GRDI2020

A Vision for Global Research Data Infrastructures

Global Research Data Infrastructures:
Towards a 10-year vision for global research data infrastructures

Final Roadmap

www.grdi2020.eu | info@GRDI2020.eu

Funded under the Seventh Framework Programme (FP7) - Infrastructures, "Capacities – Research Infrastructures" - Project Number: 246682

CAPACITIES

European Commission

Final Roadmap

The GRDI2020 10-year visionary roadmap has been compiled within the framework of the GRDI2020 (www.grdi2020.eu) project funded under the 7th Framework Programme, Capacities – GÉANT & eInfrastructures.

Principal author: Costantino Thanos, CNR-ISTI, Italy

Editor: Richard Hudson, CEO & Editor Science | Business.

Acknowledgments

The GRDI2020 Consortium would like to thank all those who have provided feedback and suggestions to the roadmap, in particular;

Malcolm Atkinson, GRDI2020 Advisory Board Member

Dan Atkins, Univ. of Michigan, Vice-President for Research Cyber-infrastructure & GRDI2020 Advisory Board Member

Antonella Fresa, DC-Net Project Coordinator, Italy

Fabrizio Gagliardi, EMEA Director, Microsoft Research Connections (External Research), Microsoft Research, UK & GRDI2020 Advisory Board Member

David Giaretta, STFC and Alliance for Permanent Access, United Kingdom & HLG-SDI Rapporteur*

Stephen M. Griffin, Program Director Information Integration and Informatics (III) cluster, National Science Foundation, US

Ray Harris, Emeritus Professor, University College London & GRDI2020 Advisory Board Member

Jane Hunter, University of Queensland, Australia

Simon Lin, Academia Sineca & GRDI2020 Advisory Board Member

Monica Marinucci , Director, Oracle Public Sector, Education & Research Business Unit & HLG-SDI Member*

Reagan Moore, Director of the Data Intensive Cyber Environments (DICE Center) & Prof. in School of Library Information Science at the Univ. of North Carolina at Chapel Hill, US

Vanderlei Perez Canhos, President Director of the Reference Center on Environmental Information (CRIA), Brazil & GRDI2020 Advisory Board Member

Laurent Romary, INRIA & Humboldt University & HLG-SDI Member*

Michael Wilson, e-Science Department, STFC Rutherford Appleton Laboratory, United Kingdom

Peter Wittenburg, Technical Director, Max Planck Institute for Psycholinguistics, The Netherlands & HLG-SDI Member*

**European Commission, Directorate General Information Society & Media High Level Expert Group on Scientific Data*

The GRDI2020 Consortium is composed of Trust-IT Services Ltd. (UK), Consiglio Nazionale delle Ricerche, - Istituto di Scienza e Tecnologie dell' Informazione A. Faedo (CNR-ISTI) (IT), ATHENA Research & Innovation Center in Information Communication & Knowledge Technologies (GR) PDC Center for High Performance Computing, Kungliga Tekniska Hoegskolan (SE)

Executive summary

The nature of science is changing. It is more global. It is increasingly multi-disciplinary. And the raw stuff of science - the data - grows ever-more abundant. This data-intensity poses new opportunities and challenges across the scientific world. It is hard to manage, physically and financially. It is hard to handle conceptually, requiring a new way of looking at the role of data. Those countries and regions that adapt quickly to this new order will benefit; the rest risk being left behind in the race to solve our societal problems and spur economic growth. For the European Union, preparing for this scientific digital challenge is a vital matter of competitiveness.

At the heart of this new science will be the Global Research Data Infrastructure - the basic computer and communications highway to link researchers and their data across borders and disciplines. GRDI is not a new kind of computer system; it is an entire digital environment, spanning countries and scientific disciplines, that lets locally controlled data systems interoperate as seamlessly as possible. It is a ubiquitous, reliable, and widely shared resource. It goes beyond mere hardware or software to include the organisational practices of the people and institutions using it, their own infrastructure, and the social norms they use to do good science at a distance. It is an enabler of a digital science ecosystem, with a bewildering variety of interactions among data management systems, digital data libraries, research libraries, data collections, data tools, and communities of research. As in a natural ecosystem, so in this digital one many balances must be maintained. It must handle knowledge freely and dynamically, so researchers can strike a balance between competition and cooperation, hoarding or sharing data. It must accommodate both the massive research team that spans the globe, as well as the creative genius who works alone. And it must dynamically adapt to changing circumstances.

Technically, this is a tall order. Here, drawing on our technical knowledge, we provide a roadmap for the obstacles to be overcome, and the possible routes to be taken. Building the GRDI is a massive undertaking. As it gets underway, we urge the scientific, industrial and policy community to incorporate these attributes into GRDI:

1. **Enable digital science ecosystems.** As the key components of GRDI are integrated, the central concept of ecosystem must be supported: the individual, dynamic communities of research, with their associated tools, networks and practices.
2. **Take into account the way science is organised and run, and the new tensions that GRDI will create in the research world..** The enterprise of science is a complex affair, with elaborate modes of behaviour and interaction. GRDI will create some winners and losers, and these organisational aspects of science must be remembered.

