Global Research Data Infrastructures:

The GRDI2020 Vision

This report is compiled by Costantino Thanos, CNR-ISTI within the framework of the GRDI2020 (www.grdi2020.eu) project funded under the 7th Framework Programme, Capacities – GÉANT & eInfrastructures.

DISCLAIMER

GRDI2020 is funded by the European Commission under the 7th Framework Programme (FP7).

The goal of GRDI2020 project, *Towards a 10-year vision for global research data infrastructures*, is to establish a framework for obtaining technological, organisational, and policy recommendations guiding the development of ecosystems of global research data infrastructures. Mobilising user communities, large initiatives, projects, leading experts, and policy makers throughout the world and involving them in GRDI2020 activities will achieve the establishment of this framework.

This document contains information on core activities, findings, and outcomes of GRDI2020. It also contains information from the distinguished experts who are in two external groups – the Advisory Board Members (AB), and the Technological and Organisational Working Groups. Any reference to content in this document should clearly indicate the authors, source, organisation, and date of publication.

The document has been produced with the funding of the European Commission. The content of this publication is the sole responsibility of the GRDI2020 Consortium and its experts, and it cannot be considered to reflect the views of the European Commission.

The European Union (EU) was established in accordance with the Treaty on the European Union (Maastricht). There are currently 27 member states of the European Union. It is based on the European Communities and the member states' cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice, and the Court of Auditors (http://europa.eu.int/).

1. The New Science Paradigm

Some areas of science are currently facing from a hundred – to a thousand-fold increase in volumes of data compared to the volumes generated only a decade ago. This data is coming from satellites, telescopes, high-throughput instruments, sensor networks, accelerators, supercomputers, simulations, and so on [1]. The availability and use of huge datasets presents both new opportunities and at the same time new challenges for scientific research.

Often referred to as a data deluge massive datasets is revolutionizing the way research is carried out and resulting in the emergence of a new fourth paradigm of science based on **data-intensive computing** [2]. New data-dominated science will lead to a new data-centric way of conceptualizing, organizing and carrying out research activities which could lead to a rethinking of new approaches to solve problems that were previously considered extremely hard or, in some cases, even impossible to solve and also lead to serendipitous discoveries.

The new availability of huge amounts of data, along with advanced tools of exploratory data analysis, data mining/machine learning and data visualization, offers a whole new way of understanding the world. One view put forward is that in the new data-rich environment correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all [3].

In order to be able to exploit these huge volumes of data, new techniques and technologies are needed. A new type of e-infrastructure, the **Research Data Infrastructure**, must be developed for harnessing the accumulating data and knowledge produced by the communities of research, optimizing the data movement across scientific disciplines, enabling large increases in multi- and inter- disciplinary science while reducing duplication of effort and resources, and integrating research data with published literature.

To make this happen several breakthroughs must be achieved in the fields of research data modelling, management and tools.

2. Research Data Infrastructures

Research Data Infrastructures can be defined as managed networked environments for digital research data consisting of services and tools that support: (i) the whole research cycle, (ii) the movement of scientific data across scientific disciplines, (iii) the creation of open linked data spaces by connecting data sets from diverse disciplines, (iv) the management of scientific workflows, (v) the interoperation between scientific data and literature, and (vi) an Integrated Science Policy Framework.

Research data infrastructures are not systems in the traditional sense of the term; they are networks that enable locally controlled and maintained digital data and library systems to interoperate more or less seamlessly. Genuine research data infrastructures should be ubiquitous, reliable, and widely shared resources operating on national and transnational scales.

A research data infrastructure should include **organizational practices**, **technical infrastructure** and **social forms** that collectively provide for the smooth operation of collaborative scientific work across multiple geographic locations. All three should be objects of design and engineering; a data infrastructure will fail if any one is ignored [4].

Another school of thought considers (data) infrastructure as a fundamentally **relational concept**. It becomes infrastructure in relation to organized (research) practices [5]. The relational property of (data) infrastructure talks about that which is between – between communities and data/publications collections mediated by services and tools. According to this school of thought the exact sense of the term (data) infrastructure and its "**betweenness"** are both theoretical and empirical questions.

In [6] (data) infrastructure emerges with the following dimensions:

- **Embeddedness**: Infrastructure is "sunk" into, inside of, other structures, social arrangements and technologies
- **Transparency**: Infrastructure is transparent to use, in the sense that it does not have to be reinvented each time or assembled for each task, but invisibly supports those tasks.
- **Reach of scope**: Infrastructure has reach beyond a single event or one-site practice.
- **Learned as part of membership**: The taken-for-grantedness of artifacts and organizational arrangements is a sine qua non of membership in a community of practice. Strangers and outsiders encounter infrastructure as a target object to be learned about. New participants acquire a naturalized familiarity with its objects as they become members.
- **Links with conventions of practice**: Infrastructure both shapes and is shaped by the conventions of a community of practice.
- **Embodiment of standards**: Modified by scope and often by conflicting conventions, infrastructure takes on transparency by plugging into other infrastructures and tools in a standardized fashion.
- **Build on an installed base**: Infrastructure does not grow de novo; it wrestles with the "inertia of the installed base" and inherits strengths and limitations from that base.
- **Becomes visible upon breakdown**: The normally invisible quality of working infrastructure becomes visible when breaks: the server is down, the bridge washes out, there is a power blackout. Even when there are back-up mechanisms or procedures, their existence further highlights the now-visible infrastructure.

Research data infrastructures should be science-and engineering-driven and when **coupled** with **high performance computational systems** increase the overall capacity and scope of scientific research. Optimization for specific applications may be necessary to support the entire research cycle but work in this area is mature in many problem domains.

Science is a global undertaking and research data are both national and global assets. There is a need for a seamless infrastructure to facilitate collaborative arrangements necessary for the intellectual and practical challenges the world faces.

