A VISION FOR GLOBAL RESEARCH DATA INFRASTRUCTURES

Costantino Thanos

Institute of Information Science and Technologies (ISTI) of the Italian National Research Council (CNR)
Email: Costantino.Thanos@isti.cnr.it

ABSTRACT

New high-throughput scientific instruments, telescopes, satellites, accelerators, supercomputers, sensor networks, and running simulations are generating massive amounts of data. In order to be able to exploit these huge volumes of data, a new type of e-infrastructure, the Global Research Data Infrastructure (GRDI), must be developed for harnessing the accumulating data and knowledge produced by the communities of research. This paper identifies the main challenges faced by the future GRDIs, defines a conceptual framework for GRDIs based on the ecosystem metaphor, describes a core set of functionality that these GRDIs must provide, and gives a set of recommendations for building the future GRDIs.

Keywords: Information networks, Distributed systems, Distributed databases, Interoperability

1 INTRODUCTION

New high-throughput scientific instruments, telescopes, satellites, accelerators, supercomputers, sensor networks, and running simulations are generating massive amounts of data. Often referred to as a data deluge, massive datasets are revolutionizing the way research is carried out, which results in the emergence of a new fourth paradigm of science based on data-intensive computing (Hey, Tansley, & Tolle, 2009). This new data-dominated science will lead to a new data-centric way of thinking, organizing, and carrying out research activities that 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, will also produce an important change in the scientific method. One view put forward is that the traditional scientific method driven by hypotheses, essentially a deductive method, will be complemented by a data-driven method, essentially an inductive one.

In order to be able to exploit these huge volumes of datasets, 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.

2 RESEARCH DATA INFRASTRUCTURES

Research data infrastructures can be defined as managed networked environments for digital data consisting of services and tools that support: the whole research cycle, the movement of scientific data across scientific disciplines, the creation of open linked data spaces by connecting datasets from diverse disciplines, the management of scientific workflows, the interoperation between scientific data and literature, and 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 item is ignored (Edwards, Jackson, Bowker, & Knobel, 2007).

Another school of thought considers (data) infrastructure as a fundamentally relational concept. It becomes infrastructure in relation to organized (research) practices (Jewett & Kling, 1991). 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.

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. Science is a global undertaking, and research data are both national and global assets. Therefore, there is a need for global research data infrastructures (GRDIs) 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 the data.

3 DATA-INTENSIVE MULTIDISCIPLINARY RESEARCH

The next generation of GRDIs faces two main challenges: to effectively and efficiently support

- data-intensive science and
- multidisciplinary science.

By data-intensive science we mean any scientific research activity whose progress is heavily dependent on careful thought about how to use data. It is 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.

By multidisciplinary approach to research problems we mean drawing appropriately from multiple disciplines in order to redefine a research problem outside of normal boundaries and reach solutions based on a new understanding of complex situations. There are several barriers to the multidisciplinary scientific approach of a 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 data moves across semantic boundaries (semantic distortion) (Bannon & Bodker, 1997).

Therefore, GRDIs must face two scale challenges:

- large scale in data volume and
- wide-scale in data contextual diversity.

4 THE RESEARCH DATA INFRASTRUCTURE LANDSCAPE

The current research data infrastructure landscape is characterized by efforts aiming to:

- **develop a number of discipline-specific research data infrastructures**
  In Europe several discipline-specific data infrastructures are underway: in the field of environmental sciences a number of projects supported by the "European Strategy Forum on Research Infrastructures (ESFRI)" are funded by the European Commission: European Multidisciplinary Seafloor Observatory (EMSO) (www.emso-eu.org), European Plate Observation System (EPOS) (www.epos-eu.org), European Contribution to a Global Ocean Observing System (ARGO) (www.argo.net), Integrated Carbon Observation System (ICOS) (www.icos-infrastructure.eu), e-Science and Technology Infrastructure for Biodiversity Data and Ecosystem Research (LifeWatch) (www.lifewatch.eu). In the marine field some important projects are underway: Data Infrastructure for Ecosystem Approach to Management of Marine Living Resources (iMARINE) (www.i-marine.eu), Ocean Data Interoperability Platform (ODIP) (www.odip.eu), Global Ocean Observing System (GOOS) (www.ioc-goos.org), Ocean Biogeographic Information System (OBIS) (www.obisproject.com), and Geo-Seas (www.geo-seas.eu). In the cultural heritage field an important project, Common Language Resources and Technology Infrastructure (CLARIN) (www.clarin.eu), aimed at making language resources available to humanities and social sciences research communities was funded by the European Commission. In the Earth sciences very relevant is the Global Earth Observation System of Systems (GEOSS) project (www.opengeospatial.org). Another important EU funded project is the European Data Infrastructure (EUDAT) (www.eudat.eu) aimed at developing a collaborative data infrastructure.

  It is worthwhile to mention two important US infrastructural projects funded by NSF: Data Observation Network for Earth (DataONE) (http://www.dataone.org/), which aims at developing a distributed framework and sustainable cyber-infrastructure that meets the needs of science for open access to Earth observational data, and EarthCube (http://earthcube.ning.com), which aims to develop a community-guided cyber-infrastructure to integrate information and data across the geosciences. Another relevant initiative is the creation of the World Data System (WDS) (www.icsu-wds.org) by the International Council for Science (ICSU) for the purpose of enabling universal and equitable access to quality-assured scientific data, data services, products, and information. In the agricultural sciences, it is worthwhile to mention the International Information System for the Agricultural Sciences and Technology (AGRIS) (http://agris.fao.org/) supported by the Food and Agriculture Organization (FAO) of the United Nations.

- **produce white papers, roadmaps reports, and high level policy reports**
  Several policy reports, white papers, and roadmaps reports have been produced by the European Commission, Scientific organizations (ESFRI, e-IRG, PARADE, KE), funding agencies (NSF, ANDS) and projects (GRDI2020). Of particular importance, the report produced by a High Level Expert Group, "Riding the Wave", (http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf) and supported by the European Commission calls for a collaborative data infrastructure that will enable scientists to use, re-use, and exploit research data to the maximum benefit of science and society. The ESFRI roadmaps (2008, 2010) (http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri-roadmap), as well as the e-IRG white papers (2009, 2011) (http://www.osiris-online.eu/RI%20Reports/e-irg_white_paper_2009_final.pdf) and (https://www.egi.eu/news-and-media/newsletters/Inspired_Summer_2011/e-IRG_White_Paper_2011.html) set out strategies for building interoperable e-infrastructures including data infrastructure components.

  The white paper "Strategy for a European Data Infrastructure" produced by the Partnership for Accessing Data in Europe (PARADE) (http://www.csc.fi/english/pages/parade) suggests a strategy for data related services and outlines a persistent, multidisciplinary European data infrastructure based on the needs of user communities. The Report "A Surfboard for Riding the Wave" produced by the Knowledge Exchange (KE) presents a four country (UK, Denmark, Germany, Netherlands) action program on research data (www.knowledge-exchange.info). Similar reports have been produced in US, the Blue Ribbon Task Force Report, on sustainable data infrastructures (http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf). Finally, some projects (GRDI2020) have produced Roadmap Reports on global research data infrastructures (www.grdi2020.eu).

