PARADE Partnership for Accessing Data in Europe

Strategy for a European Data Infrastructure

WHITE PAPER
28.09.2009

Editors

Kimmo Koski, CSC
Claudio Gheller, CINECA
Stefan Heinzel, RZG
Alison Kennedy, EPCC
Achim Streit, FZJ
Peter Wittenburg, CLARIN

Executive summary

The European collaboration in creating data infrastructures is the key enabler for state-of-the-art research. Novel disruptive innovations and scientific breakthroughs are often born when fragments of data are unified unconventionally. Accessibility and interoperability of various data repositories across geographical and disciplinary borders significantly impact the competitiveness of the European research. Data can be equated with money that has value only if it is used and circulated. As the different currencies can be stored in the globally interrelated bank infrastructures, we need persistent, highly available and compatible data infrastructures where data from various disciplines can be stored and fetched from.

Although several progressive and collaborative data initiatives, launched, e.g., by the ESFRI roadmap, are under way, the field remains compartmentalised. This hinders adjacent user communities and data service providers to learn from each other and fully benefit from the best practices. The ultimate consequences are financial due to overlapping work.

This White Paper suggests a European strategy for data related services and outlines a persistent, multidisciplinary European Data Infrastructure, based on the needs of user communities. Seamless collaboration between the user communities, data service providers, various ongoing data initiatives, industrial partners, and financial bodies is the prerequisite for a cost-effective and obliging data infrastructure. We propose a governance structure where the user communities, data service providers and funding bodies work closely together, and we present a model for roles and responsibilities of the communities and service providers. Furthermore, the congruent and consolidative interests with academic and industrial stakeholders are identified and discussed.

Building of mutual trust among all stakeholders is of utmost importance to realise the European data infrastructure strategy. The following objectives are essential:

- Develop and improve common data services
- Establish and nurture a close collaboration among user communities in a data service infrastructure
- Establish a common persistent European data service infrastructure, taking into account the requirements of the different user communities.

Sustainable national funding is fundamental since data typically preserves far beyond the technology cycles. The impacts from a permanent European Data Infrastructure include:

- Strengthening European competitiveness as the infrastructure enables lateral research and innovation
- Increasing the usability of data from high cost or long-lasting experiments, computational simulations, observations and cultural repositories
- Enhancing world-wide interoperability in science and research by establishing a leading common European data services infrastructure
- Defining standards or de-facto standards for data service providers (exchange of data, communication between service providers, meta data structure)
- Giving an excellent example of how science can meet the industry expectations.

This White Paper has been produced in co-ordination with the members of the data initiative PARADE (the Partnership for Accessing Data in Europe).

Contents

1. Introduction

The research paradigm in almost all disciplines has shifted exceedingly to data-driven methods and research questions. Research data is no longer used merely to verify a pre-defined hypothesis. Large or unconventionally combined datasets may, *per se*, suggest new hypotheses and create novel research. The evolving computational models and the growing computer power boost this development by enabling faster analysis of large datasets, which again are material for new research.

Currently, huge amounts of scientific data are stored in isolated local repositories, or even in researchers' desk-top computers. This poses a difficult dilemma, since the data accessibility is crucial for all research, regardless the focus and scale. A researcher may just want to quickly browse data to judge the potential of a spontaneous idea. On the other hand, fundamental global challenges, such as improving health conditions or supporting sustainable development under the pressure of the environmental changes, are dependent on timely access to various and often unconnected data repositories. Thus, the problem lies not only in the accessibility of data, but also in the interconnectivity and interoperability of these resources.

The third important dimension is the curation and preservation of these data sets. Once created, they are part of the scientific knowledge base. Proper preservation exceeds the data life cycle and saves costs as the same data set need not be created twice. Finally, the sheer volume of data complicates the challenge: how to manage data repositories of Petabyte scale?

This White Paper discusses a coherent strategy in dealing with data on a large scale. In contrast to current data repositories that are either geographically restricted or limited to specific disciplines, it suggests a sustainable, seamlessly integrated, physical data infrastructure in the pan-European scale, with common tools and the best practices that serve multiple user communities. Increased co-operation between various research communities, but also with industry and governmental institutions, that is facilitated by compatible data services and facilities, can make a significant contribution to realising European research potential.

1.1. Data oriented research

Modern research communities are faced with increasingly large amounts of data that stem from several sources including data from new scientific instruments, from simulations, from observations and growing volumes of data from the conversion of library resources.

The following rough estimations provide a useful baseline for the scale of data we are talking about in this White Paper:

- Annual production of refereed journal literature (~20k journals; ~2M articles): 1 Terabytes
- The Internet Archive (From 1996 to 2009): 3 Petabytes
- YouTube videos (in 2006): 600 Terabytes
- U.S. Library of Congress, available storage: 264 Exabytes
- Video monitoring systems: 50 Terabytes for a medium size city (200 cameras)
- Support for hospitals (digital libraries, video): 5 Terabytes / hospital
- Digitalisation of national cultural heritage (National Digital Libraries): Several Petabytes
- Annual production of information: 800 Exabytes.

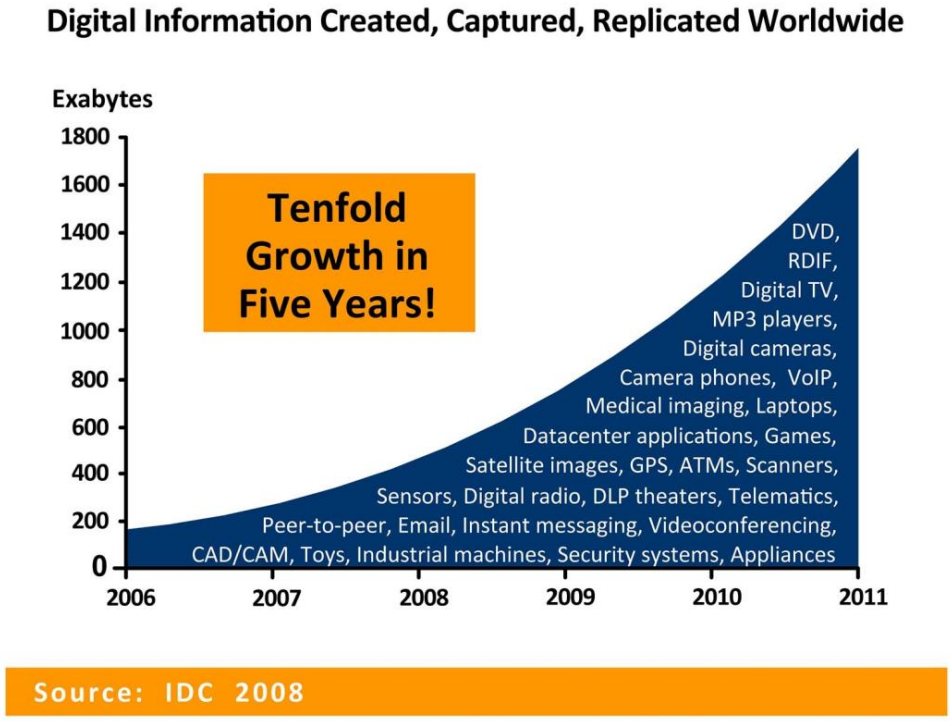


Figure 1: Digital information created worldwide.

Examples from different disciplines are listed below to demonstrate the increasing role of data. It is important to notice that they are just a few examples and do not represent any priorities. It is also notable that different disciplines can benefit from the same data sources (for example, statistical sciences analyse samples of medical data), given they are provided with a proper access. "Recycling" the data sources can considerably save costs and open new perspectives to science. To get a more accurate estimation about the data volumes, it is necessary to address the major data producers.

Bioinformatics

The EBI in the UK is one of the three primary sites for deposition of nucleotide sequence data. A new entry for this database is received every 10 seconds and data at three centres – in the USA, the UK and Japan – is synchronized every 24 hours. The EMBL database has tripled in size in the last 11 months.

Environmental science

The volume of data generated in environmental science is projected to increase dramatically during the next few years. This pattern is mirrored in the USA and elsewhere. Taking only one agency, NASA, we see the rises of data volumes of more than 10 fold in the 5 year period from 2000 to 2005 (Figure 2). The NASA EOSDIS data already exceeds 6 Petabytes.