Therefore, there is a need for **global research data infrastructures** able to interconnect the components of a distributed worldwide science ecosystem by overcoming language, policy, methodology, and social barriers. Advances in technology should enable the development of global research data infrastructures that reduce geographic, temporal, social, and National barriers in order to discover, access, and use of data.

Their ultimate goal should be to enable researchers to make the best use of the world's growing wealth of data.

The next generation of global research data infrastructures is facing two main challenges:

- To effectively and efficiently support **data-intensive Science**
- To effectively and efficiently support **multidisciplinary/interdisciplinary Science**

Data-Intensive Science

By **data-intensive science** we mean any scientific research activity whose progress is heavily dependent on careful thought about how to use data. Such research activities are characterized by:

- increasing volumes and sources of data,
- complexity of data and data queries,
- complexity of data processing,
- high dynamicity of data,
- high demand for data,
- complexity of the interaction between researchers and data, and
- importance of data for a large range of end-user tasks.

Fundamentally, data-intensive disciplines face two major challenges [7]:

- Managing and processing exponentially growing data volumes, often arriving in time-sensitive streams from arrays of sensors and instruments, or as the outputs from simulations; and
- Significantly reducing data analysis cycles so that researchers can make timely decisions.

Multidisciplinary – Interdisciplinary Science

By **multidisciplinary** approach to a research problem we mean an approach that draws appropriately from multiple disciplines in order to redefine the problem outside of normal boundaries and reach solutions based on a new understanding of complex situations.

There are several barriers to the multidisciplinary approach of behavioural and technological nature.

Among the major technological barriers we identify those that must be overcome when moving data, information, and knowledge between disciplines. There is the risk of interpreting representations in different ways caused by the loss of the interpretative context. This can lead to a phenomenon called "ontological drift" as the intended meaning becomes distorted as the information object moves across semantic boundaries (semantic distortion) [8].

A relatively similar concept is the **interdisciplinary** approach to a research problem. It involves the connection and integration of expertise belonging to different disciplines for the purpose of solving a common research problem.

Again, the barriers faced by an interdisciplinary approach are of two types: behavioural and technological.

Among the major technological barriers we identify the need for integrating data, information, and knowledge created by different disciplines. In fact, one of the major barriers to be overcome concerns the integration of activities that are taking place on different ontological foundations.

The requirements described above, imposed by data-intensive multidisciplinary-interdisciplinary science are the motivations for building the theoretical foundations of the next generation data infrastructures. To make this happen a considerable number of difficult data, application, system, organizational, and policy challenges must be successfully tackled.

The breakthrough technologies needed to address many of the critical problems in data-intensive multidisciplinary-interdisciplinary computing will come from collaborative efforts involving many domain application disciplines as well as computer science, engineering and mathematics.

3. A Strategic Vision for a Global Research Data Infrastructure

We envision that in the future several **Digital Science Ecosystems** will be established.

We use the ecosystem metaphor in order to conceptualize all the "research relationships " between the components of the science universe.

The traditional notion of an ecosystem in biological sciences describes a habitat for a variety of different species that co-exist, influence each other, and are affected by a variety of external forces. Within the ecosystem, the evolution of one species affects and is affected by the evolution of other species.

We think that a model of digital ecosystem of scientific research allows to have a better understanding of its dynamic nature. We believe that the ecological metaphor for understanding the complex network of data-intensive multidisciplinary research relationships is appropriate as it is reminiscent of the interdependence between species in biological ecosystems. It emphasizes that advances and transformations in scientific disciplines are as much a result of the broader research environment as of simply technological progress.

In the world of science, there are many factors that influence the evolution of a specific scientific discipline. By considering the digital science ecosystem as an interrelated set of data collections, services, tools, computations, technologies and communities of research we can contribute to identifying the factors that impact scientific progress.

Defining the Digital Science Ecosystem

We introduce a digital science ecosystem approach that considers a complex system composed of **Digital Data Libraries**, **Digital Data Archives**, **Digital Research Libraries**, and **Communities of Research** (see figure below).