- **make the scientific data infrastructures interoperable**
  A global effort aiming at facilitating research data sharing, use, and re-use is undertaken by the Research Data Alliance (RDA) (http://rd-alliance.org). To achieve its objectives, RDA addresses a wide range of issues including definition of new data standards and harmonization policies regarding existing standards, development of mediation techniques to solve interoperability problems, definition of new approaches to data discoverability as well as new data policy frameworks to support data openness.

- **rationalize the development processes of data infrastructures**
  Finally, some projects, for example ENVRI (www.envri.eu), funded by the European Commission, aim to draw up guidelines for the common needs of projects addressing a scientific sector and implement harmonized solutions.

- **develop new (meta) data modeling techniques, advanced data tools, and new data management systems for scientific applications**
  An intensive research activity is being undertaken by the international research community for the purpose of developing new techniques for modeling research data and metadata, new approaches to data discoverability, integration and correlation, and more advanced data tools for mining, visualizing, and analyzing research data. Efforts are also devoted to building data management systems for scientific applications.

The combined effect of all these relevant and important contributions as well as the experience gained in authoring the GRDI2020 Roadmap Report (www.grdi2020.eu) have inspired this work. We have observed, however, that a shared definition of a research data infrastructure is still missing. Currently, many different system implementations from the functional and architectural point of view are all termed as "*research data infrastructure*". We have, thus, identified the need for a conceptual framework in order to be able to characterize the functionality of the next generation of global research data infrastructures more precisely and comprehensively.

5 A CONCEPTUAL FRAMEWORK FOR A GLOBAL RESEARCH DATA INFRASTRUCTURE

Infrastructures have a supporting or enabling function. This is opposed to being especially designed to support one way of working within a specific scientific discipline. In order to be able to characterize this supporting or enabling function of a research data infrastructure, we need to define a conceptual organizational model for scientific research and identify the organizational units of this model and the interactions between them.

We have adopted the ecosystem metaphor to model the way the scientific disciplines are organized and to conceptualize all the "research relationships" between the components of a digital science ecosystem. The biological ecosystem model has been used in the past for analyzing the technological evolution, the technology ecosystem (Adomavicius, Bockstedt, Gupta, & Kauffman, 2006), for studying business relationships and strategic decision making, the business ecosystem (Iansiti & Levien, 2002), for analyzing industrial relationships, industrial ecology (Iansiti & Levien, 2004), etc. The research in digital ecosystems aims at exploiting the properties of natural ecosystems in artificial systems. 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 a digital ecosystem of scientific research would bring about a better understanding of its dynamic nature. We believe that the ecological metaphor for understanding the complex network of data-intensive multidisciplinary-interdisciplinary 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 simple technological progress. By adopting the model of a digital ecosystem of scientific research, the main challenge we face is the identification of the key properties of natural ecosystems and how these properties map to properties in digital science ecosystems. We have identified a list of properties including: cooperation and competition, landscape connectivity, self-adaptability, stability and diversity (www.digital-ecosystems.org/book/papers/t1.0.pdf), (arxiv.org/ftp/arxiv/papers/0910/0910.0646.pdf).

Cooperation and competition: Research activity is strongly dependent on finding the right balance between *cooperation* and *competition*. In particular, whereas competition works well in many research contexts, cooperation can amplify the positive effects of formalization and codification of knowledge within scientific disciplines. This means that a digital science ecosystem should be an environment where knowledge is constantly created and shared, flowing freely and dynamically where it is needed.

Landscape connectivity: By *connectivity* we mean processes that affect "communication" within a population. Sites in a landscape are connected if there are patterns or processes that link them in some way. Fragmentation can have several consequences. From a digital science ecosystem perspective, connectivity can mediate global and local research strategies. In a well connected science landscape, selection favours the globally superior, and pursuit of different research directions is discouraged, potentially leading to premature convergence. When the science landscape is fragmented, research communities may diverge, solving the same problems in different ways.

Adaptability: Natural ecosystems are often described as complex *adaptive* systems (for example, brains, individuals, economies, etc.). Digital science ecosystems are embedded in environments (technological, economic, social), upon which they are dependent. Change in the environment might require change in the science ecosystems. They must have the potential for dynamic adaptive self-organization. Constructing a digital science ecosystem requires a balance between freedom of the system to self-organize and constraint of the system to generate meaningful and useful solutions demanded by the communities of research.

Stability and diversity: A key question is how the *stability* and *diversity* properties of the natural ecosystems map to properties of the digital ecosystems. Stability is a trade-off; we want the system to respond to external changes with rapid adaptation but not to be so responsive that sudden state changes prevent control. Sustained diversity is a key requirement for dynamic adaptation. In a digital science ecosystem, diversity must be balanced against adaptive efficiency because maintaining large numbers of poorly-adapted solutions is costly.

5.1 Defining the digital science ecosystem

We introduce a complex organizational science model, based on the ecosystem approach, which is composed of a number of organizational units: discipline-specific data centers, discipline-specific data archives, research digital libraries, and communities of research.

Discipline-specific data centers
They should be designed to ensure the stewardship and provision of quality-assessed data and data services to the international science community and other stakeholders. Each discipline-specific data center 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.

Discipline-specific data archives
They should be designed to ensure the long-term preservation of research data and methods that are no longer actively used. 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.

Research digital libraries
A research digital library is a collection of electronic documents. The mission of research libraries is to acquire scientific documents, organize them, and make them available.

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.

Discipline-specific data centers, data archives, and research digital libraries support complementary phases of the research and publication process conducted by the communities of research. In fact, data centers function as a central service where researchers can both deposit data they have created and also find data they can reuse within

their own work. Data centers do not deal with data publishing aspects and suffer from a major lack of best practices and technologies in order to support a rigorous scientific communication workflow. Data archives support the phase of data long term preservation. On the other hand, research digital libraries are framed too narrowly, typically serving as repositories for the final result of the research and publication process.

So far, there is no communication among these three components of the science ecosystem; in particular, a dichotomy has been created between data centers/data archives and research libraries. This situation conflicts also with the needs of modern science, which requires an integrated support to the whole data research and publication life cycle. In fact, scientific communities and their funding bodies are talking about the need for scientists to publish their raw data sets, experimental details, analytical methods, and visualizations in addition to the traditional scholarly publications. In order to satisfy this requirement, all these three components must be interoperable, and the existing dichotomy between them must be bridged.

5.2 Science ecosystem concepts

Our organizational model of the science world, based on the ecosystem approach, includes three fundamental concepts that define the way a science ecosystem can be partitioned and the way its organizational units/components support the research and publication process through their interactions. In addition, they specify the means by which their holdings are made discoverable, interoperable, and usable.