3. **Base GRDI on scientifically sound foundations.** Existing data infrastructures in research lack scientifically sound foundations, often developed ad hoc for individual communities. It is time to develop the theoretical foundation of scientific data infrastructures.
4. **Define and implement better data models and query languages**. Too often, databases and data models in science are adapted from the business world, rather than tailored to science. New approaches are needed that more closely match the way scientists want to represent their information.
5. **Develop advanced data tools.** Researchers need better tools to help them capture, curate, analyse and visualise their data - to support the whole research cycle.
6. **Develop advanced infrastructure services.** Digital science ecosystems will need services as well as tools - services to help them find tools, discover data, integrate data sets or manage the scientific workflow.
7. **Support open linked data spaces.** Open, scientific data 'spaces' are needed to connect data sets from diverse domains, disciplines, regions and nations. Researchers should be able to navigate along links into related data sets.
8. **Support interoperability between science data and literature.** Scientific data must be unified with the literature to create a world in which the two can work together, referencing each other to speed "information velocity" and the productivity of researchers. Future scientific data infrastructures must make this happen by supporting interoperability between digital data libraries and digital research libraries.
9. **Recognise that open science requires open data.** There is an emerging consensus that e-science, to work, must be open - for researchers to join in the work, and to read and cite the work of others freely. An integrated science policy framework is needed to help manage this complex new world.
10. **Create a new international research community.** Scientifically sound data infrastructures can only be built if supported by an active international research community, capable of tackling all the scientific and technological challenges that such an enterprise implies.
11. **Create new professions.** To make best use of the huge volumes of data and network technologies we expect, new professional profiles must be created. These include data scientist, data-intensive distributed computation engineer, data curator, data archivist, and data librarian.

There are many other challenges ahead - political and organisational, as well as technical. This technology is disruptive: It will be a source of creative destruction as practises, organisations, norms, expectations and individuals bend to the new order. Some will thrive. Some will not. But the result will be a global research data infrastructure, to enable an open, extensible, reliable and evolvable 'ecosystem' for digital science. We foresee this leading science into new, as-yet unimaginable directions, empowering individuals from varied disciplines and nationalities to collaborate on solving society's greatest technical problems.

I. The data of science

In 1836, when Charles Darwin returned from his round-the-world voyage of discovery aboard HMS Beagle, his scientific haul included 770 pages of diary, 1,383 pages of geological notes, 1,529 specimens preserved in spirits and another 3,907 dried[1]. For that age, Darwin's data 'library' was astonishingly large, keeping him and his scientific colleagues in business for years.

But by today's standards, Darwin had barely begun. We know now of 10,000 types of bird and 380,000 plants – in all so far, 1.75 million different plant, animal and fungal species (and that's not even starting to count microorganisms[2]).

For this advancing frontier of knowledge, thousands of biologists across the world have amassed billions of data points. Then, to understand the spread or decline of species, they compare those data with the results of environmental monitoring stations, satellite images of land use, or databases of industrial activity. Then, they publish their findings to colleagues around the world, or share the raw data with some of them. And the information mountain grows higher.

All these data add up to a monumental challenge. Today, the biggest international effort analogous to Darwin's early collecting effort, the Global Biodiversity Information Facility[3], is developing a framework to link 10,000 databases around the world. It is a prime example of two major trends in science: **data intensity,** and **multi-disciplinarity**. For the first, biodiversity research bears the hallmarks: increasing volumes and sources of data, growing complexity of data and processing, a fast-changing and dynamic environment, a rich interaction between researchers and data, and high demand. For the second, multi-disciplinarity, it forces ecologists, biologists, meteorologists, agronomists and other specialists to work together – often across national, cultural and linguistic borders, as well as across disciplinary limits. From all this, an entirely new method for science is emerging, which some[4] have called a Fourth Paradigm – moving beyond the older scientific methods of observing, theorizing or simulating, and into a new process of looking for meaningful correlations across vast data sets. Correlation supersedes causation as the source of new knowledge; and science can advance without painstaking cause-and-effect models, grand theories or any mechanistic explanation at all[5].

Clearly, this is important. Those researchers who master these new, data-intensive and multi-disciplinary approaches will have a scientific edge; and those countries and regions that support them will gain as well. Getting this right isn't just a matter of good science; it will also underpin our economic competitiveness.

But it is also difficult. Research is now a business without borders: Scientists and engineers routinely collaborate across the globe, both in their own fields and with others. This has great benefits, but also some difficul-

ties. Take one seemingly trivial example: How many different ways is the word 'phase' used in how many sciences – in physics, psychology, astronomy, linguistics? To date, these differences haven't mattered much. But if a physicist in Geneva or Tokyo is trying to search a database maintained by astronomers in London, confusion can arise. How are we to design a search tool that can automatically broker all these different meanings? More broadly, how can we find a way to collect all the new data, work with them, analyse them, share them and then archive them for future use? And how to get this going not just in one lab or country, but across borders of nation, culture and, even, time? And how to make it affordable? "The potential benefits are enormous but the same is true for the costs," says Neelie Kroes, Vice President of the European Commission[6].