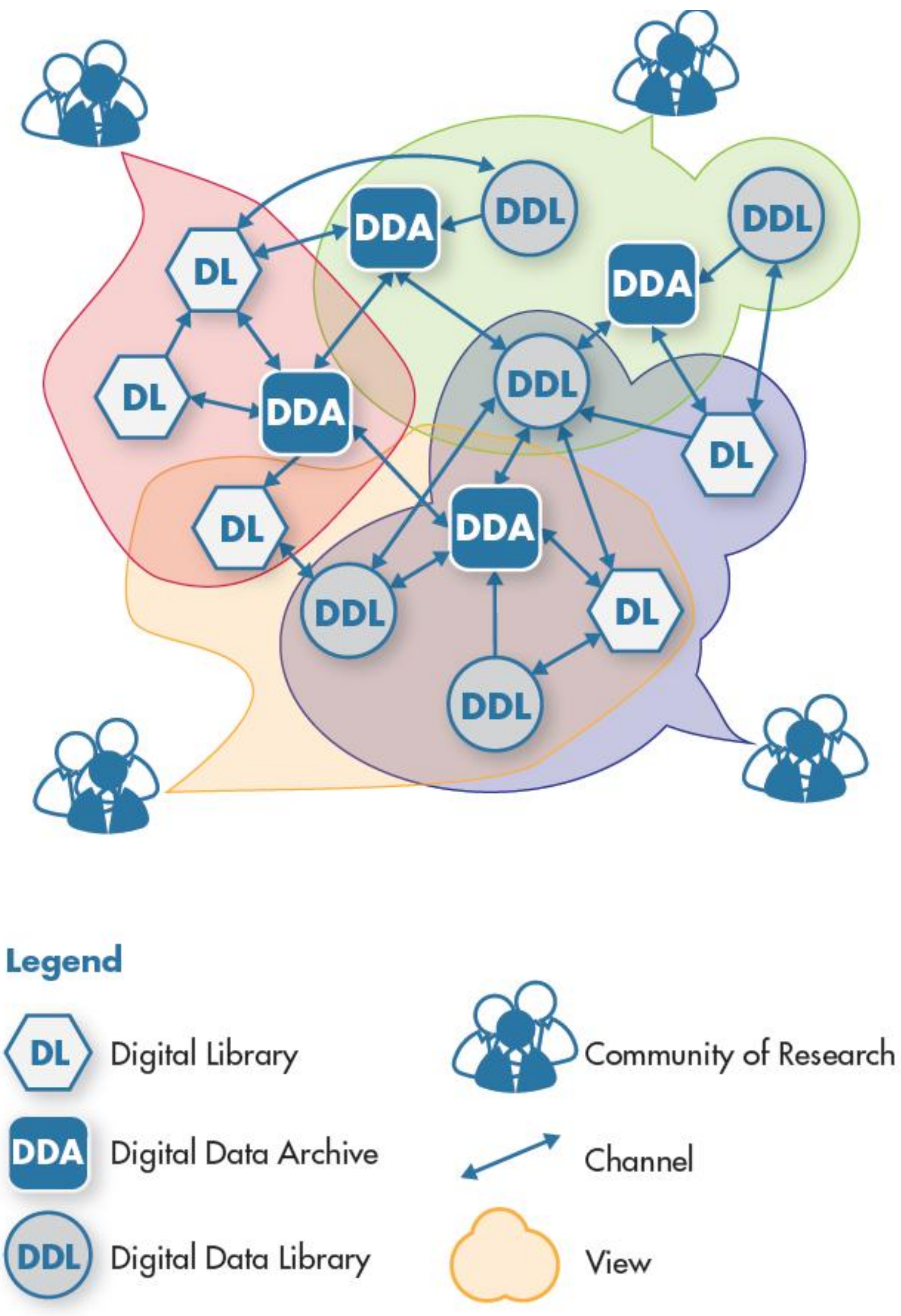

Figure 1 – GRDI2020 Digital Science Ecosystem

Digital Data Libraries

Increasingly, the volumes of data produced by high-throughput instruments and simulations are so large, and the application programs are so complex, that it is far more economical to move the end-user's programs to the data rather than moving the source data and its applications to the user's local system. The volumes of data used in large-scale applications today simply cannot be moved efficiently or economically, even over very high-bandwidth internet links.

From the organizational point of view moving end-user applications to data stores calls for the creation of Service Stations called Digital Data Libraries or Science Data Centers [9].

These should be designed to ensure the long-term stewardship and provision of quality-assessed data and data services to the international science community and other stakeholders. Each Digital Data Library will have responsibilities for curation of datasets and the applications that provide access to them, and employ technical support staff that understand the data and manage the growth, quality and inherent value of the datasets.

Digital Data Libraries fall into one of several categories [10]:

- **Research Digital Data Libraries:** contain the products of one or more focused research projects. Typically, these data require limited processing or curation. They may or may not conform to community standards for file formats, metadata structure, or content access policies. Quite often, applicable standards may be nonexistent or rudimentary because the data types are novel and the size of the user community small. Research data collections may vary greatly in size but are intended to serve a specific group for specific purpose for an identifiable period of time. There may be no intention to preserve the collection beyond the end of a project.
- **Discipline or Community Digital Data Libraries:** serve a single science or engineering disciplinary or scholarly community. These digital data libraries often establish community-level standards either by selecting from among preexisting standards or by bringing the community together to develop new standards where they are absent or inadequate.
- **Reference Digital Data Libraries:** are intended to serve large segments of the scientific and education community. Characteristic features of this category of digital data libraries are their broad scope and diverse set of user communities including scientists, students, and educators from a wide variety of disciplinary, institutional, and geographical settings. In these circumstances, conformance to robust, well-established, and comprehensive standards is essential, and the selection of standards often has the effect of creating a universal standard.
- **Specialized Service Digital Data Libraries**: while the data libraries described above are intended to provide access to data collections and a set of basic services (including collection, curation, provision, short-term preservation, and publishing), another category of data libraries will emerge offering specialized data services, i.e., data analysis, data visualization, massive data mining, etc. These will play a very important role in advancing data-intensive research.

Digital Data Archives

Scientific Data archiving refers to the long-term storage of scientific data and methods, that is, the process of moving data that is no longer actively used to a separate data storage device for long-term preservation. Data archives consist of older data that is still important and necessary for future reference, as well as data that must be retained for regulatory compliance. Provisions for long-term preservation (including means for continuously assessing what to keep and for how long) should be provided.

Data archiving is more important in some fields than others. The requirement of digital data archiving is a recent development in the history of science and has been made possible by advances in information technology allowing large amounts of data to be stored and accessed from central locations.

Data Archives should be indexed and have search capabilities so that files and parts of files can be easily located and retrieved [11].

Data Preservation issues become very important in data archiving. All storage media deteriorates over time, although some more rapidly than others. As systems change, data formats and access methods also change and it is already the case that a considerable amount of digital data cannot be retrieved because the systems and software that created these data no longer exist. Data preservation is an active area of computer science research and its importance will continue to grow as data archives become larger and more numerous.

Digital Research Libraries

A Digital Research Library is a collection of electronic documents. The mission of research libraries is to acquire information, organize it, make it available and preserve it. To meet user needs, the founders of a Digital Research Library must accomplish two general tasks: establishing the repository of electronic scholarly materials, and implementing the tools to use it. More importantly, sustainability models must be established and put into place so that scholarly information in a repository will be available to future generations of researchers. This implies *strong and enduring* organizational commitments, fiscal commitments and institutional commitments [12].