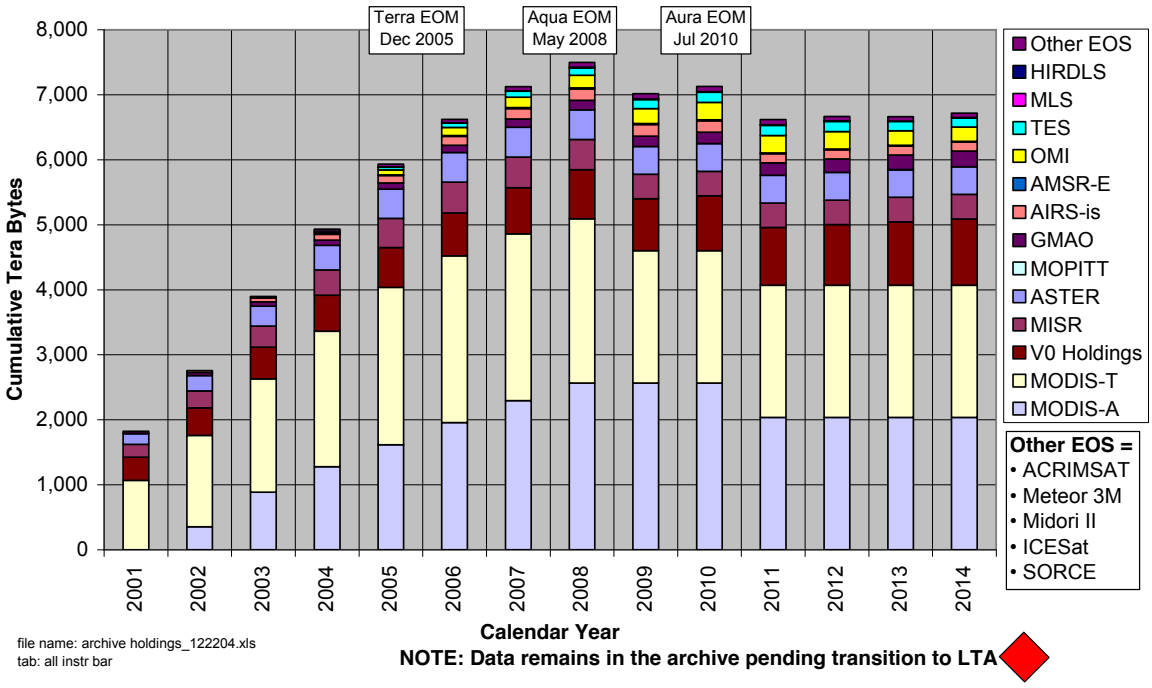


Figure 2: EOSDIS Data production evolution by element. Source: Glenn Iona, EOSDIS Element Evolution. Technical Working Group January 6-7, 2005.

Particle physics

The BaBar experiment has created what is currently the world's largest database: this is 350 Terabytes of scientific data stored in an Objectivity database. In the next few years, these numbers will be greatly exceeded when the Large Hadron Collider (LHC) at CERN in Geneva begins to generate collision data around 16 Petabytes per year. By 2015, particle physicists will be using Exabytes of storage and Petaflop/s of (non-Supercomputer) computation.

Social sciences

In the UK, the total storage requirement for the social sciences has grown from around 400 Gigabytes in 1995 to tens of Terabytes in 2008. The key issue in social sciences is not, however, the size but the accessibility. Information about societal trends is of pivotal importance in tackling challenges, such as the ageing population or in helping to define policy at a European level.

1.2. Need for data infrastructures in Europe

The challenges associated with data services on a large scale are a global issue. Noticeably in the USA and Japan, where the importance of a data services infrastructure has been understood, multiple government supported data initiatives have been launched. The global collaboration in research infrastructures, for example in particle physics or radio astronomy, shares data between researchers of different continents.

A number of new research infrastructures are being prepared in Europe due to the ESFRI roadmap, first published in autumn 2006 and updated in December 2008. The first ESFRI roadmap launched preparatory projects for over 30 new research infrastructures, construction costs varying from 10 MEUR to over 1 BEUR per infrastructure. In the updated ESFRI roadmap, 44 projects are listed. It is important to remember that the ESFRI roadmap includes

planned new infrastructures, while several existing European research infrastructures are renewing their capacity. It is estimated that there are a total of 150-200 European research infrastructures of variable size.

Even though these research infrastructures have different disciplinary ambitions, they also have a lot in common. All of them require a high quality supporting e-infrastructure in one form or another. Some of them need predominantly high performance computing or grid resources, while others also need application development, and all certainly require data services and networking capabilities. There is an enormous potential for collaboration synergy in e-infrastructure level of the existing and new research infrastructures – and at the same time, a major risk for overlapping work being done in several locations independently. This risk is not only monetary, but separate infrastructures can also prevent scientific collaboration and ultimately, slow down the European research and competitiveness. Figure 3 illustrates the horizontal e-infrastructure, combined with a number of disciplines.

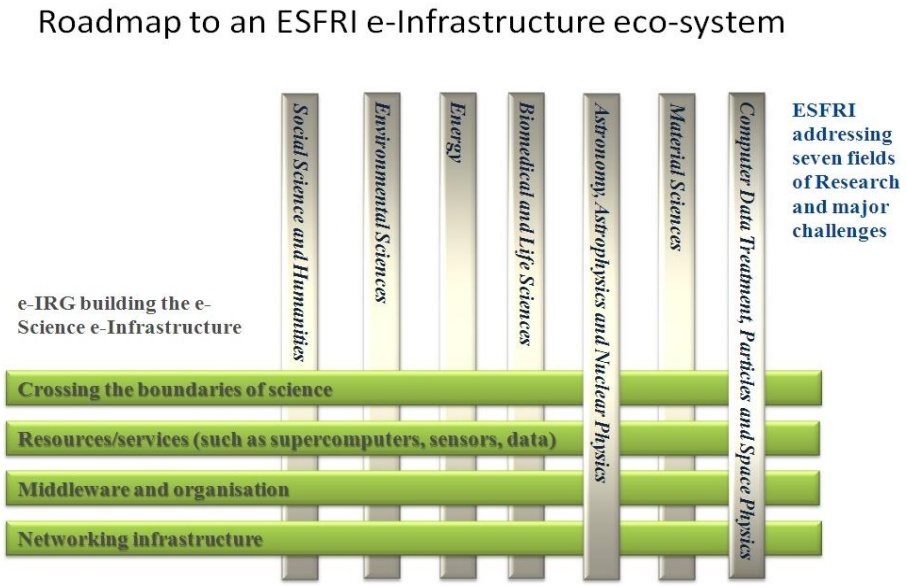


Figure 3: Linkages between research communities and e-infrastructure. Source: e-IRG.

Today, Europe benefits from efficient network collaboration, provided by GEANT. Active co-operation also labels several European HPC and grid computing projects. The emerging needs for data infrastructures and related services have been addressed by a small number of relatively new projects, derived from two data infrastructure calls of EU FP7 (see Appendix 1). In addition, a few forums to advance the data collaboration have been established. It is, however, obvious that due to the exponential growth of research data and the complexity of data management, the current activities are not sufficient. Figure 4 highlights the position of data management within the e-infrastructure.

The major factors that will influence the design of any future European data infrastructure are the requirements of the user communities. Where possible, synergies should be sought with focus on the national and regional partners to build a joint pan-European data domain.

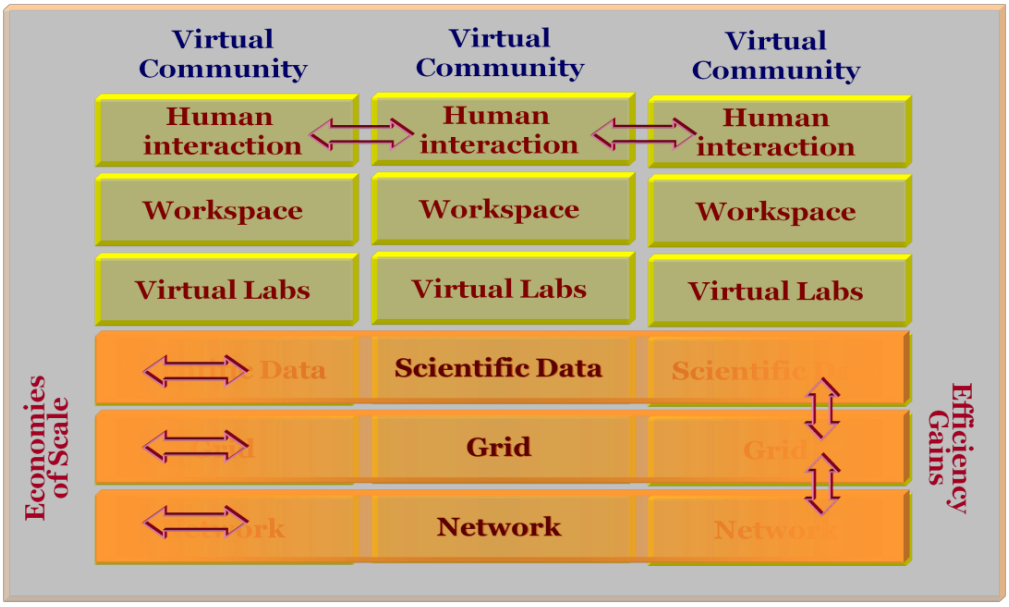


Figure 4: Scientific data as a horizontal service. Source: EC.