Getting answers to such questions is vital. Over the past decade, a growing number of advisory panels – in Brussels, Washington and beyond – have been sounding calls for regional and international action. They suggest ways to plan the support e-infrastructures, find the funding and start making this new, global, data-intensive science a matter of routine. But hitherto, many of these reports have taken the scientific user's point of view, elaborating on what such an e-infrastructure should do. This report takes a different approach: What can computer science offer to achieve the vision? Here, we start with the list of required uses as a given, and focus on the technical, organisational and policy challenges we will face in trying to build these systems. This is our special contribution to the global dialogue on data research infrastructures.

The result is a road map of the major technical trouble spots that we feel will have to be seen and conquered if we are to move forward. It results from a series of expert meetings in 2010 and 2011, organized for the European Commission by its Global Research Data Infrastructure 2020, or GRDI2020, project, part of the 7th Framework Programme. Two working groups, on technological and organisational/policy issues, developed their views in tandem, and are combined in this report. They form a research agenda, a statement by computer scientists on what questions must be answered to achieve a truly global data research infrastructure. Taken together, we believe, they constitute one of the grandest of Grand Challenges for Europe and global science.

II. How will the GRDI work?

We envision a global research data infrastructure, to enable an open, extensible, reliable and evolvable 'ecosystem' for digital science. We foresee this leading science into new, as-yet unimaginable directions, empowering individuals from varied disciplines and nationalities to collaborate on solving society's greatest technical problems.

What kind of ICT support do we imagine for this new data-intensive, multidisciplinary science?

Let's start with the most basic definition of what we are talking about, a global research data infrastructure or GRDI. First, GRDI is not a new kind of computer system; it is an entire digital environment or network, spanning countries and scientific disciplines, that lets locally controlled data systems interoperate as seamlessly as possible. It is an ubiquitous, reliable, and widely shared resource. It goes beyond mere hardware or software to include the organisational practices of the people and institutions using it, their own technical infrastructure, and the social structures and norms they use to do good science at a distance. It includes 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 data sets from diverse disciplines
* the management of scientific workflows
* the interoperation between scientific data and literature for the benefit of different communities of researchers involved in data-intensive multidisciplinary activities.

In essence, to the vast computing resources scattered across all the world's labs and technical facilities, GRDI adds reliable, efficient, and effective access. It is, in a sense, between everything - between the communities of researchers and the data and collections they use or supply. This 'betweenness' - or, less colloquially, the fundamental **relational** concept that underpins it - is its core characteristic. As others[7] have described it, a data infrastructure like this is embedded in the world of its users. It is transparent, invisibly supporting the work to be done. Its scope goes beyond that of a single project or event. It becomes part of the daily life of its users - taken for granted, almost as a form of membership, in the community that uses it. It shapes, and is shaped by, its users. It embodies technical standards, builds on the existing technology base of the group, and - alarmingly for those who rely upon it - becomes so much a part of life that when it

breaks, when a server crashes, when the power stops, life is disrupted and the invisible network suddenly becomes all too visible. Science today is an increasingly global undertaking; and that means, inevitably, that the data infrastructure to support it must also be global. The goal: to enable researchers to make the best use of the world's growing wealth of data.

Here is another way to think about it: As an ecosystem. In a natural ecosystem, varied species co-exist, influence one other, and interact with their environment. So in a digital ecosystem, a bewildering range of computer types, networks, libraries, instruments, architectures, tools and users interact. And somehow, they must all stay in balance. The digital science ecosystem must permit knowledge to flow freely and dynamically, so researchers can strike a balance between competition and cooperation, hoarding or sharing data. Its degree of connectivity – how easy or hard it is to link people and data – must be variable: connectivity, while often good, can be bad if it ends in 'group think' that stifles the creativity of a lone genius. The digital ecosystem must be able dynamically to adapt itself to changing circumstances and needs – yet still meet the recurring and basic needs of the research community. And the system must strike a balance between diversity, with lots of possible tools and solutions available, and stability: While feature-rich systems can be good, they are also expensive to maintain.

In building such a digital ecosystem, there are some main components:

Digital Data Libraries: Increasingly, the volume of data produced by high-throughput instruments and simulations is so large, and the application programs so complex, that it is cheaper to move the user's programs to the data than the data to the programs – that is, to keep most everything in a central library, while only the scientist's questions and answers move back and forth across the network. These central service stations are called Digital Data Libraries, or Science Data Centres[8]. Each of these libraries holds one or more massive datasets, manages the programs that provide access to the data, and provides staff who understand, add to and improve the data. DDLs can come in many flavours[9]. There are **Research DDLs** that hold the direct output of one or more focused research projects – typically, data without much processing. An example might be a DDL for environmental data on the Waddenzee in the Netherlands or Lake Como in Italy. **Discipline or Community DDLs** serve a larger group of scientists or engineers – for instance, providing a pan-European library for data on all major bodies of water. **Reference DDLs** are broader still, serving large segments of science or education - scientists, students, educators and others from a variety of disciplines, institutions and places; an example might be a comprehensive European environmental data centre. Finally, **Specialised Service DDLs** go beyond data to provide services – specific software tools to map all that aquatic and environmental data from across Europe, or look for correlations among the databases.