Communities of Research

Science is conducted in a dynamic, evolving landscape of communities of research organized around disciplines, methodologies, model systems, project types, research topics, technologies, theories, etc. These communities facilitate scientific progress and can provide a coherent voice for their constituents, enhancing communication and cooperation and enabling processes for quality control, standards development, and validation [13].

Digital Science Ecosystem Views

A community of research is interested in performing a research activity based on specific set of data and tools. A specific **ecosystem view** can be defined by identifying a community of research and a set of data collections, data services, and data tools necessary for the research activity undertaken by this community. Ecosystem views are materialized by **Science Gateways** or **Virtual Research Environments**.

Digital Science Ecosystem Channels

Research channels can be established across the components of a science ecosystem. The data and information exchanged between these components flow through the ecosystem channels. We classify these eco system channels according to the resulting research results.

Channels enabling Multidisciplinary/interdisciplinary research: research channels across different types of digital data libraries allow scientists belonging to different communities of research and/or to different disciplines to work together.

Channels enabling Data Preservation: channels across digital data libraries and digital data archives allow data together with the appropriate preservation information to move from short-term to long-term preservation states based on well defined provision policies.

Channels enabling Unification of Research Data with Scientific Literature*:* channels across digital data libraries and digital research libraries make it possible to merge scientific data with research literature *resulting in new document models* in which the data and text gain new functionalities. By integrating scientific data and research publications it will become possible for one to read a paper and examine the original data on which the paper's conclusions or claims are based. It will be also be possible to reuse the data and replicate the research and data analysis. Yet another possibility will be to locate all the literature referencing the data [14].

Channels enabling Cooperative research: channels across the members of communities of research allow scientific cooperation and collaboration.

Digital Science Ecosystem Services

Digital ecosystem services are necessary in order to enable researchers to efficiently and effectively carry out research activities. They include data registration, data discovery, data citation, data service/tool discovery, data search, data integration, data sharing, data linking, data transportation, data service transportation, ontology/taxonomy management, workflow management and policy management.

Global Research Data Infrastructures: The GRDI2020 Vision

We envision a Global Research Data Infrastructure as vital to the realization of an open, extensible and evolvable digital science ecosystem. A Global Research Data Infrastructure both creates and sustains a reliable operational digital science ecosystem environment.
Therefore it must support:

- the creation and maintenance of **science ecosystem views** through:
  - **Science Gateways**: A Science Gateway is a community-specific set of tools, applications, and data collections that are integrated together and accessed via a portal or a suite of applications.
  - **Virtual Research Environments:** A Virtual Research Environment (VRE) is a "technological framework", i.e., digital infrastructure and services, that together allow on-demand creation of "virtual working environments" in which "communities of research" can effectively and efficiently conduct their research activities.

- the creation and maintenance of **research channels** across the several components of the ecosystem.

This implies that all the components of the ecosystem are able, along the research channels, to exchange data and information without semantic distortions within a negotiated framework of shared policies. The final result is a highly productive "interoperable science ecosystem" that reduces fragmentation of science due to disparate data and contributes to both reducing geographic fragmentation of datasets and at the same time accelerates the rate at witch data and information can be made available and used to advance science.
In addition, a **mediation** technology capable of reconciling data and language heterogeneities associated with different scientific disciplines must be developed in order to enable the next generation data infrastructures to support ecosystem research paths.

- the creation and maintenance of **Service Environments** that enable the efficient delivery of ecosystem services. They include:
  - **Data Registration Environment**: By Data Registration Environment we mean an environment enabling researchers to make data citable as a unique piece of work and not *only as a part of a* publication. Once accepted for deposit and archived, data is assigned a "Digital Object Identifier" (DOI) for registration. A Digital Object Identifier (DOI) [15] is a unique name (not a location) within a science ecosystem and provides a system for persistent and actionable identification of data. DOIs

could logically be assigned to every single data point in a set; however in practice, the allocation of a DOI is more likely to be to a meaningful set of data. Identifiers should be assigned at the level of granularity appropriate for an envisaged functional use. The Data Registration Environment should be composed of a number of capabilities, including a specified numbering syntax, a resolution service, a data model, and an implementation mechanism determined by policies and procedures for the governance and application of DOIs.

- **Data Discovery Environment**: By Data Discovery Environment we mean an environment enabling researchers to quickly and accurately identify and find data that supports research requirements within the science ecosystem. It should be composed of a number of capabilities and tools that support the pinpointing of the location of relevant data.

- **Data Citation Environment:** By Data Citation Environment we mean an environment enabling researchers to provide a reference to data in the same way as researchers routinely provide a bibliographic reference to printed resources. Data citation is recognized as one of the key practices underpinning the recognition of data as a primary research output rather than as a by-product of research. The essential information provided by a citation links it and the cited data set. Therefore, a Data Citation Environment should support a number of capabilities including a standard for citing data sets that addresses issues of confidentiality, verification, authentication, access, technology changes, existing subfield-specific practices, and possible future extensions, as well as a naming resolution service. Data citation will ensure scholarly recognition and credit attribution.
- **Data Service/Tool Discovery Environment:** By Data Service/Tool Environment we mean an environment enabling the automatic location of data services/tools that fulfill a researcher goal. The Data Service/Tool Environment should support a number of capabilities, including ontology-based descriptions both of the researcher’s goal and the data service/tool functionality as well as a mediation support in case these descriptions use different ontologies.
- **Data Search Environment:** By Data Searching Environment we mean an environment in which researchers can identify, locate and access required data. It should support a number of capabilities that support a complex search process characterized by multiple steps, spanning multiple data sources that may require long-term sessions and continuous refinement of the search process.
- **Data Integration Environment:** By Data Integration Environment we mean an environment enabling researchers to combine data residing at different sources, and provide them with a unified view of these data. The Data Integration Environment should include capabilities and tools that support data transformation, duplicate detection and data fusion.
- **Data Sharing Environment:** By Data Sharing Environment we mean an

environment enabling the sharing of research results among the members of the Communities of Research of the science ecosystem. It should be composed of a number of capabilities and tools that support the contexts for shared data use.

