coordinating body of the national Certification Authorities (CAs) in Europe.

Federation is a group of organisations whose members have agreed to cooperate in a particular area, such as in the operation of an interorganisational AAI (a Federated AAI or an AAI Federation).

FTS - File Transfer Service

GÉANT is the pan-European data network dedicated to the research and education community. Together with Europe's national research networks (NRENs), GÉANT connects 40 million users in over 8000 institutions across 40 countries.

GÉANT 2020 - the European communications commons, where talent anywhere is able to collaborate with their peers around the world and to have instantaneous and unlimited access to any resource for knowledge creation, innovation and learning, unconstrained by the barriers of the pre-digital world. (GÉANT Expert Group)

gLite - Lightweight Middleware for Grid Computing

GN3 is the latest GÉANT project, coordinated by DANTE and co-funded by the EC, http://www.geant.net/

Grid is a system that federates, shares and coordinates distributed resources from different organisations that are not subject to centralised control, using open, general-purpose and in some cases standard protocols and interfaces to deliver non-trivial qualities of service. Grid computing is used by VOs.

HPC – High-performance computing: See capability computing.

HTC – High-throughput computing: See capacity computing.

IaaS - Infrastructure as a Service

ICT is the standard abbreviation for Information and Communication Technologies.

IETF - Internet Engineering Taskforce is a large, open, international community of network designers, operators, vendors, and researchers concerned with the evolution of Internet architecture and the smooth operation of the Internet.

IGTF - International Grid Trust Federation is a body working to establish common policies and guidelines between members of its Policy Management Authorities (PMAs) in the different regions (EUGRIDPMA is the European PMA).

IPR - Intellectual Property Rights refer to the controlled right of use of created items, so that the creator benefits from that use. Intellectual property is broken down into several types, each of which apply to different created items: copyright, designs, patents, trademarks, protection from passing off and protection of confidential information.

iRODS - Integrated Rule Orientated Data System

ISO - International Organisation for Standardisation

ITER - International Thermonuclear Experimental Reactor is a joint international research and development project that aims to demonstrate the scientific and technical feasibility of fusion power. Fusion is the energy source of the sun and the stars. Fusion research aims to demonstrate that this energy source can be used to produce electricity on Earth in a safe and environmentally benign manner, providing abundant fuel resources to meet the needs of a growing world population.

ITU - The International Telecommunication Union is the major standardisation body for ICT issues.

LHC - Large Hadron Collider is a major research infrastructure facility located at CERN in Geneva, Switzerland.

NaaS – Network as a Service

NIST - National Institute of Standards and Technology

NRENs - National Research and Education Networks are the entities responsible for procuring and operating the national network and corresponding services dedicated to the research and academic communities. NRENs are the main building blocks of GÉANT.

OGF - Open Grid Forum is an open community committed to driving the rapid evolution and adoption of applied distributed computing.

PaaS - Platform as a Service

PaNData ODI – PaNdata Open Data Infrastructure (FP7) will develop, deploy and operate an Open Data Infrastructure for the European Photon and Neutron laboratories. This will enhance all research done in the neutron and photon communities by making scientific data accessible allowing experiments to be carried out jointly in several laboratories.

PID - persistent identifiers

PII - Personally Identifiable Information

PLATON - Service Platform for e-Science

p-Medicine project - From data sharing and integration via VPH models to personalised medicine

PRACE - Partnership for Advance Computing in Europe is an initiative aiming to implement 3-5 petaflop supercomputing systems in Europe. PRACE manages extreme computing power and a select set of highly specialised services.

Repository is a storage place for digital resources. Users can easily search, access and use resources collected in a repository via an online network. A digital library is one type of repository.

RI is the common abbreviation for research infrastructure.

RTF - Reliable File Transfer

Saas - Software as a Service

SLA - Service-Level Agreement

SSH - "Secure SHell" is a protocol for securely accessing one computer from another. Despite the name, SSH allows you to run command line and graphical programs, transfer files, and even create secure virtual private networks over the Internet.

SSH - Social Sciences and Humanities

TERENA - The Trans-European Research and Education Networking Association is the association of NRENs. TERENA offers a forum in which to collaborate, innovate and share knowledge that fosters the development of Internet technology, infrastructure and services to be used by the research and education community.

VPH - Virtual Physiological Human

WLCG - Worldwide LHC Computing Grid