Digital Data Archives: As the name suggests, archives keep data and methods for the longer term, rather than active use. They can be for future reference or regulatory compliance; and they are more important in some fields than in others – for instance, a good chronological record of weather readings is vital for climatology. Indeed, the whole concept of digital data archiving is a modern

development in the history of science, made possible by advances in information technology.

Digital Research Libraries: These store the results of research, printed or not, in electronic format and organise them for long-term use – like a modern university library. Their mission is to acquire information, organise it, make it available and preserve it. And they require a long-term commitment – in financing, institutional backing and organisation[10].

All of these – digital data libraries and archives, and digital research libraries – operate in a dynamic, evolving landscape for **communities of research**. These communities are at the core of our vision of a global research data infrastructure. They speed scientific progress, by linking researchers and data in novel ways. They provide a provide a coherent voice for science, enabling their members to improve cooperation, quality control, standards development and validation[11]. They are the flesh-and-blood people in these abstract, open, extensible, and evolvable digital science ecosystems.

Further, each community has its own **ecosystem view** of the data it uses, the services it needs, and how it operates; and this view is reflected in the software, user interface and other tools the scientists in the community use daily. Each has its own:

* **Science Gateways,** a community-specific set of tools, applications and data collections that are integrated together and accessed through an online portal, or a suite of applications[12].
* **Virtual Research Environments**, the technological frameworks for data infrastructure into which all the tools, services and resources can be plugged. In essence, these are the virtual worlds in which the research communities operate.

Interconnecting all the ecosystem's components are a variety of **research channels** – networks, wireless links, or other pathways over which scientific data and work move. These channels can link different scientific disciplines, for instance connecting libraries for marine biologists and tropical storm researchers to study the interplay of weather and sea life. The channels can link data archives to libraries, moving information from long-term storage to immediate use. Or the channels can connect a published research report to the data, stored elsewhere, that underlie it. That would permit a scientist to read a paper and then look up the raw data – perhaps to double-check it, or redo the analysis completely. Or the researcher can work the other way: To look up some data and then check all the available literature that relates to it[13]. The result is that members of the research community can use the channels to exchange data without distorting the meaning – a particularly difficult task as science grows in size and complexity, and requiring new **mediation** technologies. It will reduce the fragmentation of science into micro-communities of specialists. It will, in essence, collapse time and space for research.

Continuing our GRDI map, we see the creation and maintenance of so-called **service environments** for each ecosystem. These include a series of tools so researchers can figure out what data they need, where to find it, and how to go get it. In the environment, they can register the data with unique identifiers, so they can cite specific

data sets in their work rather than published papers alone. And they can more easily copy from one database to another, send queries and get answers on remote databases, and manage their own workflows.

Ecosystem view, research channels, service environment – these are all aspects of digital life *inside* the research community. But lots of forces *outside* are also in play: social and governmental, economic and technical[14]. And these all affect the way the researchers think, work and cooperate. For a Global Research Data Infrastructure isn't just technology; it is also organisational practices and social norms for productive science. If these aren't handled right, the GRDI will fail[15].

Science is a complex human endeavour. It advances because scientific practises evolve, new ones building on old. Scientific revolutions can be seen as breakout moments, when old, well-worn paths of theory, data collection, and analysis are overturned. The development of global research data infrastructures may repr sent such a moment.

What is in a service environment?

The researchers in each digital science ecosystem need a wide range of tools and services to get their work down – taken together, a service environment. Examples of what is in it:

* Data Discovery, to quickly and accurately identify and find needed data.
* Data Registration, to identify data as unique pieces of work rather than indirect references in the footnotes of a scientific paper. Each data point or set of data would be assigned a Digital Object Identifier (DOI) to register it uniquely, so it can be found and used more easily by others. The DOI would need its own numbering syntax, a resolution service, a data model, and a way of implementing it all.
* Data Citation, to permit researchers to reference data online as easily as they do today for printed citations. For that, the environment for citation needs to provide a way to handle confidentiality, verification, authentication, access, technology changes, the practices of any particular scientific discipline, name resolution - and future extensions.
* Data Search, to more easily figure out what data a researcher needs and how to get it – across multiple data sources, in many steps, with continuous refinement of the search goal.
* Data Integration, to combine data from different sources and get a unified view of it all.
* Data Sharing, to share results with other community members.
* Data Linking, to connect data sets from diverse domains of the science ecosystem.
* Recommendations, to annotate or highlight datasets or tools that might interest other researchers.
* Ontologies and Taxonomies, to categorise data types and services in a standardised way so they can be easily found, searched and used by a wide range of researchers.
* Transportable Data, to copy data from one database to another. This would work like a shipping service, moving a package of objects from one site to another as fast as possible.
* Transportable Data Services and Tools, to operate on data and applications stored elsewhere. These would encapsulate computation into transportable objects that can be run on the remote data centres. These objects run like applications that are temporarily installed on a remote computer, perform an operation, and then are uninstalled.
* Scientific Workflow Management, to characterise and manage each step in their work, and thus coordinate it with that of other colleagues.
* Policy Management, to control who gets access to what data. This entails determining how sensitive particular data are, as well as how trustworthy a particular user can be assumed to be; and it requires a range of tools for editing, administration, graphical representation and other tasks.