- **Data Linking Environment:** By Data Linking Environment we mean an environment enabling the connection of data sets from diverse domains of the science ecosystem. It should be composed of a number of capabilities and tools that support the creation of common data spaces that allow researchers to navigate along links into related data sets.
- **Ontology/Taxonomy Management Environment:** By Ontology/Taxonomy Management Environment we mean an environment enabling a wide range of semantic science ecosystem data services. Ontologies and taxonomies provide the semantic underpinning enabling intelligent data services including data and service discovery, search, access, integration, sharing and use of research data. This type of Service Environment should contain numerous capabilities including, but not limited to ontology and taxonomy models, ontology and taxonomy metadata, and reasoning engines. These are necessary in order to efficiently create, modify, query, store, maintain, integrate, map, and align top-level and domain ontologies and taxonomies within the larger science ecosystem.
- **Transportable Data Environment**: By Transportable Data Environment we mean an environment enabling researchers to copy data from a source database to a target database. This environment should be based on a transport technology supporting the creation of *transportable modules* which function like a shipping service that moves a package of objects from one site to another at the fasted possible speed. Transportable modules enables one to rapidly copy a group of related database objects from one database to another. The physical and logical structures of the objects contained in the transportable modules being restored are re-created in the target database [16].
- **Transportable Data Services/Tools Environment**: Increasingly, the volumes of data produced by high-throughput instruments and simulations are so large, that it is much more economical to move computation to the data rather than moving the data to the computation. A Transportable Data Services/Tools Environment should support this model made possible through service-oriented architectures (SOA) that encapsulate computation into *transportable compute objects* that can be run on computers that store targeted data. SOA compute objects function like applications that are temporarily installed on a remote computer, perform an operation, and then are uninstalled [17].
- **Scientific Workflow Management Environment**: By Scientific Workflow we mean a precise description of a scientific procedure-a multi-step process to coordinate multiple tasks. Each task represents the execution of a computational process. Scientific Workflows orchestrate e-Science services so that they cooperate to efficiently perform a scientific application. A Workflow Management Service should support the creation, maintenance, and operation of scientific workflows.

- **Policy Management Environment**: By Policy Management Environment we mean an integrated set of formal semantic policies that enhances the authorization, obligation, and trust processes that permit regulated access and use of data and services (data policies). The same formal semantic policies are also used to estimate trust based on parties' properties (trust management policies). A Policy Management Environment should provide accurate and explicit policy representation and specification languages, policy editor tools, policy administrator tools, algorithms for conflict detection and resolution, and graphical tools for editing, updating, removing, and browsing policies as well as de-conflicting newly defined policies.

The ultimate aim of a Global Research Data Infrastructure is to enable **global collaboration** in key areas of science by supporting science ecosystem views, channels, and services and creating, thus, a science **collaborative** environment.

Social and Organizational Dimensions of a Research Data Infrastructure

A Digital Science Ecosystem model must also consider the fact that external environmental forces influence research advances. Specifically, three major types of external environmental forces should be considered: **social** and **governmental** forces, **economic** forces, and **technical** forces [18]. A viable vision of research data infrastructure must take into account social and organizational dimensions that accompany the collective building of any complex and extensive resource.

A robust Global Research Data Infrastructure must consist not only of a technical infrastructure but also a set of organizational practices and social forms that work together to support the full range of individual and collaborative scientific work across diverse geographic locations. A data infrastructure will fail or quickly become encumbered if any one of these three *critical* aspects is ignored. By considering data infrastructure as just a technical system to be designed, the importance of *social, institutional, organizational, legal, cultural, and other non-technical problems are marginized and the outcome is almost always flawed or less useful than originally anticipated* [4].

Tensions

New research data infrastructures are encountering and often provoking a series of **tensions** [4]. Because of the potential to upset or recast previously accepted relations and practices the development of new data infrastructures may generate, in some degree, what economists have labeled "creative destruction". This occurs when established practices, organizational norms, individual and institutional expectations adjust in a positive or negative fashion in reaction to the new possibilities and challenges posed by infrastructure. In the best circumstances, individuals and institutions take advantage of and build upon new resources. In other cases, the inertia of long-standing organizational arrangements, scholarly approaches and research practices prove to be too difficult to change with resulting disastrous consequences. Tensions should be thought of as both barriers and resources to infrastructural development, and should be engaged constructively.

A second class of tensions can be identified in instances where changing infrastructures bump up against the constraints of political economy: intellectual property rights regimes, public/private investment models, ancillary policy objectives, etc. Clearly, the next generation of research data infrastructures pose new challenges to existing regimes of intellectual property. Indeed, intellectual property concerns are likely to multiply with the advent of increasingly networked and collaborative forms of research supported by the data infrastructures [4].

Similar tensions arise in determining relationships between national policy objectives and the transnational pull of science. Put simply, where large-scale policy interests (in national economic competitiveness, security interests, global scientific leadership, etc.) stop at the borders of the nation-state, the practice of science spills into the world at large, connecting researchers and communities from multiple institutional and political locales. This state of affaires has a long history of creating tension in science and education policy, revealing in very practical terms the complications of co-funding arrangements across multiple national agencies [4].

To the extent that research data infrastructures support research collaborations across national borders, such national/transnational tensions must be carefully considered and efforts must be continually undertaken to resolve them.

4. Technological Challenges

Data Challenges

Data challenges include research data modelling, data management and data tools.
There is a need for radically new data models and query languages and tools that enable scientists to explore new methodologies, follow new data content paths, apply new techniques, and build and test new models in new ways that facilitate innovative research activities.