Science ecosystem views
A specific ecosystem view can be defined by identifying a community of research, a set of data collections distributed in one or more data centers/data archives, a set of scientific publications archived in one or more research libraries, and a set of data services and data tools registered in the ecosystem, necessary for the research activity undertaken by this community. A view creates a partition of the science ecosystem within which a specific community of research can conduct its research activities. The partitions created by the views can be, partially, overlapping, meaning that different communities of research can, partially, share collections of data, publications, and data tools.

Science ecosystem channels
The research and publication process is, essentially, composed of the following phases (http://www.jisc.ac.uk/publications/reports/2003/esciencefinalreport.aspx): (i) production by the scientists of primary raw data; (ii) analysis of these data to create secondary and results data; (iii) evaluation and refinement of these data to be reported as tertiary information for publication; (iv) feeding of the data to the publication archives through the traditional publishing process with the mediation of the peer review mechanisms; (v) determination of whether to destroy the data produced or to transfer them to an archive for a long term preservation. The transition from one phase of the research and publication process to the next one implies the movement of data in their evolving forms (raw, primary, etc.) from one organizational unit to another.

We define ecosystem channels as the conceptual conduits through which the data flow from one organizational unit to another during a transition from one phase to another. We classify these ecosystem channels according to the phase of the research and publication process they are supporting. For example, a channel through which data flow from a data center to a data archive enables data preservation as it allows the flow of the data that have been deemed to need long term preservation from a short storage to a long-term one. A channel through which data flow from a data center to a research library enables the modern scientific communication as it allows the integration of published ideas with the underlying data. A channel through which data flow from a domain-specific data center to a different domain-specific data center enables multidisciplinary-interdisciplinary research.

Science ecosystem support services
Today, scientific knowledge is network-centric, meaning that it is created by linking, correlating, integrating, aggregating, etc. information units composed of text, datasets, images, videos, sound recordings, mathematical models, workflows, presentational materials, and software packages, distributed among the different components of a science ecosystem and connected by semantic relationships and interwoven in a complex variety of ways. In turn, datasets, images, etc., in order to be interpretable, require linkage to some metadata information (contextual and

provenance information). Thus, scientific knowledge will become a network of information units. The creation of such a network-centric scientific knowledge must be supported by appropriate services.
We define ecosystem support services as networked enabled entities that provide some capability. They must make the holdings of the single components of a science ecosystem discoverable and enable the researchers to correlate, aggregate, and integrate them.

5.3 A vision for a global research data infrastructure

We envision a Global Research Data Infrastructure (GRDI) as:

- an enabler of an open, extensible, and evolvable digital science ecosystem;
- a facilitator of research data, information, knowledge, and data tools discovery;
- an enhancer of problem-solving processes.

A GRDI must support the creation, operation, and maintenance of the main science ecosystem concepts, i.e., views, channels, and support services.

5.3.1 Ecosystem views

Ecosystem views can be materialized through:

- **Science Gateways**: a science gateway is a community-specific set of tools, applications, data, and publication 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 virtual working environment, created on demand, in which communities of research can effectively and efficiently conduct their research activities. It can be viewed as a framework within which data tools and services can be plugged. A VRE is always associated with a community of research.

5.3.2 Ecosystem channels

The implementation of ecosystem channels should be based on technologies supporting the identification of the ecosystem resources (data, publications, tools, services, methods, etc.), their discovery, and transportation from one organizational unit of the science ecosystem to another. In addition, establishing a channel between two different discipline-specific organizational units of a science ecosystem implies the overcoming of the heterogeneity problems (syntactic, semantic, and pragmatic) encountered when ecosystem resources are moved between these kinds of organizational units. The ecosystem channels reduce the science fragmentation due to disparate data, reduce geographic fragmentation of datasets, and accelerate the rate at which data and information can be made available and used to advance science.

5.3.2.1 Interoperability

Data/information/knowledge, when moving between disciplines, have to cross a number of “knowledge boundaries”. A first boundary (syntactic boundary) is constituted by the different syntax of the languages used by the communities/disciplines in order to interact among them. A second boundary (semantic boundary) can arise even if a common syntax or language is present due to the fact that interpretations are often different. A third boundary (pragmatic boundary) arises when a community/discipline is trying to influence or transform the knowledge created by another community/discipline.

A useful means of representing, learning about, and transforming knowledge to resolve the consequences that exist at a given boundary is the “boundary object” (Star, 1989). A boundary object, in order to overcome a syntactic/semantic/pragmatic boundary, should establish a shared syntax, a shared means for representing and specifying differences, and a shared means for representing and specifying dependencies. We envision that one of

the most important features of the future "disciplinary data infrastructures" will be the definition and efficient implementation of a set of boundary objects (metadata models, data languages, taxonomies, ontologies, etc.) defined by the members of the disciplines being supported.

Defining boundary objects between different scientific disciplines is much more problematic. We envision that in order to enable multidisciplinary research, new methods and techniques must be developed which implement "a mediation function" between boundary objects of different disciplines. By mediation function we mean a function able to map a boundary object defined by a discipline into a semantically equivalent boundary object of another discipline. 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 (Stollberg, Cimpian, Mocan, & Fensel, 2006). We envision that one of the most important features of the future research data infrastructures will be the mediation software.

5.3.3 Ecosystem support services

The ecosystem support services should enable global collaboration in key areas of science and transform the quantitative increase in the availability of research data into a qualitative improvement in research practices and methods. They must make the holdings of the components of a digital science ecosystem discoverable, aggregable, and linkable. A core set of ecosystem support services are briefly described here below.

5.3.3.1 Data registration environment

By data registration environment we mean an environment enabling researchers to make data citable as unique pieces of work and not *only as a part of a* publication. Once accepted for deposit, data should be assigned a "Digital Object Identifier" (DOI) for registration. A DOI (Paskin, 2004) is a unique name (not a location) within a science ecosystem that provides a system for persistent and actionable identification 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 model, and an implementation mechanism determined by policies and procedures for the governance and application of DOIs.

5.3.3.2 Data citation environment

Data citation refers to the practice of providing a reference to data in the same way as researchers routinely provide a bibliographic reference to printed resources. Unfortunately, no universal standards exist for citing quantitative data. A data citation environment should include, at a minimum, five components (Altman, & King, 2007): the author of the dataset, the date the data set was published, the data set title, a unique global identifier system (LSID, DOI, URN, etc.), and a universal numeric fingerprint (UNF). The UNF is a short, fixed-length string of numbers and characters that summarize all the content in the data set such that a change in any part of the data would produce a completely different UNF. The need for introducing the fifth component is justified by the fact that unique global identifiers do not guarantee that the data does not change in any meaningful way when the data storage formats change. Together, the global unique identifier and UNF ensure permanence, verifiability, and accessibility even in the situations where the data are confidential, restricted, or proprietary. It is worthwhile to mention The DataCite (http://www.datacite.org/) international consortium, which addresses the challenges of making data citable in a harmonized, interoperable and persistent way.