III. GRDI – The Challenges

Building a Global Research Data Infrastructure will not be easy. There are challenges in the technology, organisation, system and policy. It will require several breakthroughs in data modelling, management and tools. The difficulty of all this is, we feel, often under-estimated in policy discussions. We are not pessimists; these can be overcome. But it is important to recognise what needs to be done.

It starts with top-level political and social issues. This technology has the potential to upset people – to be a source of creative destruction as practices, organisations, norms, expectations, and individuals bend to a new order. Some researchers will thrive on the unexpected insights, new colleagues and unfamiliar tools that appear; some will not. Some administrators and policy makers will encourage the many cross-disciplinary and cross-border collaborations; some not.

Emergent data infrastructures are redistribution mechanisms, reorganising resource flows across scales ranging from the local workplace or laboratory to the global economy. As a result, we already can identify provisional winners and losers. The losers are those individuals, groups, and forms of social or professional practice that fit uneasily or not at all within the emerging infrastructural paradigm; once excluded from the scientific mainstream, they may never recover – with long-term economic and social consequences. The winners may include those institutions and practises that already lead in status and power; they have the means to install and use the new systems most effectively, and thus lead their development. Thus, the losers can be orphaned by the new infrastructure; the winners empowered by it. How policy makers deal with this issue will be vital.

Another type of tension arises when the changing infrastructure bumps up against the constraints of political economy, such as public/private investment models, or legal systems. For instance, how will a world of sharing, collaborating and re-processing fit within the existing rules of intellectual property rights? How will this, in turn, fit with an emerging academic consensus that "e-science" practices should be congruent with "open science"? In our view, GRDI will enable open research - or, as the Organisation for Economic Cooperation and Development puts it, "access on equal terms for the international research community at the lowest possible cost, preferably at no more than the marginal cost of dissemination." This runs counter, of course, to many private and commercial publishing models today[16].

And how will GRDI affect the relationship between national policy objectives and the transnational pull of science? Many important policy interests – a country's economic competitiveness, security or scientific leadership – are determined nationally. But science, especially when amplified by GRDI, spills into the world at large, connecting researchers and communities anywhere. This already shows up in the complications we now see in the various co-funding arrangements across multiple national agencies. How much more difficult will national, and European Union, funding rules become as GRDI advances?

These are a few of the legal, economic and social policy challenges ahead. But there is another kind of 'policy' challenge as well: internal GRDI policy, in the broadest sense of the word, for who access is controlled, trustworthiness of users determined, or quality of service tailored to each user or use. These kinds of policies should be built into a single, coherent framework - an **Integrated Science Policy Framework.**

A host of other issues, more obviously linked to bits and bandwidth, need addressing, as well.

The Data Challenges

Modeling

Scientists have long suffered from a legacy of the history of computing: It wasn't meant for them, in the first place. Today, the most common type of database system, the relational database management system, was developed for business, not science. It was written to store information about customers, such as airline passengers – their names, addresses, credit card details, flight plans, seat assignments and the like. This works for some scientific purposes, like storing data on individual galaxies, spectra, proteins or events. But it isn't ideal for complex time-sequences, spatial relations or data relationships. It does not support the core scientific data type: the N-dimensional array; simulating arrays on top of conventional data tables is difficult. Graphs and sequences are important to biology and genomics; again, not easily managed in the standard commercial system. And solid modelling needs a 'mesh' data model. In short, the one-size-fits-all method of standard databases is not optimised for science; instead, a mix of specialised systems is needed.

There are other problems in data modelling. For instance, **metadata** are the descriptive information about the various scientific attributes being studied: for instance, names, units, precision, accuracy, and data layout. These attributes can be as valuable as the data themselves. But if scientists are to share more work on GRDI, their systems must preserve the metadata, generate new, and translate appropriately from one system to another. Data **provenance**, or lineage, is another challenge: Where did the information come from? How was it updated? Another problem: Data **context**. As humans, we share a great deal of unspoken knowledge about the context we work in; computers cannot. How do we retain that context as we work more and more through remote computer systems? The **uncertainty** of any data is also important to understand and retain – error, incompleteness, inconsistency, ambiguity; the typical commercial database simply cannot handle it. And then there is data **quality**: how 'good', useful, or appropriate are the data for the task in hand? Again, all these difficulties in scientific data modelling will have to be addressed if GRDI is to advance.

Management

Science needs well managed data – easy to get, find, store and analyse. But as data intensity rises, so do the challenges of management. **Curation of data**, for instance, is converting raw measurements to finished information: translating sensor information into standard data types, correcting for calibration information,

and the like. Some do this work inside the 'home' system; some do it 'off-site' on specialised facilities. Either way, with a GRDI, good communications between data gatherers and curators is going to be important. Another problem is **data integration** – combining data from different sources into a unified view. The applications for this have to be modelled correctly; researchers' queries must be processed correctly; and ways must be developed to handle inconsistent data sources, and to reason from the queries.