Data Modeling Challenges

There is a need for radically new approaches to research data modelling. In fact, the current data models (relational model) and management systems (relational database management systems) were originally developed by the database research community for business and commercial data applications. Research data has completely different characteristics from business/commercial data and thus the current database technology is inadequate to handle it efficiently and effectively.
There is a need for data models and query languages that:

- more closely match the data representation needs of the several scientific disciplines;
- describe discipline-specific aspects (metadata models);
- represent and query data provenance information;
- represent and query data contextual information;
- represent and manage data uncertainty;
- represent and query data quality information.

Data Management Challenges

If research data are well organized, documented, preserved and accessible, and their accuracy and validity is controlled all times, the result is high quality data, efficient research, findings based on solid evidence and the saving of time and resources. Researchers themselves benefit greatly from good data management. It should be planned before research starts and may not necessarily incur significant time or costs if it is engrained in standard research practice. A Data Management Plan helps researchers consider, when research is being designed and planned, how data will be managed during the research process and shared afterwards with the wider research community [19].

Data Tools Challenges

Currently, the available data tools for most scientific disciplinary research are not adequate. It is essential to build better data management tools and computational algorithms in order to make scientists more productive. Better data and computational tools are essential to visualize, analyze, and catalog the enormous research datasets that enable a data-driven research.

Scientists will need better analysis algorithms that can handle extremely large datasets with approximate algorithms (ones with near-linear execution time), they will need parallel algorithms that can apply many processors and many disks to the problem to meet CPU-density and bandwidth-density demands, and they will need the ability to "steer" long-running computations in order to prioritize the production of data that is more likely to be of interest [20].

Scientists will need better data mining algorithms to automatically extract valid, authentic and actionable patterns, trends and knowledge from large data sets. Data mining algorithms such as automatic decision tree classifiers, data clusters, Bayesian predictions, association discovery, sequence clustering, time series, neural networks, logistic regression when integrated directly in database engines will increase the scientist's ability to discover interesting patterns in their observations and experiments [20].

Large observational data sets, the results of massive numerical computations, and high-dimensional theoretical work all share one need: *visualization*. Observational data sets such as astronomic surveys, seismic sensor output, tectonic drift data, ephemeris data, protein shapes, and so on, are infeasible to comprehend without rendering them for the human visual system [20].

In essence, scientists need advanced tools that enable them to follow new paths, try new techniques, build new models and test them in new ways that facilitate innovative multidisciplinary/interdisciplinary activities and support the whole research cycle.

5. Organizational Challenges

From the organizational point of view a research data infrastructure must support the Research and Publication Process.

This process is composed of the following phases: (i) the original scientist produces, through research activity, primary, raw data; (ii) this data is analysed to create secondary data, results data; (iii) this is then evaluated, refined, to be reported as tertiary information for publication; (iv) with the mediation of the pre-print and peer review mechanisms, this then goes into the traditional publishing process and feeds publication archives. An alternative to phase (i), is when a scientist may perform research based on data, i.e. using data to make new discoveries or to obtain further insights [21].

Primary data is archived into dynamic digitally **curated** data repositories (**Digital Data Libraries**). By curated data we mean that this data is associated with metadata and kept dynamic with annotations and links to other research.

In research organizations, two roles are rapidly growing in importance: the **data archivist** and the **data curator.**

Data archivist: in general people in this role need to interact with the data provider to prepare data for archiving (such as generating metadata which will ensure that it can be found, and can be rendered or used in the future).

Data curator: people in this role need to keep data dynamic with annotations and linked to other research as well as continuously reviewing the information in their care, though they may still maintain archival responsibilities. They should also take an active role in promoting and adding value to data holdings, and managing the value of collections.

Static digital data is stored into **Digital Data Archives** for long-term preservation.

The relationship between constantly curated, evolving datasets and those in static digital archives is one that needs to be explored, through research and accumulation of practical experience.

Publications are archived into publication archives (**Digital Research Libraries).**

Future research data infrastructures must guarantee interoperability between Digital Data Libraries, Digital Data Archives and Digital Research Libraries in order to be able to support the scientific scholarly and research processes.

6. System Challenges

System challenges include open and extensible architectures, virtual research environments, science gateways, interoperability and mediation software, new computing and programming paradigms (cloud computing and MapReduce).

Virtual Research Environment (VRE)

By Virtual Research Environment (VRE) we mean the “technological framework”, i.e., digital infrastructure and services that enable the creation of “virtual working environments” in which “communities of research” can effectively and efficiently conduct their research activities. It can be viewed as a framework within which tools, services, and resources can be plugged. Next generation research data infrastructures must provide the necessary architectural and management tools in order to be able to build, support, and maintain VREs.

Science Gateways

By Science Gateway we mean a community-specific set of tools, applications, and data collections that are integrated together and accessed within a scientific data infrastructure via a portal or a suite of applications [22].

Science gateways can support a variety of capabilities including workflows, visualization as well as resource discovery and job execution services.

Interoperability and Mediation Software

We adopt the IEEE definition of interoperability - “*The ability of two or more systems or components to exchange information and to use the information that has been exchanged*”

There are three main problems that hamper the interoperability between two entities [23]:

The Heterogeneity Problem

During the data exchange process between two entities (data producer and data consumer) different sources of heterogeneity can be encountered depending on: how data are requested by the consumer entity; the use of different terminologies; how data will be represented; the semantic meaning of each data; and how data are actually transported over a network.

The Logical Inconsistency Problem

Some logical inconsistencies may arise between functional descriptions of services (producer) and requests (consumer). When the exchanged information specify the functionality of a service or what is required to satisfy a consumer request, some inconsistencies may arise between the logical relationships of these descriptions.

The Usage Inconsistency Problem

Usage inconsistency means that the consumer's goal, that is, the objectives that she/he wants to achieve by using the producer's resources cannot be reached. In order for a consumer to productively use exchanged data, the data must be complemented with descriptive information that provides the context, provenance, lineage, and other information that gives additional meaning. The descriptive information should be modelled by **purpose-oriented** metadata models.