1.3. The target audience

This White Paper addresses several stakeholder groups. The importance of joint activities and strategic planning to develop data infrastructures and related services is highlighted. The message should reach governments, research councils and other funding bodies, and political decision makers.

1.4. The contents of the White Paper

The purpose of this White Paper is to contribute to the European data management strategy. It supports the upcoming strategic plans of the ESFRI and e-IRG data task forces, and suggests a permanent, multidisciplinary European Data Infrastructure. The following aspects are highlighted:

1. The definition of the scientific case, proposed by the user communities, is the core of the strategy work and it determines the motivation for further development.
2. The proposal of the services must address the requirements of the scientific case.
3. The identification of the stakeholders in the data domain, both in science and culture, is required for investigating further collaboration possibilities and potential overlaps in Europe.
4. Proactive communication of the policy efforts will help in understanding the targets and interests of different countries and groups.
5. The summary of the core development needs should be produced for data services to approach an efficient European Data Infrastructure.

The following objectives are set to support this work:

- Identify the data needs of user communities which make the major impact for their research.

  Common data needs for multiple communities and community-specific requirements both have a key role. It is important to identify these two different parts to be able to provide

the common services together with a target of cost-efficiency through increased volume and synergy, and still assist the communities in their specific areas.

- Provide strong 'horizontal' data services aiming at building a European Data Infrastructure that supports, e.g., the ESFRI roadmap.

  Providing the horizontal data services that span the various research infrastructures and national borders requires understanding of user needs and collaboration attitude among all stakeholders. Utilization of existing infrastructures, selective upgrades and pan-European resource sharing are among the key opportunities to reach the goals.

- Stimulate the European collaboration.

  Trust and collaboration need to be nurtured in various levels: between research communities and partners providing ICT services as well as between different research communities themselves. Bringing the people from different backgrounds together promotes service innovation in the data domain.

- Increase the commitment through policy impact.

  To impact European decision making through building a discussion forum between user communities, e-infrastructure providers, EC member countries and European commission is a challenging task. If successful, it can make a major improvement. Links between different stakeholders, from science to culture, will bring the data users closer together and allow the sharing of the best practices in a wide scope. Bridging the policy work of the different expert groups, specialized to data, can further stimulate European collaboration.

- Increase visibility of the data development needs in Europe.

  The impact of data on research is immense, which should be made aware to the decision makers including governments and other funding organisations. Due to the complexity of the topic, this task is not easy. Thus, major dissemination effort is required from the European data community at large: both the research and e-infrastructure side of the data domain.

- Show the possible way of co-operation with other communities.

  The Data Infrastructure and its services form the most promising technology that may define good use cases and best practices of collaboration between the scientific community, administration and governmental institutions and science and industry.

- Work out the necessary standards or de facto standards between different service providers.

  The European effort should be directed also towards a unified way of exchanging data, common access interfaces, similar metadata structure, and data mining procedures.

2. Scientific case for data preservation

Increasingly, research disciplines understand that they need to organise themselves into structured communities to become acquainted with e-research practices based on a true cyberspace scenario. As structured communities, they can formulate the key pillars of discipline-specific data infrastructures. Furthermore, they can set requirements for a generic e-infrastructure, offering data services that they cannot or should not realise themselves. Although the exact borderline between the generic and discipline-specific infrastructure aspects has not yet been settled, we can assume that long-term preservation of research data and related services call for cross-disciplinary efforts, since the mechanisms needed are predominantly discipline-independent. In this chapter, we will briefly describe the necessity of a structured community as partner for data services, the requirements that these communities are typically raising, and finally, the functions that a data services federation needs to offer.

Some communities, for example in the humanities and social sciences, have a long history of unraveling the challenges of long-term data preservation and of executing services on such data. The volume of data that they, in general, need to manage and preserve has been relatively small until now. Also the re-usage of such data by other researchers and communities has been very limited. However, the humanities are currently facing a radical change deriving from two sources: a) the need to store high-resolution, lossless media information amounting easily to petabytes and b) an enormous increase in the internal and external complexity of the data. For these reasons, the value of digital humanities and societal data to research is better acknowledged. This emphasizes the need of long-term preservation and persistent access methods to such data from a wide range of research communities.

In domains that practise large scale experimentation or simulation, the volume of generated data is so great and the costs of repeating the experiments so high, that the future availability of raw data for analyses has become ultimately important. Data may need to be retained indefinitely, until scientists have the capabilities to fully analyse and interpret the meaning of anomalous data patterns.

Across Europe, there is a growing number of structured communities in different research domains who are facing similar challenges with regard to data preservation services. These communities have overlapping interests and objectives, but often no history of cross-disciplinary collaboration. In the following section, we explain how a data service infrastructure, comprising a network of computing and storage centres can offer generic data services and support both communities and discipline-oriented research infrastructures.

2.1. Structured communities

Advanced data management of research data is not the primary focus of researchers, although in many disciplines, access to "old" data is a prerequisite for respected publications. According to W. Spek (Alliance for Permanent Access), 30% of requests for earth observation data addresses old data that is used for comparison, and this number tends to increase. In the humanities, the relevance of "old" data is even bigger. Nevertheless, researchers, in general, cannot be responsible for proper data management, given the increasing amount and complexity of data. Researchers need to rely on a backbone of community centres, which take over tasks related to data management, data access and data curation, based on a trust relationship. These community centres have to organise their services in such a way that for a given identifier, the researchers get the exact resource or resource fragment or the expected service. These centres take care that data is described by machine-readable metadata and identified by persistent identifiers. All these are aspects, which the individual researchers normally will not be able to set up.

It is not surprising that many of the distributed research infrastructures that have been launched by the ESFRI process are talking of a new type of network consisting of strong community centres[1] and providing the necessary services in a reliable and stable manner.

These community centres need to focus on discipline related data services, while having expertise on how to contact an Appication Programming Interface (API) of a Persistent Identifier (PID)[2] service of a generic service layer for example. But these community centres can not generally offer long-term preservation services in a cost-efficient way. At bit-stream level, the preservation service is typically discipline-independent and not requiring discipline-specific expertise. What the centres of a discipline-oriented infrastructure are expecting is a data service infrastructure that offers:

- A trusted domain for long-term data preservation accompanied with related data services
- A trusted domain that can be used for compute-intensive services to process the stored data.

2.2. Requirements

The future European Data Infrastructure will be required to fill the gaps, which currently exist in the service landscape. The goal of such a data infrastructure is to respond to the requirements of the various discipline-oriented initiatives and research communities by providing a reliable and stable infrastructure on 24h/7d basis. Figure 5 outlines the wishes of nine different communities or initiatives, as they were presented at a recent meeting of PARADE partners. The collected wishes can certainly be extended to other communities.

	CLARIN	LIFE-WATCH	DIFRE	ELIXIR	INCF	Climate	Space	STFC	EDIFIS
Long-term data archiving, authenticity control	👍	👍	👍	👍	👍	👍	👍	👍	👍
Discovery, access, data mining, virtual integration, curation	👍	👍	👍	👍	👍		👍	👍	👍
Data Processing and workflows	👍			👍	👍	👍	👍		
Data federation	👍		👍	👍	👍	👍	👍	👍	👍
High availability, high reliability	👍	👍					👍	👍	
Authentication,Authorization, Accounting	👍	👍	👍			👍	👍	👍	
Persistent identifiers	👍	👍	👍			👍	👍	👍	
(Component) metadata	👍	👍	👍			👍		👍	
SOA, web services	👍	👍	👍	👍	👍	👍			
Interoperability and standards	👍	👍	👍	👍	👍	👍			
Network of domain nodes	👍		👍	👍		👍	👍		

Figure 5: User community requirements.