5.3.3.3 Data findability environment

In today's network-centric data era one big challenge faced by researchers is pinpointing the location of relevant data. By data findability environment we mean an environment enabling researchers to quickly and accurately find data that support specific research requirements. Currently, there is a conceptual shift from search to findability as the traditional search paradigm is characterized by a lack of context, where the search is conducted in an independent way from professional profiles, context, provenance, and work goals. In the context of a digital science ecosystem, searching for data/information/knowledge is better served by findability. A data findability environment

should be supported by search and query capabilities that exploit semantic data models, semantically rich metadata descriptions contained in data categorization/classification schemes, data dictionaries/data inventories, and metadata registries as well as professional profiles, work goals, context, and provenance.

5.3.3.4 Data tool/service findability environment

Acceptance of the open science principle entails open access not only to research data but also to data tools/services that enable researchers to effectively conduct their research activities. By data tool/service findability environment we mean an environment enabling automatic location of data tools/services that fulfill a research goal. It must support the description of a published data tool/service. The description should be given at three different levels (Keller, Lara, Lausen, Polleres, & Fensel, 2005): at the first level only the static characteristics of the tool/service are described, i.e., only what the tool/service can provide but no longer under which circumstances a concrete tool/service can actually be provided; at the second level the dynamic characteristics of the tool/service are described, i.e., what input information is required for providing a concrete service and what conditions it must fulfill (pre-conditions) and what conditions the object delivered fulfills (post-conditions); and at the third level the deployment capabilities of the hosting operational environment are described. In addition, the description of the researchers' needs must be supported. Several modeling approaches to the description of data tool/service and user needs can be adopted. They include: keyword-based representation models, controlled vocabularies, ontologies, and full-fledged logic. The data tool/service findability environment should also support a data tool/service location process. In addition, as data tool/service findability is based on matching abstract goal and service descriptions, this environment should also provide a mediation support; such mediation support should establish a mapping between the controlled vocabularies/ontologies used to describe goals and services.

5.3.3.5 Data federation

Data federation is an umbrella term for a wide range of decentralized data practices. It covers data integration, data harmonization, and data linking.

5.3.3.5.1 Data integration environment

Data integration is the process of combining data residing at different sources and providing the user with a unified view of these data. Data integration has two broad goals (Bleiholder, & Naumann, 2008): increasing the completeness and the conciseness of data that are available to users and applications. An increase in completeness is achieved by adding more data sources to the system. An increase in conciseness is achieved by removing redundant data, by fusing duplicate entries, and merging common attributes into one. The data integration systems are characterized by an architecture based on a global schema and a set of sources. The sources contain the real data while the global schema provides a reconciled, integrated, and virtual view of the underlying sources.

Data integration is a three-step process: data transformation, duplicate detection, and data fusion.

Data Transformation is concerned with the transformation of the data present in the sources into a common representation (renaming, restructuring). Data from the data sources must be transformed to conform to the global schema of an integrated information system. Modelling the relation between the sources and the global schema is therefore a crucial aspect. Two basic approaches have been proposed for this purpose (Lenzerini, 2002). The first approach, called global-as-view (or schema integration), requires the global schema to be expressed in terms of the data sources. The second approach, called local-as-view (or schema mapping), requires the global schema to be specified independently from the sources, and the relationships between the global schema and the sources are established by defining every source as a view over the global schema.

Data Detection regards the identification of multiple, possibly inconsistent representations of the same real-world objects. The result of the duplicate detection step is the assignment of an object-ID to each representation. Two representations with the same object-ID indicate duplicates.

Data Fusion combines and fuses the duplicate representations into a single representation while inconsistencies in the data are resolved. Several data fusion strategies have been defined. They include: *Conflict-ignoring*, *conflict avoiding,* and *conflict-resolution* strategies.

5.3.3.5.2 Data harmonization environment

Data harmonization is the process of comparing similar conceptual and logical data models to determine the common data elements, similar data elements, and dissimilar data elements, in order to produce a resulting unified data model that can be used consistently across organizational units and data warehouses. A data harmonization environment should support a number of harmonization tools able to: import data models into the tool from various representation formats; linguistically and semantically analyze the components of multiple data models to determine equivalence, similarity, and dissimilarity; construct multi-modal views of the data models to assist in comparison analysis; harmonize and visualize models with multiple users simultaneously and to jointly create the resultant data model; and extract common data elements across models to produce a new resultant skeleton model that can be further refined by a manual process.

5.3.3.5.3 Data linking environment

In the context of a science ecosystem, linking data becomes an imperative as it allows researchers to benefit from the use of multiple datasets created by different communities of research including the connection of publications with the subject data. Linking data refers to the capability, supported by a research data infrastructure, of publishing data on a scientific data space in such a way that they are machine-readable, their meaning is explicitly defined, they are linked to other external data sets, and they can in turn be linked to and from external data sets. A data infrastructure supporting a data space does not act as a data integration system as this requires semantic integration before any service can be provided; instead it follows a co-existence approach, i.e., to provide base functionality over all data sets, regardless of how integrated they are (Franklin, Halevy, & Maier, 2005).

A research data infrastructure should offer the possibility for researchers to start browsing in one data set and then navigating along links into related data sets, or it can support data search engines that crawl the data space by following links between data sets and provide expressive query capabilities over aggregated data. To achieve this, the data infrastructure must support the creation of typed links between data from different sources. Data providers willing to add their data to a scientific data space that allows data to be discovered and used by various applications must publish them according to some principles.

A set of best practices for publishing structured data on the Web referred to as “Linked Data” was introduced by Tim Berners-Lee (Bizer, Heath, & Berners-Lee, 2009) and has become known as the linked data principles. These principles are the following:

- Use URIs as names for things,
- Use HTTP URIs so that people can look up those names,
- When someone looks up a URI, provide useful information, using recommended standards (RDF, SPARQL), and
- Include links to other URIs so that they can discover more things.

The linked data principles enable the implementation of generic applications that operate over the complete data space because the resulting web of linked data is based on standards for the identification of entities, retrieval of entity descriptions, and parsing of descriptions in RDF as a common data model. Linked data is not just a vision - the linked data principles are applied in various domains including e-science, libraries, and scholar communication.

A wide range of generic linked data tools, such as linked data browsers and linked data search engines already have been developed. What is still missing is the closer integration of linked data features into the scientific work environments that are used within the different scientific disciplines. From a system perspective, a data infrastructure, in order to support a linking data capability, should provide:

- a registry service whose purpose is to manage a collection of identifiers and make it actionable and interoperable, where that collection can include identifiers from many other controlled collections;

- a name resolution system, which resolves the data identifiers into the information necessary to locate and access them; and
- automated or semi-automated generation of links.

5.3.4 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. 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. 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 is marginalized, and the outcome is almost always flawed or less useful than originally anticipated (Edwards, Jackson, Bowker, & Knobel, 2007).

In the social sciences, one of the biggest worries is data privacy. The tension between individuals' interest in protecting their privacy and companies'/organizations' interest in exploiting personal information could be resolved by giving people more control. In the information realm, loss of privacy is usually associated with failure in controlling the access to data, the flow of data, or the purposes for which data are employed. Data infrastructures should support efficient privacy schemes in order to protect sensitive datasets, designing algorithms that satisfy a given privacy definition while producing useful outputs and control anti-social behaviors.

