TECHNOLOGICAL AND ORGANISATIONAL ASPECTS OF GLOBAL RESEARCH DATA INFRASTRUCTURES TOWARDS YEAR 2020

Fotis Karagiannis[1*], Dimitra Keramida[1], Yannis Ioannidis[1], Erwin Laure[2], Dejan Vitlacil[2], and Faith Short[2]

[1]ATHENA Research and Innovation Center in Information, Communication and Knowledge Technologies, Artemidos 6 & Epidavrou, 151 25 Maroussi, Greece
**Email:* fotis.karayannis@gmail.com*; yannis@di.uoa.gr; d.keramida@di.uoa.gr*
[2]PDC Center for High Performance Computing, CSC School of Computer Science and Communication KTH Royal Institute of Technology, Teknikringen 14, Stockholm, Sweden
Email: erwinl@pdc.kth.se; vitlacil@kth.se; faith@kth.se

ABSTRACT

A general-purpose Global Research Data Infrastructure (GRDI) for all sciences and research purposes is not conceivable for the next decade as there are too many discipline-specific modalities that currently prevail for such generalisation efforts to be effective. On the other hand, a more pragmatic approach is to start from what currently exists, identify best practices and key issues, and promote effective inter-domain collaboration among different components forming an ecosystem. This will promote interoperability, data exchange, data preservation, and distributed access (among others). This ecosystem of interoperable research data infrastructures will be composed of regional, disciplinary, and multidisciplinary components, such as libraries, archives, and data centres, offering data services for both primary datasets and publications. The ecosystem will support data-intensive science and research and stimulate the interaction among all its elements, thus promoting multidisciplinary and interdisciplinary science. This special issue includes a set of independent papers from renowned experts on organisational and technological issues related to GRDIs. These documents feed into and compliment the GRDI2020 roadmap, which supports a Global Research Data Infrastructure ecosystem.

Keywords: Data Infrastructures, Data Management, Global Data Research Infrastructure (GRDI), Technological, Organisational, Policy, Analysis, Curation, Quality, Preservation, Interoperability, Linking, Discovery, Interoperability, Preservation, Provenance, Trust, Quality, Curation, Sustainability, Governance.

1 INTRODUCTION

According to Wikipedia, a data infrastructure is a digital infrastructure promoting data sharing and consumption, and similar to other infrastructures, it is a structure needed for the operation of a society as well as the services and facilities necessary for an economy to function, the data economy in this case. A research data infrastructure is a data infrastructure for research and scientists in particular. Because research and science are global, the need for GRDIs is becoming more evident. A general-purpose GRDI for all sciences and research purposes is not deemed possible in the next decades as there are too many discipline-specific modalities that currently prevail for such generalisation efforts to be effective. A more pragmatic approach is to start from what currently exists, identify best practices and promote effective inter-domain collaboration among different GRDIs or GRDI components forming an ecosystem. GRDI components are libraries, archives, and data centres, along with their services. This ecosystem of interoperable research data infrastructures will be composed of regional, disciplinary, and multidisciplinary GRDIs or GRDI components, offering data services for both primary datasets and publications.

This special issue aspires to improve the current understanding of the technological, organisational, and policy challenges in the development of interoperable Global Research Data Infrastructures (GRDIs). It is based on the work of the GRDI2020 EU-funded project (http://www.grdi2020.eu/) that finished in year 2012. Its main target has been to identify and outline key issues, challenges, and priorities towards interoperating GRDIs and the suggestion of possible actions to tackle those issues. Its main outputs are a set of reports including a roadmap document (Thanos, 2012), anchored on sound technical and organisational recommendations.

Given the global nature of GRDI2020, the reports and the roadmap are relevant to a wide audience and set of stakeholders that span beyond the European continent: from global policy makers, funding agencies, and

research service providers (e-Infrastructures and Research Infrastructures) to developers and research user communities at large. e-Infrastructure service providers refer to the networking, computing, data, and other electronic resource providers while Research Infrastructures service providers refer to the providers of research facilities including the ESFRI roadmap projects (http://ec.europa.eu/research/esfri/).