CLARIN: language resources and technology community (www.clarin.eu); LifeWatch: community studying biodiversity (www.lifewatch.eu); DIFRE: community studying nuclear fusion; ELIXIR: genetic biologist community (www.elixir-europe.org); INCF: neuroinformatics community (www.incf.org); Climate: climate research community (www.ngdc.noaa.gov/wdc/); Space: European space research community (/www.esa.int/esaSC/120377_index_1_m.html#subhead0); STFC: Science & Technology Facilities Council (www.scitech.ac.uk); EDIFIS: European Data Infrastructures for Innovative Science

[1] The terms that are used to denote these community centres and their function may differ slightly.
[2] Persistent Identifier services based on a system, such as the Handle System (www.handle.net).

The table indicates the major requirements for a persistent data service infrastructure. In short, they are:

1. Long-term data preservation including authenticity and other checks that guarantee the data quality
2. Data access and data curation services of various types that are based on the fact that data is already stored on the preservation nodes and that these nodes are also equipped with large computational capacities
3. Some communities explicitly requested the usage of computational capacities, which is extending to grid/cloud services where two aspects were mentioned: (a) high-performance requirements resulting from specific applications and (b) execution of smaller applications that are used by many users in parallel
4. Almost all initiatives speak about data distribution, data federation and data grid solutions which are not only meant for data preservation purposes, but also for access optimization[3].

From the point of view of the business requirements, 24/7 service and high availability must be expected. Data resources must be stored redundantly so that the individual nodes do not need to reach the 100% service availability. Instead, the network as a whole ensures the high availability.

Based on the community requirements detailed above the general functionality required from a future European Data Infrastructure can be defined as follows

- Open deposit, allowing community centres to easily store data
- Bit-stream preservation, i.e., ensuring that data authenticity will be guaranteed for specified number of years
- Format and content migration, executing CPU intensive transformations on large data sets on the command of the communities
- PID service, allowing centres to register a huge amount of persistent identifiers with the required granularity and data extensions and resolve them
- Metadata support to allow effective management
- Maintaining proper access rights as the basis of all trust
- A variety of access and curation services that will vary between disciplines and over time
- Execution services that allow a large group of researchers to operate on the stored date
- Security to ensure that only partners of the trust federation will have access
- High availability of service so that researchers can rely on them
- Regular quality assessment to ensure adherence to all agreements
- Authentication should be supported by distributed AAI mechanisms enabling single identity and single sign-on
- The principles of Services Oriented Architecture should be followed
- Achieving a higher degree of interoperability at format and semantic level is one of the major goals of most of the initiatives.

[3] Yet this aspect is not completely clear, since moving large amounts of data through our networks will block them.

3. Services provision

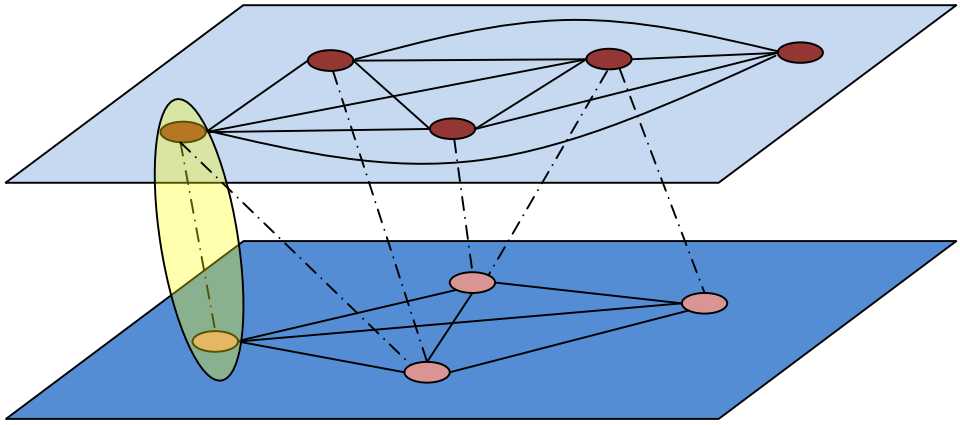
The services to be provided by a European Data Infrastructure are defined by carefully considering the requirements of the user communities. Furthermore, the structure of these services takes advantage of the commonalities between the community requirements.

By considering which services are required by multiple communities, a common set of core services is identified as the basis for the data service infrastructure. Services that lie outside this common set are considered community specific. Since the user community needs will vary over time, the data service infrastructure needs to exhibit a flexible attitude.

Consequently, the proposed data service infrastructure can be viewed as a two layer system, as shown in Figure 6. The underlying layer of common core services is seen as the data infrastructure layer. This infrastructure layer is formed from standard services and it provides a stable basis upon which the user communities can build on. The second layer, the community infrastructure layer, is formed by community specific services. Each supported community is free to deploy and manage their services within this layer. This provides the user communities with a great deal of flexibility regarding their services while they can simultaneously rely on the stable base of the generic data infrastructure layer.

The boundaries between the data service and community layers are dynamic since it is possible that some services will migrate between the layers even during the lifetime of the project, but certainly, in its operational phase. The types of data services, their level of support and stability should, however, be seen as more stable in the infrastructure layer than in the community layer.

The community layer formed from community specific services



The Infrastructure layer consisting of common data services

Figure 6: The two layer model showing the upper community layer and lower infrastructure layer.

3.1. Responsibilities

The two major actors within the data infrastructure are the Communities themselves, in the form of Community Service Providers (CSPs), and the Data Service Providers (DSPs). Defining the responsibilities of these two actors is simplified when considering the two layer model. Some overlap still remains.

Data Services Providers

The Data Services Providers are responsible for the deployment, maintenance and management of the data infrastructure. They provide and manage the underlying storage capacity and adjacent technology. This ensures that migrations of the storage hardware etc. will be transparent to the supported communities and allows the communities themselves to concentrate on innovative research rather than the intricacies of various aspects of data management.

This dual mode ensures that the underlying common services are equally supported for all communities. It is also envisaged that some support will be needed for the higher level, community specific services. In addition, associated members may ask access to certain services while their data is stored outside of the standard infrastructure. Allowing such use cases ensures a high degree of flexibility and eases the integration of further communities.

Communities

The communities are responsible for the deployment and management of the community specific services. These services may be located at one of the Community Service Providers or at the Data Services Providers.

The management of access control and resource provisioning is also covered by the communities, thus allowing the communities themselves to best manage their allocated resources. However, mechanisms need to be established that not only replicate resources, but guarantee that access permissions are maintained for all copies.

Services and tools developed within the individual communities may be incorporated into the common data services infrastructure when they are of use to other communities or when they turn out to be used by a large group of researchers.

3.2. Structure

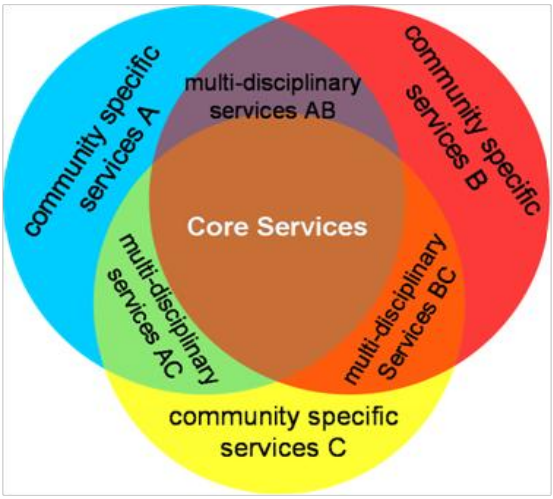
The overall infrastructure will consist of several types of centres, each of which can perform one or more roles. The major types are as follows:

1) *Community Service Providers* – community services with little support levels for core data services such as preservation
2) *Data Service Providers* – core data services and "outsourced" community services with high support level
3) *External sites* – a variety of service centres of different kinds offering data for researchers, such as libraries and archives or even industry offering cost effective storage services, for example.

The heterogeneous nature of the Data Services Providers means that some centres will be better suited to provide certain services and support certain communities. Considerations about the centre's ability to provide, i.e., tape based storage and archiving, the proximity of large computing facilities for data analysis, and the relation to the communities themselves will be taken into account when defining which centres provide which services.

By matching the services and communities rather than attempting to force all supporting sites to provide all services, a more intelligent use of the available resources is achieved. The proposed structure can be seen in Figure 7. Core services are provided to all communities while higher level services are provided at either a community specific or inter-community level.