New research data infrastructures are encountering and often provoking a series of tensions. This occurs when established practices, organizational norms, and individual and institutional expectations adjust in a positive or negative fashion in reaction to the new possibilities and challenges posed by infrastructure. 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. (Edwards, Jackson, Bowker, & Knobel, 2007). 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. Tensions should be thought of as both barriers and resources to infrastructural development and should be engaged constructively.

6 TECHNOLOGICAL CHALLENGES

There are several technological challenges that must be tackled in order to be able to develop the future GRDIs. They include research data modeling and management as well as system challenges.

6.1 Data modeling challenges

There is a need for radically new approaches to research data modelling. In fact, the current data models (relational models) and management systems (relational database management systems) were developed by the database research community for business/commercial data applications. Research data have completely different characteristics from business/commercial data, and thus the current database technology is inadequate to handle them 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; and
- represent and query data quality information.

While most scientific users can use relational tables and have been forced to do so by current systems, we can find only a few users for whom tables are a natural data model that closely matches their data. Conventional tabular (relational) database systems are adequate for analysing objects (galaxies, spectra, proteins, events, etc.), but the support for time-sequence, spatial, text and other data types is awkward. For some scientific disciplines (astronomy, oceanography, fusion, and remote sensing) an array data model is more appropriate. Database systems have not traditionally supported science's core data type: the N-dimensional array. Simulating arrays on top of tables is difficult and results in poor performance. Some other disciplines, i.e., biology and genomics, consider graphs and sequences more appropriate for their needs. Lastly, solid modelling applications want a mesh data model. The net result is that "one size will not fit all", and science users will need a mix of specialized database management systems (Stonebraker, Becla, DeWitt, Lim, Maier, Ratzesberger, et al., 2009).

6.1.1 Metadata modeling

Metadata are the descriptive information about data that explain the measured attributes, their names, units, precision, accuracy, data layout, and ideally a great deal more. Most importantly, metadata include the data lineage that describes how the data were measured, acquired, or computed. The metadata are as valuable as the data themselves (Gray, Liu, Szalay, Nieto-Santisteban, DeWitt, & Heber, 2005). If the data are to be analysed by generic tools, the tools need to "understand" the data. The tool will want to know the metadata. If scientists are to read data collected by others, then the data must be carefully documented and must be published in forms that allow easy access and automated manipulation. In the next generation of data infrastructures, there will be powerful tools to make it easy to capture, organize, analyse, visualize, and publish data. The tools will do data mining and machine learning on the data and will make it easy to script workflows and analyse the data. Good metadata for the inputs are essential to make these tools automatic. Preserving and augmenting these metadata as part of the processing (data lineage) will be a key benefit for next generation tools.

All the derived data that the scientist produces must also be carefully documented and published in forms that allow easy access. Ideally, much of these metadata would be automatically generated and managed as part of the workflow, reducing the scientist's intellectual burden. The use of purpose-oriented descriptive data models is of paramount importance to achieve data usability. The type of descriptive information to be provided by the data producer depends very much on the requirements imposed by the data consumer tasks. For example, if the consumer entity wants to perform a data analysis task on the imported information, then the quality of the information is of paramount importance; without such information the task of data analysis cannot be performed. Consequently, if a researcher is willing to export/publish the data produced, its possible uses by the potential users must be carefully taken into account, and it must be endowed with appropriate descriptive information. Appropriate purpose-oriented metadata models to represent the descriptive information must be chosen and used.

6.1.2 Data provenance modeling

In its most general form, provenance (also sometimes called lineage) captures where data came from and how they have been updated over time. Provenance can serve a number of important functions (Ikeda, & Widom, 2010):
Explanation: Users may be particularly interested in or wary of specific portions of a derived data set. Provenance supports "drilling down" to examine the sources and the evolution of data elements of interest, enabling a deeper understanding of the data.
Verification: Derived data may appear suspect due to possible bugs in data processing and manipulation, due to stale data, or even due to maliciousness. Provenance enables auditing on how data were produced, either to verify their correctness or to identify the erroneous or out-dated source data or processing nodes that are responsible for erroneous or out-dated data.
Recomputation / Repeatability: Having found out-dated or incorrect source data or buggy processing nodes, users may want to correct the errors and propagate the corrections forward to all "downstream" data that are affected. Provenance helps to re-compute only those data elements that are affected by the corrections.

There has been a large body of very interesting work in lineage and provenance over the past two decades. Nevertheless, there are still many limitations and open areas. Specifically, the primary focus is on *modelling and capturing provenance*: How is provenance information represented? How is it generated? There has been considerably less work on *querying provenance*: What can we do with provenance information once we've captured it? In the long-term, the development of a standard open representation and query model is necessary. It is worthwhile to mention the "Open Provenance Model" (Moreau, Freire, Futrelle, McGrath, Myers, & Paulson, 2008), a community-driven model for provenance that allows provenance to be exchanged between systems.

6.1.3 Data context modeling

Context is a poorly used source of information in our computing environments. As a result, we have an impoverished understanding of what context is and how it can be used. Contextual information is any information that can be used to characterize the situation of a digital information object. In essence, this information documents the relationship of the data to their environment. Context is the set of all contextual information that can be used to characterize the situation of a digital information object. Several context modelling approaches exist and are classified by the scheme of data structures that are used to exchange contextual information in the respective system (Strang, & Linnhoff-Poppien, 2004): key-value models, mark-up scheme models, object oriented models, logic based models, and ontology based models. Future research data infrastructures should be context-aware, i.e., they should use context to provide relevant information and/or services to the user, where relevancy depends on the user's task.

6.1.4 Data uncertainty modeling

As models of the real world, scientific databases are often permeated with forms of uncertainty, including *imprecision, incompleteness, vagueness, inconsistency, and ambiguity*. Uncertainty is the quantitative estimation of error presenting data; all measurements contain some uncertainty generated through systematic error and/or random error. Acknowledging the uncertainty of data is an important component of reporting the results of scientific investigation (http://visionlearning.com/en/library/Process-of-Science/49/Uncertainty-Error-and-Confidence/157). Essentially, all scientific data are imprecise, and without exception science researchers have requested a database management system that supports uncertain data elements. Of course, current commercial products do not support uncertainty. There is, among science users, a universal consensus on requirements in this area. Some of them request a simple model of uncertainty while others request a more sophisticated model. There has been a significant amount of work in areas variously known as "*uncertain, probabilistic, fuzzy, approximate, incomplete, and imprecise*" data management. Undoubtedly, the development of suitable database theory to deal with uncertain database information remains a challenge that has yet to be met.