In summary, the list of topics that are addressed and the corresponding experts are as follows:

Table 1. Topics and experts

Topic	***Author***
Data Security	**Diego Lopez,** *Independent Consultant, Spain*
Data Discovery	**Gerhard Weikum**, *Research Director, Max-Planck-Institut für Informatik, Germany*
Funding, Sustainability, & Governance	**Matti Heikkurinen**, *Director, Emergence Tech Ltd., UK*
Data Policy	**Mark Parsons**, *National Snow and Ice Data Center, University of Colorado, USA*
Data Storage	**Erwin Laure / Dejan Vitlacil**, *Director of PDC HPC center / System administrator, KTH Royal Institute of Technology, Sweden*
Data Provenance & Trust	**Stratis Viglas**, *University of Edinburgh, UK*
Data Quality & Curation	**Kevin Ashley,** *Director, Digital Curation Center, UK*
Data Preservation	**Carlo Meghini**, *Prime Researcher, CNR, Italy*
Data Use – VREs	**Leonardo Candela**, *Researcher, Networked Multimedia Information Systems, CNR-ISTI, Italy*
Education & Training	**David Fergusson**, *Deputy Director Training, Outreach and Education, National e-Science Centre, UK*
Data Linking	**Chris Bizer**, *Prof. Dr, Research Group Data and Web Science, School of Business Informatics and Mathematics, University of Mannheim, Germany*

Each of these topics is analysed in terms of the state-of-the-art, the vision for the future, the challenges faced, and the proposed way to overcome these challenges. A list of recommendations for related stakeholders is also provided at the end.

2 SUMMARY OF EXPERTS' PAPERS

In their paper addressing **Data Storage and Management,** Laure and Vitlacil stress the need for a common globally interoperable distributed data system, formed out of data centres, that incorporates emerging technologies and new scientific data activities. The main challenge is to define common certification and auditing frameworks that will allow storage providers and data communities to build a viable partnership based on trust. To achieve this, it is necessary to find a long-term commitment model that will give financial, legal, and organisational guarantees of digital information preservation.

Pagano discusses **Data Interoperability**, which, while a paramount issue regarding global research data infrastructures, is still a challenge for open research. His ten-year vision is one of a unified information space, virtual or physical, based on the data infrastructure, that will give seamless access to heterogeneous data that were originally scattered across a number of independent data sources. As a highly challenging and multifaceted task, data interoperability subsumes many challenges and research topics. These include the lack of a common problem definition, coping with variety, enabling data reuse, agreeing on common standards, and developing comprehensive approaches. These challenges make it fundamental to develop a shared and participative strategy about how to approach the topic.

In his paper on **Data Discovery**, Weikum expects a quantum leap by the year 2020. There should be richer support for capturing the information needs of advanced users in a semantic representation, which will refer to entities and their relationships rather than keywords and pages. The most important limitation for this is the lack of semantic understanding of content as well as users' questions (or more generally, users' information needs). To address the challenges facing the 2020 vision, several research directions are proposed, including search for knowledge, search as a service, and personalisation.

Gionis develops a vision for **Data Analysis** that involves a global research data infrastructure in which many different datasets of extremely large scale are collected and stored and sophisticated data-analysis techniques can be applied to these datasets. A major challenge for this is to support collecting, storing, and analysing very large and heterogeneous datasets. It is important to start implementing such a system by addressing the easy questions first and then moving on to the more challenging issues and also by taking advantage of solutions that are already proposed and implemented. During this process, standardisation and open development environments are very important.

Data Provenance and Trust are discussed in the paper by Viglas. His vision for the future is the existence of platforms for the standardisation of all aspects of the digital life cycle, similar to the already existing Digital Object Identifier. In other words, there will be a way to uniquely identify digital artefacts as well as digital signatures of physical artefacts (e.g., individuals) and digitally captured workflows associated with the transformations of digital artefacts. There are plenty of challenges and limitations on the way to these goals. Provenance and trust will need to be retrofitted to existing infrastructures; sustainable and complete ways of recording and tracking provenance will need to be developed; and assigning trust will also need to be developed. The paper also points out the need for a standardisation body on provenance and trust as well as a canonical provenance- and trust-aware system.

Meghini discusses **Digital Preservation**, a relatively young discipline, whose importance becomes more and more apparent as the amount of knowledge encoded exclusively in digital form grows. In the future, the increase in demand will stimulate the creation of industrial-scale preservation services, leading in turn to the development of a new profession in digital preservation with a well-defined role and its own qualifications and training. Some of the paper's recommendations include the seamless integration of preservation into the life cycle of digital objects, the development of automatic techniques for obtaining preservation metadata automatically or semi-automatically from the objects themselves (analysis of multimedia content) or from external sources, and the promotion of sustainability of preservation by supporting the sharing of services and knowledge required.

Data Quality, discussed in the paper by Ashley, is an area that touches on every aspect of the research data landscape and is therefore appropriate to be examined in the context of planning for future research data infrastructures. Ashley presents a set of characteristics for the data quality systems of the future as well as a number of challenges towards this goal, illustrated through present-day use cases and the proposed actions for

overcoming them. Foremost among these are the education of data managers, curators, specialists, or carers (the terms vary, but the jobs are much the same) in generic techniques of data quality and other aspects of data curation.