Linking data also becomes more important. This entails publishing data for the scientific community so that it can be machine-readable, clearly defined and linked to and from other data sets – or, to use a simplified analogy, rather like Web surfing. The goal: For a researcher to start browsing in one data set, and then navigate along the links to related data sets. Or for a search engine to crawl through the data space by following all the links. This is no small task.

For an individual researcher, the quality of data management is vital. If the data are well organised, documented and preserved, and if their accuracy and validity is controlled, the result is efficient research, solid evidence and convincing science. Indeed, even before researchers begin their work, they should draw up a data management plan for how their data will be managed during the project and shared with others afterwards[17].

Services and Tools

Better software and other data tools for science would make a difference: they would raise productivity, open new research paths, allow new kinds of modelling, and make multi-disciplinary work easier. For instance, they will need better **algorithms** to analyse extremely large data sets 'approximately' rather than exactly, to analyse problems with many processors at once, and to 'steer' long-running computations so priority can be given to the most important data. They will also need better ways of **visualising data**, so it is easier to validate models, interpret information, play 'what if' scenarios, form hypotheses and look at data from multiple perspectives. In many fields, data is arriving too fast to see or understand it. Brute force navigation through databases is not enough. Rather, a technological leap is needed to structure and prioritise information. New **data mining** techniques are needed to predict trends and behaviours. These must handle, not just batch or 'one-shot' data sets, but the continuous streams so common in astronomy, climatology and other data-intensive sciences.

The System Challenges

The system - the way all the components and users interact with one another - is a top-level development challenge for GRDI. As explained earlier, GRDI needs science gateways: the community-specific set of tools, applications and data collections that fit together and can be accessed through an online portal. And it needs Virtual Research Environments, the broad framework into which all the components are plugged.

With them go **interoperability and mediation** software. Interoperability, according to the Institute of

Electrical and Electronics Engineers, is "the ability of two or more systems or components to exchange information and to use the information." Sounds straightforward, but many problems hamper interoperability. *Heterogeneity* arises when the way a users asks for information differs from one system to another, when models differ for the way information objects are represented, or when the semantics of the systems differ. *Logical inconsistencies* can arise in the functional descriptions of the data-supplier's services and the user's requests. And *usage inconsistency* arises when the researcher simply cannot use the information that gets delivered - perhaps because the quality of the information retrieved isn't what the user needs, or the data are incomplete.

In practical terms, how do these interoperability problems get solved? A key is mediation software. In essence, this captures the user's requirements, then brokers an information 'deal' with the distributed information sources[18]. It maps the information structures, properties and relationships between the two systems, verifies whether the data strings actually match, and checks for consistency between the logical frameworks of the two systems. Mediation is nothing new, of course. XML, RDF and other well-known information models are already used on the Web to exchange all kinds of information; and several discipline-specific models have been developed. But for GRDI to work, an **integrated mediation framework** is needed to handle all kinds of heterogeneities and inconsistencies[19].

A bit of good news is that, already in the marketplace, some useful paradigms of programming and networking are emerging that could make life easier for GRDI. One is the programming concept of **scientific workflow**. This is a precise description of a scientific procedure – a multi-step process to coordinate multiple tasks, like a script[20]. Each task in the script is a specific computational process, such as running a programme, submitting a query, or using a service on the Web. The data output from one task feeds into the next task, all according to the script. This approach can be used for repetitive but vital scientific tasks - such as working with the torrents of data from a synchrotron. It automates what the researcher would do with the data, freeing science from data drudgery.

Another useful paradigm is **cloud computing**. This is a model, now widely promoted in the ICT industry, to organise large-scale distributed computing. In essence, it takes the key data, storage, software and services that a user needs, and pools and shares them over the Internet - in the 'cloud', rather than on the desktop. When somebody needs to find a fact, draw a graph, analyse a problem, they connect to the cloud. It is massively scalable. It delivers different levels of service to different users; and these services can be configured on the fly. It is driven by economies of scale: The more users, the cheaper per user it becomes. We envision that the future Digital Data Libraries of GRDI, mentioned earlier, will be based on cloud philosophy and technology. Each scientific community of practice will have its own cloud or clouds; the federation of these clouds will allow collaboration among communities of practice. In short, it is a route to more multidisciplinary science.

MapReduce is another helpful paradigm. This is a programming framework to break big problems down into sub-problems that can be distributed to other computers to work on. As they produce

their answers, the results are then re-assembled to give the big answer. For instance, a social historian might ask how many times did the politically charged phrase 'the social question' appear in popular publications before, during and after Europe's 1848 revolutions? Clearly, multiple sources and multiple tasks are required to answer that seemingly straightforward question. MapReduce is one way to handle it, hiding the messy details of parallelisation, fault-tolerance, data distribution, and load balancing that, in this example, the social historian does not want to hear about. It would be suitable for many of the large computational problems encountered in data-intensive scientific applications[21].

The Organisational Challenges

Science is an increasingly complex business - and GRDI must make it easier. In the standard pipeline, science starts with a researcher gathering raw data, analysing to create results data, evaluating to understand and report, then pushing through the long and tortuous publication process. Then, the data end up in archives. All these steps are reflected in the basic GRDI components discussed earlier: The Digital Data Libraries to hold the primary information, Digital Data Archives to preserve the information long-term, and Digital Research Libraries to manage the published results. For GRDI to work, the components must be interoperable. This is no small task, given the fact that they are, and probably always will be, owned, managed and funded by different organisations. Again, this is why a broad, international framework for GRDI is needed.