A high degree of redundancy is achieved by ensuring that several service providers share the responsibility for providing each core service.



This diagram indicates the service landscape that is needed to be established. Community specific services will primarily be offered by community centres. Shared services and services requiring core IT expertise and a high level of persistence and high availability will be offered by data service centres.

Figure 7: Service landscape showing the core and community regions.

4. Targeting an European strategy for data

4.1. Research driving the development of services

As indicated, any e-infrastructure needs to be designed in close connection with the research communities. This is in particular true for the development of data related services that are relevant for almost all research disciplines. Efficient collaboration, in which the needs are demonstrated by qualified research and implemented by high quality e-infrastructure experts, can result in increased synergy between the services offered to different user communities. The goals of such collaboration include the establishment of a trust relationship, state-of-the-art quality and cost-efficiency in the utilization of data infrastructures, and increased development effort due to sharing of the effort of competent people.

The focus in developing the data infrastructures needs to be on the researchers' needs. Building up the trust between research and e-infrastructure providers is of utmost importance in order to be successful. Efficient collaboration is needed not only between research and IT experts, but also between the different disciplines, who can share part of their data related work. By investing in building a solid trust between all the stakeholders, it is possible to create a multidisciplinary community targeting together to solve the data related challenges – something that is currently missing in the European e-Infrastructure domain.

Part of the required services are common to most disciplines, such as distributed authentication systems, assigning persistent identifiers (PID) and search functions, to name a few. Some services might be used only by part of the user communities, but they still could use synergy. Part of the data services are discipline-specific and cannot be shared between other communities. In these cases, the decision whether these services should be implemented by research community centres or people providing specialised data services needs to be made. Figure 8 illustrates the different coverage of the services.

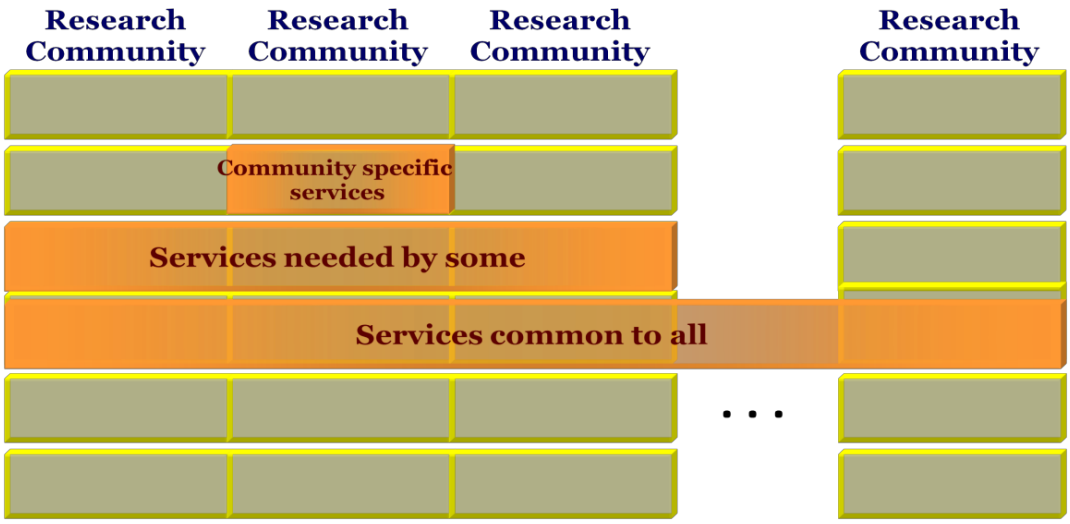


Figure 8: The coverage of the services varies. Collaboration between research communities and e-infrastructure providers can be achieved in each case.

4.2. Why a European strategy is needed?

Today, a number of activities aim to solve the data related challenges indicated in the previous chapters. In some areas, solutions are available while in others managing and accessing data remains to be the most severe bottleneck to overcome. Europe has large

disciplinary centres for some research areas, such as EBI or CERN, who have long experience in managing data, but still need major investments to address their growing requirements. The same applies to many other stakeholders, but often with less experience available. National HPC centres play a role in providing part of the experience and infrastructure, as also do libraries, archives and other authorities dealing with non-scientific data[4]. The efforts are often scattered, serious gaps in data service provisioning can be recognised and overlapping work is definitely done in many areas. Thus, there exists an obvious need for a European level strategy to align the activities for data services and to promote collaboration.

Another urgent incentive for a European data strategy effort is the preparation of new ESFRI research infrastructures, all of which need data related services. Large experimental facilities are currently being planned and partly built, and avoidance of the creation of many incompatible data environments is crucial. Activities gathering together the major research communities with their requirements and e-infrastructure providers with multidisciplinary data services need to be started immediately.

Several stakeholders have noticed the need for drafting a strategy for data services. It is pivotal to bring the different stakeholders together and link the strategy efforts together to target a joint European strategy.

[4] For some disciplines in particular in the humanities libraries and archives are storing the data required for their research of course.

5. Roles of the different stakeholders

5.1. Motivation to work together

The growing importance of data infrastructures drive a number of stakeholders to look for collaboration. They include:

- User communities
- Service providers such as national or international centres
- Consortia of national infrastructure (like NRENs, NGIs)
- Governments and research councils
- European Union
- European projects working with e-infrastructure
- Policy groups, such as e-IRG and ESFRI
- Several working groups and taskforces, e.g, in TERENA
- Industry, both as a user, vendor and service provider
- Standardisation bodies, e.g., NESSI, OGF and OGS.

The roles of different actors need to be defined clearly and smooth and constant interaction between these roles must be ensured.

The policy and strategy work related to data infrastructures is currently being organised by multiple consortia and joint efforts. Resources available for implementation are, however, limited, suggesting that the synchronization of the activities is needed to achieve sufficient impact. Strong consortia and organisations with solid base are required for provision of sustainable services.

National funding organisations have a key role – most of the data related development and operations are highly dependent on sustainable national funding. E.g. long-term storage needs, such digital preservation of national heritage, inherently require financial commitments. As a consequence, the targets of national stakeholders, such as local governments, are of utmost importance.

The following sections discuss a few notable stakeholders that complement or support an effort for a future European Data Infrastructure.

5.2. Alliance for Permanent Access (APA) and a European Data Infrastructure

The main **mission of APA** is to develop a shared vision and framework for a sustainable organisational infrastructure for permanent access to scientific information. It describes the following main objectives:

- Aligning and enhancing permanent information infrastructures in Europe
- Building collaboration and relationships between many of the relevant stakeholders
- Joint advocacy and representation of the views on permanent access
- Increasing impact and mass of permanent information infrastructures.

Operationally, APA has organised an interaction process that has been culminating in three annual conferences where also the interaction with communities has been taken up. The most recent APA conference brought together a number of different communities and experts to discuss the question whether we can afford keeping the records of science accessible. Thus, the focus was on costs and business models. Currently, APA is turning discussion into concrete plans for grant applications that will tackle essential problems in maintaining data accessible.

From the community perspective, the approaches and goals of a data infrastructure and APA are overlapping and partially complementary. While a future European Data Infrastructure should aim towards establishing a stable and persistent operational preservation and access services that offer high availability and reliability, APA is more focused on network building at various levels and methodology evaluation. APA also addresses some critical issues, such as how to deal with a cross-disciplinary metadata landscape that exhibits much structural and semantic heterogeneity. New methodologies will affect the functioning of the infrastructure, once they have demonstrated their robustness and may change the roles of the participating actors.

APA is an alliance of strong research organisations and large institutions, in particular from the domain of libraries that to a certain extent address questions which are similar to those which are to be addressed by the data infrastructure described here. With respect to membership, there is some overlap except for the big libraries, which are mostly participating in APA.

Since similar questions are addressed by the two initiatives and since the interests of the participating institutions are overlapping, it must be possible to define a joint roadmap. At the end, the overall goals and agreements need to be formulated and signed by the research organisations, since finally, they need to cover a substantial fraction of the costs when the seed funding from the EC will stop.

5.3. Other initiatives

There are numerous consortia involved in data domain, most of which are presented in Appendix 1.

Most of the initiatives are relevant through their role in standardisation, policy effort or development activities. Tools for long-term preservation are designed, implemented and validated, e.g., by CASPAR. The EUROPEANA project is being carried out by big libraries and archives in Europe to create a joint collection space accessible by one portal. DRIVER targets to generate a joint metadata and search space from contributions from many partners. E-IRG data management task force will draft policies and give recommendations for European data collaboration. Various projects driven by single user communities and related to community specific services develop tools and best practices that could be reused in designing multidisciplinary services. These are just few examples from a long list of potential synergy.