6.1.5 Data quality modeling

Quality of data is a complex concept, the definition of which is not straightforward. There is no common or agreed definition or measure for data quality apart from such general notions as *fitness for use*. The consequences of poor data quality are often experienced in every scientific discipline but without making the necessary connections to its causes (Batini & Scannapieco, 2006). Awareness of the importance of improving the quality of data is increasing in all scientific fields. In order to fully understand the concept, researchers have traditionally identified a number of specific quality *dimensions*. A dimension or characteristic captures a specific facet of quality. The more commonly referenced dimensions include *accuracy, completeness, and consistency*. An important aspect of data is how often they vary in time - *stable* data and *time-variable* data. In order to capture aspects concerning temporal variability of data, different data quality dimensions need to be introduced. The principal time-related dimensions are *currency, timeliness, and volatility.*

Data quality dimensions are not independent of each other, but correlations exist among them. If one dimension is considered more important than the others for a specific application, then the choice of favouring it may imply negative consequences for the others. Establishing trade-offs among dimensions is an interesting problem. The core set of data quality dimensions introduced here is shared by most proposals for data quality dimensions in research literature. However, for specific categories of data and for specific scientific disciplines, it may be appropriate to have more specific sets of dimensions. As an example, for geographic information systems specific, standard sets of data quality dimensions are under investigation (ISO, 2005).

6.2 Research data management challenges

In the area of research data management, data inputs in the order of multiple petabytes are expected in the near future. The current database technology is not suitable to fulfill the demanding requirements of research data management. The database technology has evolved having in mind a business paradigm. The users of current science databases have several problems with the current relational database technology (Becla & Lim, 2008): (i) research data very rarely naturally fit into a relational table-based model; (ii) the most frequently performed operations are not supported, and complex analytics such as time series analysis are impossible to express in SQL; (iii) provenance information is not supported; (iv) uncertainty management is not supported; and (v) there is insufficient scalability as none of the existing DBMSs offers multi-petabyte level scalability.

Some directions of research in database technology are indicated in (Kersten, Idreos, Manegold, & Liarou, 2011): (i) array support: DBMSs should support multi-dimensional arrays as first-class citizens next to relational tables; (ii) one-minute database kernels: they should be able to identify and avoid performance degradation by answering queries only partially within strict time bounds; (iii) multi-scale query processing: in order to be able to quickly explore large datasets, it is necessary to partition the database according to scientific interest and resource availability; (iv) post-processing result sets: the often huge results returned should not be thrown at the user directly but passed through an analytical processing pipeline to condense the information for human consumption; (v) query morphing: given the imprecision of the queries, the system should aid in hinting at proximity results using data distributions looked upon during query evaluation; (vi) queries as answers: the database system should be able to provide starting points and guidance for data exploration instead of returning results for a random or badly formulated query, and it should return query suggestions for more effective exploration; and (vii) the system should support transparent data ingestion from and seamless integration of scientific file repositories.

It is worthwhile to mention the efforts towards the development of a scientific database management system, SciDB. This is an open-source DBMS oriented towards the data management needs of scientists. It mixes statistical and linear algebra operations with data management ones, using a natural nested multi-dimensional array data model. SciDB is used by the large synoptic survey telescope (LSST). Another important effort towards the development of a scientific DBMS is the MonetDB. It is an open source column-oriented database management system. It was designed to provide high performance on complex queries against large databases, e.g., combining tables with hundreds of columns and multi-million rows.

6.3 Infrastructural management challenges

A research data infrastructure, in order to be able to efficiently and effectively play its role as has been described in Section 4, must also tackle some management challenges, which include ontology, workflow, and policy management.

6.3.1 Ontology management

Ontologies/taxonomies constitute a key technology enabling a wide range of science ecosystem services. In fact, ontologies provide the semantic underpinning enabling intelligent search, access, integration, sharing, and use of research data. The services to be provided by a research data infrastructure require extensive use of ontologies and also include data/information discovery, data service/tool discovery, data classification, data exchangeability and interoperability, dealing with inconsistent information, etc. In addition, in several communities of research, ontologies are considered to be the ideal formal tool to provide a shared conceptualization of the domain of interest. Ontologies were initially developed by the artificial intelligence community to facilitate knowledge sharing and reuse. An ontology consists of a set of concepts, axioms, and relationships that describe a domain of interest. Currently, ontologies are produced in large numbers and exhibit great complexity. Therefore, ontology management is a needed capability of a global research data infrastructure in order to be able to effectively support the semantic science ecosystem services.

There are three main components of an ontology management capability: ontology model, ontology metadata, and reasoning engine (Bloehdorn, Haase, Huang, Sure, Voelker, van Hermalen, et al., 2009). The ontology model maintains a set of references to the ontologies, their respective contents, and corresponding instance data sources.

The ontology metadata store maintains metadata information about the ontologies themselves. The reasoning engine operates on top of the ontology model, giving the set of axioms an interpretation and thus deducing new facts from the given primitives.

In a science ecosystem, individual data sources, service functionalities, etc. are often described based on different, heterogeneous ontologies. To enable interoperability across these sources, it is necessary to specify how the resource residing at a particular node corresponds to a resource residing at another node. This is formally done using the notion of mapping. Although in the area of ontology languages the Web Ontology Language (OWL) has become a de facto standard for representing and using ontologies, there is no agreement yet on the nature and the right formalism for defining mappings between ontologies.

In the context of a science ecosystem, ontologies are not standalone artifacts. They relate to each other in ways that might affect their meaning, and are inherently distributed in a network of interlinked semantic resources, taking into account in particular their dynamics, modularity, and contextual dependencies. It is important to investigate the entire development and evolution lifecycle of networked ontologies that enable complex, semantic applications. For each science ecosystem, we envision the need for building a top-level ontology, domain-independent, supplemented by several domain ontologies, one for each community of research belonging to the same science ecosystem. The top-level ontology would confine itself to the specification of such high level general (domain-independent) categories as: time, space, inherence, instantiation, identity, measure, quantity, functional dependence, process, event, attribute, boundary, and so on. The top-level ontology would be designed to serve as a common neutral backbone (Smith, 2003). It would be supplemented by the work of ontologists working on more specialized domains, e.g., the domain ontologies. Such ontologies should extend or specify the top-level ontology with axioms and definitions to the objects in some given domain. Each community of research should develop its own domain ontology. A research data infrastructure should have the capability of managing and maintaining the alignment of the top-level and domain ontologies of the science ecosystem.

6.3.2 Scientific workflow management

Today, research collaborations are becoming more and more geographically dispersed and often exploit heterogeneous tools, compare data from different sources, and use machines distributed across several institutions throughout the world. Therefore, the task of running and coordinating a scientific application across several administrative domains remains extremely complex. Scientific workflow is a key component in a research data infrastructure as it orchestrates e-science services so that they co-operate to implement efficiently a scientific application. A workflow is a precise description of a scientific procedure – a multi-step process to coordinate multiple tasks, acting like a sophisticated script. Each task represents the execution of a computational process, such as running a program, submitting a query to a database, submitting a job to a compute cloud or grid, or invoking a service over the Web to use a remote resource. Data output from one task is consumed by subsequent tasks according to a predefined graph topology that "orchestrates" the flow of data.