But there are other considerations. For instance, two relatively new professions will become important for GRDI: the data archivist and the data curator. The archivist needs to work with the researchers producing data to be sure they can be easily archived and retrieved - for instance, generating the metadata specific to different disciplines and tasks. The data curator works on the information itself, annotating it, linking it to other research, continually reviewing it. They should also take an active role in promoting and adding value to the holdings. The relationship between constantly curated, evolving datasets and those in the digital archives needs to be explored, through research and accumulation of practical experience.

GRDI - EU projects already underway

There are already many e-Infrastructure projects underway in Europe that, taken together, are attempting to address many of the GRDI challenges we describe. Here is a partial list:

CLARIN (Common Language Resources and Technology Infrastructure) aims to provide a research infrastructure that offers single-point access to data, tools, resources and human expertise pertaining to language.

DARIAH (Digital Research Infrastructures for the Arts and Humanities) aims to improve research opportunities and outcomes through linking digital source material of many kinds, and to exchange knowledge expertise, methodologies and practices across domains and disciplines.

ECRIN (European Clinical Research Infrastructures Network) aims to develop and provide a European Infrastructure for clinical research. ECRIN supports multinational clinical trials.

BBMRI (Biobanking and Biomolecular Resources Research Infrastructure) aims to prepare and operate a pan-European infrastructure for biomedical and biological research, providing and managing access to biological resources required for health-related research in Europe.

CESSDA (Council of European Social Science Data Archives) aims to develop open, extensible, and evolvable networked data collections as well as a new generation of tools and services for data discovery, integration, visualisation, analysis and preservation.

ELIXIR (European Life Sciences Infrastructure for Biological Information) aims to construct and operate a sustainable infrastructure for biological information in Europe, to support life science research and its translation to medicine and the environment, the bio-industries and society.

D4Science-II aims to provide mechanisms for facilitating interoperation of the D4Science e-Infrastructure with diverse other data e-Infrastructures that are running autonomously, thus creating the core of an e-Infrastructure Ecosystem. For instance, D4Science-II will bring together several scientific e-Infrastructures in areas such as biodiversity, fishery resources management and high energy physics.

OpenAIRE aims to support the implementation of the open access policies of the European Commission's research programmes. It establishes infrastructure and help desks for researchers in all member-states, and is building a portal and exploring data management services in five scientific disciplines.

IV. Conclusions

Great science is performed at our leading research universities and institutes. We have the full-time equivalent of 1.5 million researchers in the EU - ahead of the US, though behind a rapidly growing China. The EU produces nearly twice as many doctorates as the US, and its share of the world's top-cited published research has been rising over the past decade. But the overall picture and current trends are worrisome: in patents, technology trade, scientific excellence and many other measures of innovation our position is weakening. As the European Commission's competitiveness report put it in 2011: "There is a widening gap between the EU and its world competitors" in major innovation measures[22].

In this context, GRDI is a challenge. If handled well, it can strengthen Europe's scientific base - placing us among the 'winners' who use GRDI to build on their existing strengths. If handled badly, we will be among the 'orphans' left isolated in a world of scientific collaboration and advance. With this in mind, several reflections are offered to guide Europe's role in GRDI development.

1. Enable science ecosystems

Several discipline-specific components of GRDI, such as Digital Data Libraries, Digital Data Archives and Digital Research Libraries, are already under development or will be developed in the near future. It is vital that these systems be able to interwork, and form ecosystems for both individual disciplines and for multidisciplinary work. The key concept they enable: digital science ecosystems - the communities of research, with associated tools, networks and practises - that will populate the GRDI.

2.Take into account the way science is organised and run, and the new tensions that GRDI will create in the research world.

Science is a social activity, with its own norms and organisational structures. A GRDI is not just a collection of fancy hardware and software; it also affects the way people work, collaborate and manage - across disciplines and cultures. It will create winners and losers - organisations and individuals that adapt quickly to the new world and prosper, and those who do not. New tensions will arise. The development of GRDI should take these science organisational aspects into consideration.

3. Base GRDI on scientifically sound foundations

Existing data infrastructures in research lack scientifically sound foundations, most often using application-

specific software of limited long-term value without a consistent computer science perspective. The research world tends to purpose-build software for each discipline, rather than devise common requirements that permit re-use of software across many disciplines. It is time to develop the theoretical foundation of scientific data infrastructures. This will permit us to develop generic data infrastructure technology and incorporate it into industrial-strength systems.

4. Define and implement better data models and query languages

Radically new approaches to scientific data modelling are required. In fact, the most common, relational data models were intended for business, not science. Scientific data have completely different characteristics. Data models and query languages are needed that more closely match the way scientists want to represent their information, that describe discipline-specific metadata, and that represent and query data-provenance information, contextual information, data uncertainty and data quality. Formally defined data models and data languages will allow the development of automated data tools and services, as well as generic software.