Data Security is discussed in the paper by Lopez who feels that in the near future, Data Security will no longer be a matter of keeping "the inside" out of reach of "the outside". Security must become pervasive and must be dynamically associated with data themselves and their metadata so that the entities in the different ecosystems can apply the policies they consider relevant. The main challenges to realizing the above are essentially related to the evolution of the AAA (Authentication-Authorisation-Accounting) infrastructures themselves. It is also important to take into account the consolidation of security metadata models and to explore new patterns for integrating the two former models (AAA evolution and security metadata) with the rest of the data infrastructure. Recommendations include promoting the integration of different identity technologies in the research and academic community and collaborating in the development of standard methods for the new frontiers in data service integration with security infrastructures.

A high-level goal for the **Funding, Sustainability, and Governance** models of 2020 is developed by Heikkurinen who advocates the emergence of a common conceptual model capturing most of the value network supported by the GRDIs. Most of the challenges facing this goal rely fundamentally on building a communication network that reaches the majority of GRDI stakeholders and has high enough visibility and credibility to influence the regulatory processes surrounding research data management. Therefore, GRDI users and service providers should develop a common strategy for approaching the political decision makers.

Another important issue regarding data infrastructures is **Open Access**, discussed by Parsons. Ten years from now, all research data should be readily discoverable, and the vast majority of data should be open and in the public domain. There are several challenges currently facing open access. Implementation of the principle of full and open access is highly variable at the national level. In some nations, the national policy is entirely inconsistent with the principle. Furthermore, there is huge variability in attitudes towards data sharing across research disciplines. Work is needed to harmonise policy across national jurisdictions in accordance with common principles of openness and ethical use. Moreover, research communities need to define and develop norms of ethical, collaborative data sharing.

Candela presents **Virtual Research Environments** (VREs), which represent innovative working environments that aim to enhance the cooperation and collaboration among researchers in all modern research scenarios. His vision for the future is that regardless of geographical location, scientists will be able to use their Web browsers to seamlessly access data, software, and processing resources that are managed by diverse systems in separate administration domains. The challenges to fully achieving this are related to large-scale integration and interoperability, sustainability, and adoption. VREs should be designed to promote uptake, ensure usability, and guarantee sustainability. Resources and systems such as Internet, grid, and data infrastructures (e.g., GEANT, DataONE, OpenAIRE) should be considered building blocks to develop VREs.

Data Linking technologies are discussed in the paper by Bizer. The term Linked Data refers to a set of best practices for publishing structured data on the Web. There are indicators that linked-data architecture is suitable for extending the Web within the global scientific data space. These indicators include the increasing global adoption of linked-data technologies for sharing scientific, library, and e-government data as well as the first generation of linked-data discovery tools, such as linked-data search engines. By 2020, scientists will navigate along RDF links between different scientific datasets as well as between publications and supporting data. The current challenges that hinder the adoption of linked data technologies are related to data quality, data interoperability, and the lack of integration of linked data features into the scientific work environments that are used within the different scientific disciplines.

And finally, Fergusson identifies **Education and Training** as major underpinning activities. While much effort has been put into devising training regimes for distributed and high-performance computing, the focus tends to be more on computational aspects and less on data-related issues. One reason for this is certainly the complexity involved with data aspects that relate to all of the topics mentioned above. Another reason is the inherent heterogeneity and domain-specific issues involved in data handling that often lead to significant domain customization of any education and training program. Fergusson proposes including data-related and e-Infrastructure-related aspects in the curricula of the majority of disciplines, and he also proposes educating trainers accordingly. Because training material is itself data, the issues discussed above also relate to training

material. Efforts should be invested in curating and sharing data material to increase its quality for the whole community.

The papers collected in this publication are intended to stimulate discussions, outline strategic directions, and set foundations for the implementation of Global Research Data Infrastructures, keeping in mind the ten-year visions outlined here and in the GRDI2020 roadmap.

3 ACKNOWLEDGEMENTS

We are grateful for the financial support from the European Commission (Grant agreement no.: 246682) and would like to acknowledge all the GRDI2020 consortium members along with the reviewers.

4 REFERENCES

Thanos, C. (2012) GRDI2020 Roadmap Report "Global Research Data Infrastructures: The GRDI2020 Vision" - Final Release March 2012. Retrieved May 5, 2013 from the World Wide Web: http://bit.ly/18kLWFw.