Workflow systems generally have three components: an execution platform, a visual design suite, and a development kit (Goble & De Roure, 2009). The platform executes the workflow on behalf of applications and handles common crosscutting concerns. The design suite provides a visual scripting application for authoring and sharing workflows and preparing the components that are to be incorporated as executable steps. The development kit enables developers to extend the capabilities of the system and enables workflows to be embedded into applications, Web portals, or databases. Scientific workflows liberate scientists from the drudgery of routine data processing so they can concentrate on scientific discovery. They shoulder the burden of routine tasks, they represent the computational protocols needed to undertake data-centric science, and they open up the use of processes and data resources to a much wider group of scientists and scientific application developers.

Workflows offer techniques to support the new paradigm of data-centric science. In fact, as first class citizens in data-centric science, they can be generated and transformed dynamically to meet the requirements at hand. In a landscape of data in considerable flux, workflows provide robustness, accountability, and full auditing. Workflows enable data-centric science to be a collaborative endeavor on multiple levels. They enable scientists to collaborate over shared data and shared services, and they grant non-developers access to sophisticated code and applications without the need to install and operate them. Consequently, scientists can use the best applications, not just the ones with which they are familiar. Multidisciplinary workflows promote even broader collaboration. In this sense, a

workflow system is a framework for reusing a community's tools and datasets that represent the original codes and overcomes diverse coding styles. Although the impact of workflow tools on data-centric science is potentially profound, many challenges exist over and above the engineering issues inherent in large-scale distributed software. There are a confusing number of workflow platforms with various capabilities and purposes and little compliance with standards.

6.3.3 Policy management

The need for using semantic policies in science ecosystem environments is widely recognized. In fact, the interactions among the different components of a science ecosystem should be governed by formal semantic policies. It is important to adopt a broad notion of policy, encompassing not only access control policies but also trust, quality of service, and others. Policies are means to dynamically regulate the behavior of system components without changing code and without requiring the consent or cooperation of the components being governed (Damianou, Dulay, Lupu, & Sloman, 2001). Policies that constrain the behavior of system components are becoming an increasingly popular approach to the dynamic adjustability of applications in academia and industry. Policies are pervasive in distributed and networked environments, for example in Web and Grid applications. They play crucial roles in enhancing security, privacy, and usability of distributed services.

Multiple approaches for policy specification have been proposed that range from formal policy languages that can be processed and interpreted easily and directly by a computer to rule-based policy notation using if-then-else format and to the representation of policies as entries in a table consisting of multiple attributes (Bonatti & Olmedilla, 2007). There are also ongoing standardization efforts toward common policy information models and frameworks. Policy language provides a framework for specifying both authorization policies and obligation policies. It is worthwhile to mention some significant policy specification tools, such as KAoS, Ponder, and Rei. All the different kinds of policies should eventually be integrated into a single coherent framework so that (i) this policy framework can be implemented and maintained by a research data infrastructure and (ii) the policies themselves can be harmonized and synchronized.

7 RECOMMENDATIONS

Future research data infrastructures must enable science ecosystems.
Several discipline-specific digital data libraries, digital data archives, and digital research libraries are under development or will be developed in the near future. These systems must be able to interwork and constitute disciplinary and/or multidisciplinary ecosystems. The next generation of global research data infrastructures must enable the creation of efficient and effective science ecosystems.

Science social and 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.
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 critical aspects is ignored. New research data infrastructures are encountering and often provoking a series of tensions. Tensions should be thought of as both barriers and resources to infrastructural development and should be engaged constructively.

Global research data infrastructures must be based on scientifically sound foundations.
It is widely recognized that current database technology is not adequate to support data-intensive multidisciplinary research. Existing research data infrastructures suffer from the following main limitations:

- not based on scientifically sound foundations,
- application-specific software of limited long term value coupled with the absence of a consistent computer science perspective, and
- discipline-specific.

Science re-builds rather than re-uses software and has not yet come up with a set of common requirements. It is time to develop the theoretical foundations of research data infrastructures. They will allow the development of generic data infrastructure technology and incorporate it into industrial-strength systems.

Formal models and query languages for data, metadata, provenance, context, uncertainty, and quality must be defined and implemented.

Radically new approaches to scientific data modeling are required. In fact, the data models (relational models) developed by the database research community are appropriate for business/commercial data applications. Scientific data has completely different characteristics from business/commercial data, and therefore current database technology is inadequate to handle it efficiently and effectively. 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, and represent and query data quality information are necessary. Formally defined data models and data languages will allow the development of automatized data tools and services (i.e., mediation software, curation, etc.) as well as generic software.

New advanced data tools (data analysis, massive data mining, data visualization) must be developed.

Current data management tools as well as data tools are completely inadequate for most science disciplines (Gray, 2009). It is essential to build better tools and services in order to make scientists more productive, tools helping them to capture, curate, analyze, and visualize their data, in essence tools and services that support the whole research cycle. Advanced tools and services are needed so as to enable scientists to follow new paths, try new techniques, build new models, and test them in new ways that facilitate innovative multidisciplinary/interdisciplinary activities.

New advanced infrastructural services (data findability, data tool findability, data federation (integration/aggregation/correlation/harmonization/linking) as well as infrastructural management services (workflow management, ontology/taxonomy management, policy management, etc.) must be developed.

The ultimate aim of a global research data infrastructure is to enable global collaboration in key areas of science. Therefore, the infrastructural services must achieve the conditions needed to facilitate effective collaboration among geographically and institutionally separated communities of research. To this end a global research data infrastructure must provide advanced support services that make the components of a science ecosystem interoperable and their holdings discoverable and usable.

Future research data infrastructures must support open linked data spaces.

A research data infrastructure must lower the barrier to publishing and accessing data leading, therefore, to the creation of open scientific data spaces by connecting data sets from diverse domains, disciplines, regions, and nations. Researchers should be able to navigate along links into related data sets.

Future research data infrastructures must support interoperation between science data and literature.

In the future all scientific literature and data will be on-line. Scientific data must be unified with literature to create a world in which the data and the literature interoperate with each other. Such a capability will increase the "information velocity" of the sciences and will improve the scientific productivity of researchers. Future research data infrastructures must make this happen by supporting the interoperation between data centers and research libraries.

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.

There is an emerging consensus among the members of the academic research community that "e-science" practices should be congruent with "open science". The open science principle entails not only open access to data but also to scientific analyses, methods, etc. This principle has not only a technological dimension but also policy and legal dimensions. Policies and laws must deal with legal jurisdictional boundaries and they must be integrated into a shared science policy framework.

A new international research community must be created.

The building of scientifically sound research data infrastructures can only be achieved if supported by an active new international research community capable of tackling all the scientific and technological challenges that such an enterprise implies. This community would embrace two main components:

- researchers who use data-intensive methods and tools (biologists, astronomers, etc.) and
- researchers who create or enable these models and methods (computer scientists, mathematicians, engineers, etc.).

So far, these two components have operated in isolation and have only come together sporadically. We firmly believe that the development of the new data-intensive multidisciplinary science must spring from a synergetic action between these two components. We firmly believe that without the creation and active involvement of such a research community it is illusory to think about the development of the data-intensive multidisciplinary science.