5. Develop advanced data tools

Current data management tools are inadequate for science. Researchers need better tools to capture, curate, analyse and visualise their data - to support the whole research cycle. These will permit researchers to follow new paths, try new techniques, build new models and test them in new ways. The result: innovative multidisciplinary and interdisciplinary work.

6. Develop advanced infrastructure services

The aim of GRDI is simple: to make it easier for scientists to collaborate globally. But of course, that is a difficult task. GRDI must support a new and growing range of data-intensive services, managed and shared within digital science ecosystems whose members have common needs. This goes beyond individual data tools, to include service offerings for finding tools, discovering needed data, integrating data sets, moving them across the network, managing the scientific workflow and the like. GRDI must provide the advanced support services that makes these components interoperable, usable and visible.

7. Support open linked data spaces

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

8. Support interoperability between science data and literature

In the future virtually all scientific literature and data will be on-line. The data must be unified with the literature to create a world in which they can interoperate. This will increase the "information velocity" of the sciences and will improve the scientific productivity of researchers. Future scientific data infrastructures must make this happen by supporting interoperability between digital data libraries and digital research libraries.

9. Recognise that open science requires open data

There is an emerging consensus in the research world that e-science must be open - both to join in the work, and to read and cite the work of others freely. It must provide open access not only to data, but also to scientific analyses and methods. This principle goes beyond a technical requirement; it must be reflected in law and policy, as well. The resulting principles of "open science" and "open access" must be widely accepted and incorporated into an integrated Science Policy Framework.

10. Create a new international research community

Scientifically sound data infrastructures can only be built if supported by an active international research community, capable of tackling all the scientific and technological challenges that such an enterprise implies. This community would embrace researchers who use data-intensive methods and tools (for instance, biologists and astronomers) and researchers who create or enable these models and methods (computer scientists, mathematicians, engineers, and the like.) So far, these two types of researchers have operated most often in isolation from one other. We believe that the development of a new data-intensive, multidisciplinary science must spring from synergetic action between these two groups. In fact, without it, there can be no data-intensive, multidisciplinary science. It will be an empty illusion.

11. Create new professions

To make best use of the huge volumes of data and network technologies we expect, new professional profiles must be created. These include data scientist, data-intensive distributed computation engineer, data curator, data archivist, and data librarian. They must be able to operate in the fast-moving world of network and data technologies. Education and training to enable them to use and manage the data and the infrastructures must be defined and put in action.

REFERENCES

1 Desmond, Adrian and James Moore. Darwin. Penguin Books, London, 1992.

2 Pennisi, Elizabeth. "Filling Gaps in Global Biodiversity Estimates." Science 330: 24 (1 October 2010.)

3 www.gbif.org.

4 Hey, Tony; Stewart Tansley and Kristin Tolle, Eds. "The Fourth Paradigm: Data-Intensive Scientific Discovery." Microsoft Research. Redmond, Wash: 2009.

5 Anderson, C. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." Wired Magazine, 06/27/08.

6 European Commission, High Level Expert Group on Scientific Data. "Riding the Wave: How Europe can gain from the rising tide of scientific information." October 2010.

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

8 Gray, J., D. Liu, A.Szalay, D. DeWitt, and G. Heber. "Scientific data management in the coming decade." SIGMOD Record Vol. 34, No. 4, Dec. 2005.

9 National Science Board. "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century." 2005.

10 Graham, P."The Digital Research Library: Tasks and Commitments." www.csdl.tamu.edu/DL95/papers/graham/graham.htl

11 National Science Foundation Cyberinfrastructure Council. "NSF's Cyberinfrastructure Vision for 21st Century Discovery." January 20, 2006, Version 5.0.

12 "Special Issue: Science Gateways – Common Community Interfaces to Grid Resources," in Concurrency and Computation: Practice and Experience. Wiley Interscience: 2007.

13 Hey, op.cit.

14 Adomavicius, G., J. Bockstedt, A. Gupta and R. Kauffman. "Understanding Patterns of Technology Evolution: An Ecosystem Perspective." Proc. Of the 39th Hawaii International Conference on System Sciences. 2006.

15 Edwards, P., S. Jackson, G. Bowker and C. Knobel. "Understanding Infrastructure: Dynamics, Tensions, and Design." Report of a Workshop on "History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures." January 2007.

16 Organisation for Economic Cooperation and Development. "OECD Recommendations of the Council Concerning Access to Research Data from Public Funding." Dec. 14, 2006.

17 Van den Eynden, V., 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.

18 G. Wiederhold. "Mediators in the Architecture of Future Information Systems." IEEE Computer, Vol. 25, 1992.

19 M. Stollberg, E. Cimpian, A. Mocan and D. Fensel. "A Semantic Web Mediation Architecture." Proc. Of the 1st Canadian Semantic Web Working Symposium. Springer, 2006.

20 K. Thearling. "An Introduction to Data Mining." http://www.thearling.com/dmintro/dmintro_2.htm.

21 Wiederhold, op. cit.

22 European Commission, Directorate-General for Research and Innovation. 'Innovation Union Competitiveness Report 2011." January 2011.

Disclaimer

GRDI2020 is funded 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/).