Mediation Software

The main concept enabling interoperability of data/services/policies is mediation. This concept has been used to cope with many dimensions of heterogeneity spanning data language syntaxes, data models, and semantics. The mediation concept is implemented by a mediator, which is a software device capable of establishing interoperability of resources by resolving heterogeneities and inconsistencies. It supports a mediation schema capturing user requirements, and an intermediation function between this schema and the distributed information sources [24].

There are four main mediation scenarios:

*Mediation of data structures***:** permits data to be exchanged according to syntactic, structural and semantic matching.

*Mediation of functionalities***:** makes it possible to overcome mismatching of functional descriptions of two entities that are expressed in terms of pre- and post-conditions.

*Mediation of policies logics***:** employs techniques to solve policy mismatches.

*Mediation of protocols***:** makes it possible to overcome behavioural mismatches among protocols run by interacting parties.

The ultimate aim should be the definition and implementation of an "*integrated mediation framework*" capable of providing the means to handle and resolve all kinds of heterogeneities and inconsistencies that might hamper the effective usage of the resources of a global research data infrastructure [25].

We envision that one of the most important features of the future research data infrastructures will be the mediation software.

Infrastructural Services

An infrastructural service is defined as a network-enabled entity that provides some capability. Entities are network-enabled when they are accessible from other computers than the one they are residing on. Research data infrastructures must provide a sufficient measure of network-enabled "support services" in order to achieve the conditions needed to facilitate effective collaboration among geographically and institutionally separated communities of research. A support service should be [26]:

Shareable: it must be able to be used by any set of users in any context consistent with its overall goals.

Common: it must present a common consistent interface to all users, accessible by a standard procedure. The term "common" may be synonymous with the term "standard".

Enabling: it must provide the basis for any user or set of users to create, develop, and implement any applications, utilities, or services consistent with its goals.

Enduring: it must be capable of lasting for an extensive period of time. It must have the capability of changing incrementally and in an economical feasible fashion to meet the slight changes of the environment, but be consistent with the broader world view. In addition, it must change in a fashion that is transparent to the users.

Scale: the service can add any number of users or uses and can by its very nature expand in a structured manner in order to ensure consistent levels of service.

Economically sustainable: it must have economic viability.

A Global Research Data Infrastructure must provide support services for:

- data registration,
- data discovery,
- data citation,
- data service/tool discovery,
- data search,
- data sharing,
- data linking,
- data federation (including data harmonization, data fusion, data integration),
- recommendation,
- scientific workflow management,
- data transportation,
- service transportation, and
- policy management.

7. New Computing and Programming Paradigms

Cloud Computing

Cloud Computing is a new term for a long-held dream of computing as a utility, which has recently emerged as a commercial reality.

By Cloud Computing it is meant a large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualized, dynamically-scalable, managed computing power, storage, platforms, and services are delivered on demand to external customers over the Internet [27]. The key points of this definition are: (i) Cloud Computing is a specialized distributed computing paradigm; (ii) it is massively scalable; (iii) it can be encapsulated as an abstract entity that delivers different levels of services to customers outside the Cloud; (iv) it is driven by economies of scale; (v) the services can be dynamically configured (via virtualization or other approaches) and delivered on demand.

We envision that the future Digital Data Libraries (Science Data Centers) will be based on cloud philosophy and technology. Each scientific community of practice will have its own Cloud(s); the federation of these Clouds will allow collaboration among communities of practice enabling thus multidisciplinary research.

MapReduce

Many data-intensive applications require hundreds of special-purpose computations that process large amounts of raw data. MapReduce is a programming model and an associated implementation for processing and generating large data sets while hiding lower level details of parallelization, fault-tolerance, data distribution, and load balancing. The MapReduce programming model has been successfully used for many different purposes. This success can be attributed to several reasons. First, the model is easy to use; second, a large variety of problems are easily expressible as MapReduce computations; and third, several implementations of MapReduce have been developed that scale to large clusters of machines comprising thousands of machines.

These implementations make efficient use of these machine resources and therefore are suitable for use on many of the large computational problems encountered in data-intensive applications [28].

8. Policy Challenges

The need for using semantic policies in science ecosystem environments is widely recognized.

It is important to adopt a broad notion of policy, encompassing not only access control policies, but also those related to trust, quality of service, and others. Policies should eventually be integrated into a single coherent framework, so that the framework can be implemented and maintained by a research data infrastructure, and that individual policies themselves can be harmonized and synchronized as needed.

Open Data – Open Science

There is an emerging consensus among the members of the academic research community that "e-science" practices should be congruent with "open science". We envision that the future research data infrastructures will constitute infrastructures for open scientific research.

"Openness means access on equal terms for the international research community at the lowest possible cost, preferably at no more than the marginal cost of dissemination. Open access to research data from public funding should be easy, user-friendly and preferable Internet-based" [29].

The Open Data Principle has three dimensions: policy, legal, and technological**.** Technology must render physical and semantic barriers irrelevant, while policies and laws must address and supplant outdated legal jurisdictional boundaries.

The principles of open science data and open science can be widely accepted only if realized within a shared Integrated **Science Policy Framework** to be implemented and manifested by global research data infrastructures.

9. Recommendations

1. Future Scientific Data Infrastructures must enable Science Ecosystems.
2. Science organizational aspects should be taken in due consideration when designing global research data infrastructures as well as potential tensions which could be faced or provoked by them.
3. Global Research Data Infrastructures must be based on scientifically sound foundations.
4. Formal models and query languages for data, metadata, provenance, context, uncertainty and quality must be defined and implemented.
5. New advanced data tools (data analysis, massive data mining, data visualization) must be developed.
6. New advanced infrastructural services (data discovery, tool discovery, data integration, data/service transportation, workflow management, ontology/taxonomy management, policy management, etc) must be developed.
7. Future Research Data Infrastructures must support open linked data spaces.
8. Future Research Data Infrastructures must support interoperation between science data and literature.
9. The principles of open science and open data in order to be widely accepted must be realized within an Integrated Science Policy Framework to be implemented and enforced by global research data infrastructures.
10. A new international research community must be created.
11. New Professional Profiles must be created.