New professional profiles must be created.

In order to be able to exploit the huge volumes of data available and the expected new data and network technologies, new professional profiles must be created: data scientist, data-intensive distributed computation engineer, data curator, data archivist, and data librarian. These professionals must be capable of operating in the fast moving world of network and data technologies. Education and training activities to enable them to use and manage the data and the infrastructures must be defined and put in action.

8 ACKNOWLEDGEMENTS

This work has been funded by the Commission of the European Union under the Seventh Framework Program (FP7) – Infrastructures, "Capacities – Research Infrastructures" – Project Title: GRDI2020, Number: 246682

9 RERERENCES

Altman, M. & King, G. (2007) A Proposed Standard for the Scholarly Citation of Quantitative Data. *D-Lib Magazine 13*(3/4).

Adomavicius, G., Bockstedt, J., Gupta, A., & Kauffman, R. (2006) Understanding Patterns of Technology Evolution: An Ecosystem Perspective. *Proc. of the 39th Annual Hawaii International Conference* (*HICSS'06*), Hawaii, US.

Bannon, L. & Bodker, S. (1997) Constructing Common Information Spaces. *Proceedings of the Fifth European Conference on Computer-Supported Cooperative Work* (*ECSCW'97),* Lancaster, UK.

Batini, C. & Scannapieco, M. (2006) *Data Quality: Concepts, Methodologies, and Techniques*, New York, US: Springer.

Becla, J. & Lim, K-T. (2008) Report from the 2nd Workshop on Extremely Large Databases (XLDB). *Data Science Journal 7*, pp 196-208.

Bizer, C., Heath, T., & Berners-Lee, T. (2009) Linked Data – The Story so Far. Special Issue of the *International Journal on Semantic Web and Information Systems 5*(3), pp 1–22.

Bleiholder, J. & Naumann, F. (2008) Data Fusion. *ACM Computing Surveys 41*(1).

Bloehdorn, S., Haase, P., Huang, Z., Sure, Y., Voelker, J., van Hermalen, F., & Studer, R. (2009) Ontology Management. In Davies J. F., Grobelnik, M., & Mladenic, D. (Eds.), *Semantic Knowledge Management*, Heidelberg, Germany: Springer-Verlag.

Bonatti, P. & Olmedilla, D. (2007) Rule-Based Policy Representation and Reasoning for the Semantic Web. Antoniou, G. et al. (Eds.) *Reasoning Web 2007*, Heidelberg, Germany: Springer-Verlag.

Damianou, N., Dulay, N., Lupu, E., & Sloman, M. (2001) The Ponder Policy Specification Language. *Proceedings of Workshop on Policies for Distributed Systems and Networks*, Bristol, UK .

Edwards, P., Jackson, S., Bowker, G., & Knobel, C. (Eds.) (2007) Understanding Infrastructure: Dynamics, Tensions, and Design. Final Report of the *Workshop on History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures*, Ann Arbor, MI: National Science Foundation, Grant # 0630263.

Franklin, M., Halevy, A., & Maier, D. (2005) From Databases to Dataspaces: A New Abstraction for Information Management. *ACM SIGMOD Record 34*(4), pp 27-33.

Goble, C. & De Roure, D. (2009) The Impact of Workflows on the Data-centric Research. In Hey, T., Tansley, S., & Tolle, K. (Eds.), *The Fourth Paradigm: Data Intensive Scientific Discovery*, Redmond, WA: Microsoft Research.

Gray, J. (2009) Jim Gray on eScience: A Transformed Scientific Method. In Hey, T., Tansley, S., & Tolle, K. (Eds.), *The Fourth Paradigm: Data Intensive Scientific Discovery*. Redmond, WA: Microsoft Research.

Gray, J., Liu, D., Szalay, A., Nieto-Santisteban, M., DeWitt, D., & Heber, G. (2005) Scientific Data Management in the Coming Decade. *Technical Report, MSR-TR 2005-10*, Microsoft Research, Redmond, WA.

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

Iansiti, M. & Levien, R. (2002) The New Operational Dynamics of Business Ecosystems: Implications for Policy, Operations and Technology Strategy. *Working Paper 03-030*, Harvard Business School, Cambridge, MA.

Iansiti, M. & Levien, R. (2004) Strategy as Ecology. *Harvard Business Review 8*(3).

Ikeda, R. & Widom, J. (2010) Panda: A System for Provenance and Data. *IEEE Data Engineering Bulletin, Special Issue on Data Provenance 33*(3), pp 42-49.

Jewett, T. & Kling, R. (1991) The Dynamics of Computerization in a Social Science Research Team: a Case Study of Infrastructure, Strategies, and Skills. *Social Science Computer Review 9*(2), pp 246-275.

Keller, U., Lara, U., Lausen, H., Polleres, A., & Fensel, D. (2005) Automatic Location of Services. *Proceedings of the 2nd European Semantic Web Conference (ESWC)*, Heraklion, Greece.

Kersten, M., Idreos, S., Manegold, S., & Liarou, E. (2011) The Researcher's Guide to the Data Deluge: Querying a Scientific Database in Just a Few Seconds. *Proceedings of the VLDB Endowment 4*(9).

Lenzerini, M. (2002) Data Integration: A Theoretical Perspective. *Proceedings of Symposium on Principles of Database Systems (PODS)*, Madison, Wisconsin US.

Moreau, L., Freire, J., Futrelle, J., McGrath, R., Myers, J., & Paulson, P. (2008) The Open Provenance Model: An Overview. *Proceedings of the International Provenance and Annotation Workshop (IPAW)*, Salt-Lake City, Utah, US.

Paskin, N. (2004) Digital Object Identifier for Scientific Data. *The 19th International CODATA Conference*, Berlin, Germany.

Smith, B. (2003) Ontology. In Floridi, L. (Ed.), *Philosophy of Computing and Information*, Oxford, UK: Wiley Blackwell.

Star, S. L. (1989) The Structure of Ill-Structured Solutions: Boundary Objects and Heterogeneous Distributed Problem Solving. *Distributed Artificial Intelligence 2*, San Francisco, CA: Morgan Kauffmann Publishers.

Stollberg, M., Cimpian, E., Mocan, A., & Fensel, D. (2006) A Semantic Web Mediation Architecture. *Proc. of the 1st Canadian Semantic Web Working Symposium (CSWWS 2006)*, Quebec City, Canada.

Stonebraker, M., Becla, J., DeWitt, D., Lim, K., Maier, D., Ratzesberger, O., & Zdonik, S. (2009) Requirements for Science Data Bases and SciDB. *Proceedings of Conference on Innovative Data Systems Research (CIDR 2009),* Asilomar, California, US.

Strang, T. & Linnhoff-Poppien, C. (2004) A Context Modeling Survey. Workshop on Advanced Context Modeling, Reasoning and Management associated with the *Sixth International Conference on Ubiquitous Computing (UbiComp 2004)*, Nottingham, UK.

(Article history: Received 5 December 2012, Accepted 1 September 2013, Available online 19 September 2013)