However, collaboration should not be limited to Europe. Especially the USA has been very active in developing national collaboration for data. National Science Foundation (NSF), national libraries, archives and large research agencies, such as NASA, have allocated significant funding to improve services for data, stimulate collaboration between stakeholders and guarantee long-term preservation and data access.

5.4. Increasing interest of policy makers

The importance of data management is highlighted in various strategy groups. Examples of recent policy work in this area, with recommendations to improve the European collaboration and utilisation of e-infrastructure, include:

- ESFRI data taskforce conclusions, to be published in autumn 2009
- e-IRG data taskforce strategy paper, to be published in October 2009
- ERA expert group recommendations for e-infrastructure, to be published in October 2009.

In addition, the importance of data has been confirmed in various national research infrastructure roadmaps and national e-science strategies. The European Union has promoted the data related collaboration in Europe in various ways, for example, through two specific calls for data infrastructures, several calls directed to data repositories, preservation of scientific and other kind of data, development of libraries and archives etc.

Due to the complexity of data management and especially due to the unsolved challenge to guarantee the long-term preservation of data, development in this area has high priorities in different countries, among research communities and the EU. In many cases, for example in gene technology, data itself can be much more valuable than the infrastructure used for preservation and access of it.

The need to manage data is a common challenge for all. Computing capacity or application development is needed in many projects, but data services, such as preserving, accessing and sharing it, is required in all of them. In addition, in many countries the legislation sets preconditions to part of data, for example by defining the minimum time to store it or setting the authorisation levels of the users. Open access policies are developed in some areas, while much of data has still limitations.

5.5. Trans-European access and global collaboration

With increasing multinational participation in research infrastructures, the need to access data becomes increasingly trans-European or even global. Research infrastructures in the ESFRI Roadmap are a good example of this. Another major example is the Large Hadron Collider (LHC) experiment in CERN, in which data is not only shared within Europe, but even distributed globally. Some of the largest telescopes are physically placed in Chile, but data is globally used for research.

The global collaboration with stakeholders from multiple continents in order to share scientific instruments is increasingly common. The costs of a major research infrastructure are high, and there are many parallel preparatory projects or construction activities going on. This results in a need to share costs among multiple countries – overlapping infrastructures cannot be justified easily.

5.6. Commercial and industrial stakeholders

Data and associated services have an outstanding value for industry, commercial and financial activities, as well as for federated national organizations. Such players can be divided into data users and data service providers. The users' requirements and supported services are compatible to what we expect for the research and science community. Major differences are represented by the policies of access and usage of data and services. Property rights, restricted access, clearance levels, though present also for science, are often mandatory for commercial applications. Quality and reliability of services are to be endorsed. Commercial and governmental data services can be considered and supported.

All these aspects must be addressed and governed in such a way that they comply with the data infrastructure design in terms of cost-effectiveness, technological and architectural choices and, even more crucially, in the acceptance and adoption of a "golden trust rule", which drives the interplay between the different actors.

Commercial data service providers are increasingly diffused and popular. Google, with services like "Maps", "Books", "Earth", "Sky" and the not yet available, but already announced, "Science" services, represents a paradigmatic example: huge data repository, enforced security and availability, intelligent search functions, which match users' interests and needs in a fast and effective way. Such kind of services are provided, often with a more specific characterisation, by Amazon, eBay, Microsoft and the other hundreds of minor players in the field of Internet based data services.

Although widely used, all these services face difficulties related to issues like privacy, intellectual property or copyrights, free sharing and exchange of information. Furthermore, their cost is not a negligible concern. All these issues must be considered and faced by a European Data Infrastructure.

With respect to industrial and commercial data producers, a large fraction of their data can be of general interest, both for science and for further commercial exploitation. Many different examples can be provided. Weather forecast services, after a short proprietary period, usually release data to the community that can use it, for instance, for studies on climate changes or natural risk assessment and prevention. Oil companies use advanced seismographic techniques together with sophisticated numerical modelling to map and characterise the subsoil searching for oilfields. This data could provide priceless information to geophysicists and geologists.

In this framework, a major issue is the data-property and non-disclosure clauses, which can forbid the access to part or all of the information and block a free distribution and sharing of the data. A further concern is how to deal with a possible further commercial exploitation of this data.

In conclusion, industrial and governmental stakeholders can be extremely interesting from different aspects for the European data infrastructure. However, they pose a number of problems and issues, which usually are normally not present in the scientific and research community. This requires a specific analysis and feasibility study to be undertaken, solving any possible legal concern and ambiguity, before such stakeholders can be fully integrated.

6. Governance

Building a sustainable European Data Infrastructure involves the establishment of effective governance. This chapter suggests a general governance model for such an initiative. An appropriate governance structure is essential to ensure that processes are put in place, during the inception phase, to address decisions concerning of the roles, responsibilities, policies and standards that will direct the setting up of the European Data Infrastructure.

We see three major groups of stakeholders that should be represented:

- The user communities who will determine the core activities to be undertaken
- The service providers who will focus on the technological aspects of the data service infrastructure to meet these user community requirements
- The funding bodies and research agencies who will ensure that the infrastructure operates cost-effectively and can evolve into a persistent infrastructure.

These three stakeholder groups form boards that will work out strategic and policy decisions, but they will not be involved in the daily operations. This will be the responsibility of the Executive Board (EB), a body of well-respected experts drawn from all three stakeholder groups. The EB is fully responsible for the success of the infrastructure as a whole and has ultimate responsibility for all aspects of the implementation of the infrastructure, Due to the size of the intended project, the EB will be supported by a professional office.

In addition, an International Advisory Board, composed of individuals considered visionary in their field, is to be appointed. This board will provide constructive criticism on the development of the infrastructure, its strategic direction and on the quality and relevancy of the services offered and provide advice on the adoption of best practice from any comparable data infrastructures elsewhere in the world.

The proposed structure is outlined in the figure below:

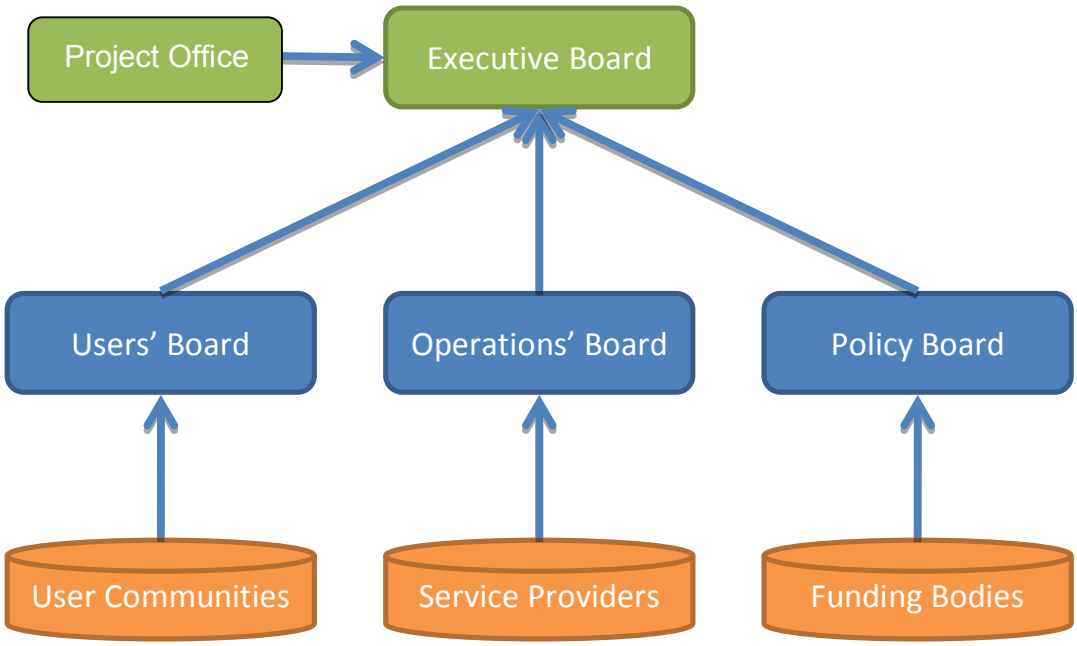


Figure 9: Proposed governance structure.