10. References

[1] G. Bell, T. Hey, and A. Szalay, “Beyond the Data Deluge”, *Science,* **323**, 1297-1298, March 2009

[2] T. Hey, S. Tansley and K. Tolle (Eds.), *The Fourth Paradigm: Data Intensive Scientific Discovery*. Redmond, WA: Microsoft, 2009.

[3] C. Anderson, “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete”, Wired Magazine: 16.07 . Retrieved from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory june 2011

[4] P. Edwards, S. Jackson, G. Bowker and C. Knobel, *Understanding Infrastructure: Dynamics, Tensions, and Design*, Final Report of the Workshop on History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures, Jan. 2007. Retrieved from http://hdl.handle.net/2027.42/49353 June 2011

[5] T. Jewett, R. Kling, “The dynamics of computerization in a social science research team: a case study of infrastructure, strategies, and skills”, *Social Science Computer Review*, **9** (2), 246-275, 1991.

[6] S. Star, and K. Ruhleder, “Steps toward an ecology of infrastructure: Design and access for large information spaces”, *Information Systems Research*, **7** (1), 111-134, 1996.

[7] I. Gordon, P. Greenfiels, A. Szalay and R. Williams, “Data Intensive Computing in the 21st Century”, *Computer*, **41** (4), 30-32, April 2008.

[8] L. Bannon, and S. Bodker, “Constructing Common Information Spaces” in: J. Hughes, T. Rodden, W. Prinz and K. Schmidt (eds.), *ECSCW’97: Proceedings of the Fifth European Conference on Computer-Supported Cooperative Work*, Kluwer Academic Publishers, 1997.

[9] J. Gray, D. Liu, A. Szalay, D. DeWitt and G. Heber, “Scientific Data management in the Coming Decade”, *SIGMOD Record*, **34** (4), 35-41, Dec. 2005

[10] National Science Board, *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*, National Science Foundation, 2005 Retrieved from http://www.nsf.gov/pubs/2005/nsb0540/ June 2011

[11] *What is Data Archiving*? [Definition] Retrieved from http://searchdatabackup.techtarget.com/definition/data-archiving June 2011

[12] P. Graham, *The Digital Research Library: Tasks and Commitments*, 1995. Retrieved from www.csdl.tamu.edu/DL95/papers/graham/graham.html June 2011

[13] *NSF' Cyberinfrastructure Vision for 21st Century Discovery*, NSF Cyberinfrastructure Council, March 2007.
Retrieved from
http://www.nsf.gov/pubs/2007/nsf0728/index.jsp
June 2011

[14] "Jim Gray on eScience: A Transformed Scientific Method", in: *The Fourth Paradigm: Data Intensive Scientific Discover* [2], xix-xxxiii.

[15] N. Paskin ,"Digital Object Identifiers for scientific data", *Data Science Journal,* **4***,* 12-20, March 2005

[16] "Moving Large Volumes of Data Using Transportable Modules" in Oracle® Warehouse Builder Data Modeling, ETL, and Data Quality Guide" 11g Release 2 (11.2)
Retrieved from
http://download.oracle.com/docs/cd/E14072_01/owb.112/e10935/trans_mod.htm
June 2011

[17] S. Kahn, "On the Future of Genomic Data", *Science*, **331** (6018), 728-729.

[18] G. Adomavicius, J. Bockstedt, A. Gupta and R. Kauffman, "Understanding Patterns of Technology Evolution: An Ecosystem Perspective", in: *System Sciences, 2006. HICSS '06. Proceedings of the 39th Annual Hawaii International Conference on* , vol.8, pp. 189a, 04-07 Jan. 2006.

[19] V. Van den Eynden, L. Corti, M. Woollard, L. Bishop and L. Horton, *Managing and Sharing Data: Best Practices for Researchers*. UK Data Archive, University of Essex, May 2011

[20] Microsoft Draft Roadmap "Towards 2020 Science"
Retrieved from: http://www. Jyu.edu.cn

[21] P. Lord and A. Macdonald, *Data Curation for e-Science in the UK*: *an audit to establish requirements for future curation and provision*. E-science Curation Report, 2003.

[22] N. Wilkins-Diehr, "Special Issue: Science Gateways – Common Community Interfaces to Grid Resources", *Concurrency and Computation: Practice and Experience*, **19** (6) 743-749, April 2007.

[23] C. Thanos, *Interoperability: A Holistic Approach*, Manuscript, 2010.

[24] G. Wiederhold, "Mediators in the architecture of future information systems", *Computer*, **25** (3), 38-49, March 1992.

[25] M. Stollberg, E. Cimpian, A. Mocan and D. Fensel, "A Semantic Web Mediation Architecture", in: *Proceedings of the 1st Canadian Semantic Web Working Symposium (CSWWS 2006*), 22 pp., Springer, 2006.

[26] O. Nanseth and E. Monteiro, *Understanding Information Infrastructure,* Manuscript, 1998.

[27] I. Foster, I.Yong Zao Raicu, and S. Lu, "Cloud Computing and Grid Computing 360-Degree Compared", in: *Proc. IEEE Grid Computing Environments Workshop*, IEEE Press, 2008, pp.10

[28] J. Dean and S. Ghemwat, "MapReduce: Semplified data Processing on Large Clusters", in: *OSDI 2004, USENIX Symposium on Operating System Design and Implementation*, 137-150.
Retrieved from
http://www.usenix.org/events/osdi04/tech/dean.html
June 2011

[29] OECD Recommendations of the Council concerning Access to Research Data from Public Funding" C (2006)184, Dec. 14, 2006.