This structure is designed to:

- Ensure effective management of the Data Infrastructure
- Ensure clearly defined channels of communication, in particular to facilitate user input
- Establish clear procedures for taking decisions and resolving conflicts effectively and quickly
- Ensure the infrastructure operates within its agreed budget and according to administrative, financial and legal principles defined by European and national regulations and by the membership instruments of the Infrastructure itself
- Ensure that the stakeholders (including providers of external services) conform to their obligations
- Manage any intellectual property and liability issues, taking due account of the rights of the stakeholders and data depositors.

7. Conclusions

The competitiveness of European research needs a strong pan-European data service infrastructure that addresses multiple research infrastructures from different disciplines. The prime drivers for the international collaboration in data infrastructures include:

- Better quality of services through a wide collaboration with well-coordinated effort
- Cost efficiency through sharing of infrastructure investment and jointly working for developing and maintaining the services
- Utilization of the best practices developed for data management, accessing and curation in pan-European scope
- Increased security by managing multiple copies in geographically distant locations
- Sustainability of the provided approach
- Multi-disciplinary approach of the provided data infrastructures and data services.

To succeed in this target, it is crucial that different stakeholders collaborate in an open and efficient way. The service development should be clearly driven by the user communities, their needs and priorities. They need to be committed to define ***what*** is being deployed. The service providers, such as national data centres and other partners providing e-infrastructure and related services, are liable to best respond to the demand of the different communities. As some compromises are inevitable (e.g., in cases of conflicting priorities) the centres are responsible to define ***how*** the services can be technically implemented and adapted to the local conditions.

The collaboration between researchers and providers of ICT services has not always been easy. This needs to be changed dramatically. We must not establish just good collaboration between the research and e-infrastructure, but seek for building a mutual **trust** between all stakeholder groups.

The user communities do little with pure storage capacity. They are interested in **complete services** which enable data to be stored, curated, accessed, analysed and updated in a secure and persistent way. Services to support the utilisation of data infrastructures include persistent identifications to globally located data, content migration to adapt with new media, long-term archiving and metadata management, just to name a few examples. A complete service needs to be integrated with other ICT requirements of the user communities.

Linking the strategy work of different stakeholders together and defining a common roadmap is a vital condition for this kind of collaboration. Even though it is not possible to merge the vast area together, different data initiatives should complement each other.

The European Union has funded a number of projects that target in developing pan-European data infrastructures and repositories. The continuation of these services is uncertain as the funding is not necessarily guaranteed after the project ends. Europe needs **sustainable** data services. The key actors are governments that can allocate a sufficient funding to maintain the services beyond the project termination. Continuous funding is pivotal for data infrastructures since data preserves far beyond a typical lifetime of any ICT system. The requirements for data preservation can exceed 50 or 100 years, which calls for a new kind of definition for sustainability.

Europe has built a solid collaboration for research networking through GEANT. Various successful activities in HPC and grid computing have been carried out, and sustainable organisations, such as PRACE for the high-end computing and EGI for the grid collaboration, have been established. However, a same level of collaboration addressing data is still missing in Europe. There is an urgent demand for a strong sustainable organisation to develop multi-disciplinary data services with a target of building an efficient European Data Infrastructure.

Appendix 1: Stakeholders

CASPAR - Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval
www.casparpreserves.eu/

CASPAR intentions relevant to:

- Implement, extend, and validate the OAIS reference model (ISO:14721:2003)
- Enhance the techniques for capturing Representation Information and other preservation related information for content objects
- Design virtualisation services supporting long-term digital resource preservation, despite changes in the underlying computing (hardware and software) and storage systems, and the Designated Communities.
- Integrate digital rights management, authentication, and accreditation as standard features of CASPAR.
- Research more sophisticated access to and use of preserved digital resources including intuitive query and browsing mechanisms

CASPAR thus designs, implements and validates tools for Long-Term Preservation. These CASPAR key component software are distributes distributed with the GNU LGPL license, and can thus be utilized freely

Planets, Preservation and Long-term Access through Networked Services
www.planets-project.eu/

The PLANETS project's relevance lies in its software, tools and services that will help institutions, organisations and individuals to preserve and access their data in the long-term. PLANETS provides: Planets will deliver:

- **Preservation Planning** services that empower organisations to define, evaluate, and execute preservation (Planning Tool Plato)
- Methodologies, tools and services for the **Characterisation** of digital objects (PRONOM 7 Format registry**,** Pdf/a validator**)**
- Innovative solutions for **Preservation Actions** tools which will transform and emulate obsolete digital assets
- An **Interoperability Framework** to seamlessly integrate tools and services in a distributed service network (Fedora Repository Integration**)**
- A **Testbed** to provide a consistent and coherent evidence-base for the objective evaluation of different protocols, tools, services and complete preservation plans

A future data infrastructure should undertake, use and support the usage of these component tools.

DPE, DigitalPreservationEurope
www.digitalpreservationeurope.eu/

The DPE project is very relevant since within the project much knowledge has been gathered about major issues which are core to a European data project, for instance:

- Concepts of trust with regards to digital repositories
- the DRAMBORA toolkit to help design and develop systems and workflows that can help build trusted digital repositories
- Skills needed to undertake a thorough assessment of digital repositories using the DRAMBORA toolkit

SHAMAN, Sustaining Heritage Access through Multivalent ArchiviNg
www.shaman-ip.eu/

SHAMAN's application solution environments for analysing, ingesting, managing, accessing and reusing information objects and data is of interest to a wider audience. The need comes

from the fact that the current approaches to migration do not scale; are not extensible; cannot be automated; cannot support scientific data processes; and cannot guarantee authenticity and integrity. The SHAMAN's preservation architecture preserves the ability to manipulate the original encoding format of a digital entity. This enables logical migration of the data and services

- Multivalent presentation tool
- Client side workflows
- Ingest workflows

SHAMAN solution is built on the Multivalent technology (Java) and the media engine are archived as an iRODS collection.

A European infrastructure will need to manage and automate the preservation of petabytes of information over a 50-100 year period, and the approach needs to be extensible to potentially any data and support provenance of electronic data. SHAMAN provides the way: the *digital entity remains unchanged*, while making it possible to apply new operations.

D4SCIENCE, DIstributed colLaboratories infrastructure on Grid enabled technology 4 science
www.d4science.eu/

D4Science is an e-Infrastructure that provides facilities for creating Virtual Research Environments (VREs) based on shared computational, data and service resources offered by EGEE and DILIGENT at a European level, as well as on data and domain-specific service resources offered by large international organisations. It is offering secure sharing, dynamic allocation of resources (where the concept of resources was extended to content and services in addition to computing and storage) and configurable GUIs for scientific communities. The D4Science infrastructure interoperates sites based on:

- gLite services providing computing and storage grid resources;
- gCube services providing on demand and dynamic Virtual Research Environments.

e-IRG Data management Task Force
www.e-irg.eu/images/stories/dissemination/e-irg_newsletter_1_feb09.pdf

e-IRG discussions and results are strongly relevant for any future European data infrastructure since the initiative describing the pillars of a generic e-Infrastructure landscape for Europe and data services certainly are one of the core pillars. Cose interaction with e-IRG is essential and offer its position papers to the initiative for commenting. In addition we need to refer to the activities of the ESFRI Task Force on Repositories and the e-IRG Task Force on Data Management. Both defined or are in the process of defining recommendations or constraints for activities that will be undertaken by a data infrastructure .

Europeana
www.europeana.eu/portal/

The EUROPEANA project is being carried out by large libraries and archives in Europe to create a joint collection space accessible by one portal. It is certainly a model for the type of collaboration that is required to organise such a large group of institutions.

DRIVER II, Digital Repository Infrastructure Vision for European Research
www.driver-repository.eu/

Also the DRIVER project is of interest because (1) It tries to generate a joint metadata and search space from contributions from many partners. (2) The metadata element mapping and selection solution presented by DRIVER could be useful to more communities.

PARSE.Insight, Permanent Access to the Records of Science in Europe
www.parse-insight.eu/

Parse.Insight is a highly interesting project, since it wanted to get an overview about the needs of researchers with respect to long-term preservation. The results of the survey, the Parse.Insight experts interpretations of the results would be of interest to a future data infrastructure.

EURO-VO-AIDA, The European Virtual Observatory-Astronomical Infrastructure for Data Access
www.euro-vo.org/pub/

EURO VO AIDA provides metadata and resource space where all results are mapped on a unified coordinate system .

GENESI-DR, Ground European Network for Earth Science Interoperations - Digital Repositories
www.genesi-dr.eu/

GENESI-DR will operate, validate and optimise the integrated access and use available data, information, products and knowledge originating from space, airborne and insitu sensors from all digital repositories dispersed all over Europe.

METAFOR, Common Metadata for Climate Modelling Digital Repositories
http://metaforclimate.eu/

Creating a joint metadata domain is a topic would be of interest to a future data infrastructure.only in so far that some core elements relevant for preservation and non-specialist access should be extracted from the provided domain specific metadata descriptions. Proper methods need to be discussed and implemented. DublinCore could be a basic to start with, but results from projects such as METAFOR, ECHO etc need to be studied.

HELIO, Heliophysics Integrated Observatory
www.helio-vo.eu/

HELIO is designed around a Service-oriented Architecture. The initial infrastructure will include services based on metadata and data servers deployed by the European Grid of Solar Observations (EGSO). Processing and storage services will allow the user to explore the data and create the products that meet stringent standards of interoperability. These capabilities will be orchestrated with the data and metadata services using the Taverna workflow tool.

Activities in the US

NARA, U.S. National Archives and Records Administration
http://www.archives.gov/

NARA preserves 1%-3% of all documents and materials created in the course of business conducted by the United States Federal government. Electronic documents are available through the online service also to non-US citizens. Only a small amount of the records is electronic and NARA launch a project called ERA, Electronic Records Archives. It is the National Archives and Records Administration's strategic initiative to preserve and provide long-term access to uniquely valuable electronic records of the U.S. Government, and to transition government-wide management of the lifecycle of all records into the realm of e-government. In 2010, the National Archives intends to make the system available to the public. Ultimately, the Archives expects the system to be able to preserve and provide access to ever-increasing volumes of important electronic records of the Federal government, even long after the hardware and software used to create them has become obsolete.

LOCKSS, Lots Of Copies Keep Stuff Safe
www.lockss.org/lockss/Home

LOKSS is an excellent example for a lean software package that solves the data replication task to a certain extent. However, LOKSS is not the only package that tackled the replication task, iRods is another such package offering storage resource virtualization as one of its core pillars.

NDIIPP, National Digital Information Infrastructure and Preservation Program
http://www.digitalpreservation.gov/

The mission of the NDIIPP is to develop a national strategy to collect, archive and preserve the burgeoning amounts of digital content, especially materials that are created only in digital formats. NDIIPP is based on an understanding that digital stewardship on a national scale depends on public and private communities working together. The Library has built a preservation network of over 130 partners from across the nation to tackle the challenge, and is working with them on a variety of initiatives. The Program focuses on three areas:

- Capturing, preserving, and making available significant digital content.
- Building and strengthening a network of partners
- Developing a technical infrastructure of tools and services

Chronopolis
http://chronopolis.sdsc.edu/.

A key goal of the Chronopolis project is to provide cross-domain collection sharing for long-term preservation. Using existing high-speed educational and research networks and mass-scale storage infrastructure investments, the partnership is designed to leverage the data storage capabilities at SDSC, NCAR, and UMIACS to provide a preservation data grid that emphasizes heterogeneous and highly redundant data storage systems

Blue Ribbon Task Force
http://brtf.sdsc.edu/about.html

The **Blue Ribbon Task Force on Sustainable Digital Preservation and Access** was created in late 2007. During the next two years, the BRTF-SDPA will explore the sustainability challenge with the goal of delivering specific recommendations that are economically viable of use to a broad audience, from individuals to institutions and corporations to cultural heritage centers. Broadly speaking, economic sustainable digital preservation will require new models for channeling resources to preservation activities; efficient organisation that will make these efforts affordable; and recognition by key decision-makers for the need to preserve, with appropriate incentives to spur action.

TeraGrid
http://www.teragrid.org/about/index.html

TeraGrid is an open scientific discovery infrastructure combining leadership class resources at eleven partner sites to create an integrated, persistent computational resource.
Using high-performance network connections, the TeraGrid integrates high-performance computers, data resources and tools, and high-end experimental facilities around the country. Currently, TeraGrid resources include more than a petaflop of computing capability and more than 30 petabytes of online and archival data storage, with rapid access and retrieval over high-performance networks. Researchers can also access more than 100 discipline-specific databases. With this combination of resources, the TeraGrid is the world's largest, most comprehensive distributed cyberinfrastructure for open scientific research.

Appendix 2: List of Acronyms

AAI	Authentication and Authorization Infrastructure
APA	Alliance for Permanent Access
API	Application Programming Interface
Astro-WISE	Astronomical Wide-field Imaging System for Europe
BaBar	A High Energy Physics experiment located at SLAC National Accelerator Laboratory, near Stanford University, in California.
CASPAR	Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval
CLARIN	Common Language Resources and Technology Initiative
CPU	Central Processing Unit
DIFRE	Community studying nuclear fusion
DPE	Digital Preservation Europe
DRIVER	Digital Repository Infrastructure Vision for European Research
EBI	European Bioinformatics Institute
EC	European Commission
ECMWF	European Centre for Medium Range Weather Forecasting
EDIFIS	European Data Infrastructures for Innovative Science
EGEE	Enabling Grids for E-sciencE
EGI	European Grid Initiative
e-IRG	e-Infrastructure Reflection Group
Elixir	European Life Sciences Infrastructure For Biological Information
ENES	European Network for Earth System Modelling
ERA	European Research Area
ESA	European Space Agency
ESFRI	European Strategy Forum on Research Infrastructures
ETSI	European Telecommunications Standards Institute
EUROPEANA	European Digital Library
EURO-VO	European Virtual Observatory
GEANT	European multi-gigabit computer network for research and education
HPC	High Performance Computing
ICT	Information and Communication Technology
INCF	International Neuroinformatics Coordinating Facility
IMPACT	Improving Access to Text
IPR	Intellectual Property Rights
ISO	International Organization for Standardization
LHC	Large Hadron Collider
LifeWatch	E-science and technology infrastructure for biodiversity data and observatories
LIGO	Laser Interferometer Gravitational Observatory
NASA	National Aeronautics and Space Administration
NESSI	Networked European Software and Services Initiative
NGI	National Grid Initiative
NREN	National Research and Education Network
NSF	National Science Foundation
OAIS	Open Archival Information System
OGF	Open Grid Forum
PARADE	Partnership for Accessing Data in Europe
PID	Persistent Identifier
PRACE	Partnership for Advanced Computing in Europe
SOA	Service Oriented Architecture
SSO	Single sign-on
TERENA	Trans-European Research and Education Networking Association
VISTA	Visible and Infrared Survey Telescope for Astronomy

Appendix 3: Organizations contributing to the White Paper

Partnership for Advanced Data in Europe (PARADE) is a consortium targeting to build efficient services addressing data management needs of multiple research communities. PARADE consists of several user communities and national partners, who work together to improve European collaboration in data infrastructures. The work aims at linking with various European initiatives addressing data with an intention to work together to build a pan-European collaboration.

The following organizations have been contributing to this White Paper:

Astro-WISE / EURO-VO	Astronomical Wide-field Imaging System for Europe / The European Virtual Observatory
BSC	Barcelona Supercomputer Center - Centro Nacional de Supercomputacion
CINECA	Consorzio Interuniversitario
CLARIN	Common Language Resources and Technology Infrastructure
CSC	CSC – IT Center for Science Ltd.
DIFRE	Data Initiative for Fusion Research in Europe
ELIXIR	European Life Sciences infrastructure for Biological Information in Europe
EMBL	European Molecular Biology Laboratory
ENES	European Network for Earth System Modelling
EPCC	Edinburgh Parallel Computing Centre
ETHZ	Eidgenössische Technische Hochschule Zürich - Swiss National Supercomputing Centre (CSCS)
HIP	University of Helsinki
JÜLICH	Forschungszentrum Jülich GmbH
Lifewatch	e-Science and Technology Infrastructure for Biodiversity Research
NCF	Stichting Nationale Computer Faciliteiten
PDC, KTH	Parallell Dator Centrum, Den Kungliga Tekniska Högskolan
PSCN	Instytut Chemii Bioorganicznej Pan W Poznaniu
RUG	The Donald Smits Centre of Information Technology of the University of Groningen/ TARGET
RZG	Rechenzentrum Garching of the Max Planck Society and the IPP
SARA	Stichting Academisch Rekencentrum Amsterdam
SNIC	Swedish National Infrastructure for Computing
STFC	Science and Technology Facilities Council
UNINETT	UNINETT Sigma AS