

Rising to the Big Data Challenge

GRDI2020 Workshop on Global Research Data Infrastructures 2020

Research is increasingly global and multi-disciplinary. Advances in computing technology are empowering scientists to collect massive amounts of data, marking an important step towards solving complex problems like global climate change and uncovering secrets hidden in genes. Data intensive computing capabilities are fundamental for advancing data-intensive sciences, as well as huge volumes of complex data related to energy, health and national security. The challenge of global food security through the twenty-first century is inextricably bound up with other global issues, most notably climate change, population growth and the need to sustainably manage the world's rapidly growing demand for energy and water, calling for cross-cutting approaches. Interdisciplinary studies have the potential to empower scientific teams to previously impossible experiments, pushing the boundaries of our knowledge in areas as diverse as brain simulations, astronomy and seismography to forecast the effects of earthquakes. "Big data" is also transforming the research landscape for the humanities and social sciences. What new, computationally-based research methods might we apply now that we have massive databases of materials, from digitised books, newspapers and music to transactional data like web searches and sensor data?

Data-intensive research poses both new opportunities and challenges, calling for a new way of dealing with data. For the European Union, preparing for this scientific digital challenge is fundamental for competitiveness. The Research Data Infrastructure envisioned by GRDI2020¹ is an entire digital environment enabling a variety of interactions among data management systems, digital data libraries, research libraries, data collections, data tools and communities of research alongside the organisational practices of the people and institutions using it. Emphasis is placed on handling knowledge freely and dynamically to enable research communities to strike a balance between competition and co-operation.

The GRDI2020 Workshop on 'Global Research Data Infrastructures: The Big Data Challenge', 18-19 October 2011 in Brussels², brings to the table forward-thinking policy makers and world-renown experts from diverse fields to chart a course for GRDI. Questions addressed during the Workshop span the role of international cooperation in shaping visions for GRDI; the policy measures underpinning this sea-change; organisational and technological challenges that need overcoming and innovation that is being spurred by new approaches to scientific data management and distributed computing.

Data & Knowledge for Global Challenges

Global food security and environmental monitoring are among the multi-disciplinary and cross-cutting approaches to grand global challenges explored during the GRDI2020 Workshop. The Food and Agricultural Organisation (FAO), the United Nation's specialised agency, is chartered with ensuring that the world's knowledge of food and agriculture is available to those who need it when they need it and in a form which they can access and use. The talk by Johannes Keizer, Information Systems Officer at FAO, presents the CIARD Movement which the agency has facilitated to foster cooperation between all stakeholders, develop

¹ GRDI2020 – A 10 year vision for Global Research Data Infrastructures – www.grdi2020.eu

² Agenda, Speaker Details & Registration at <http://tinyurl.com/3lwp4jd>

capacity and create a technical framework for data and information sharing. As Johannes Keizer explains, “the technical framework leverages common standards and approaches, as well as the common development and use of tools and services. To achieve the goals set, special emphasis will also be placed on semantic web and linked open data”.

Seismology uses a range of applications to tackle critical social issues like earthquake detection, hazard assessment, volcanic eruption, and tsunami-warning systems. Global and regional seismological networks with a commitment to long-term operation, and pools of portable instruments for shorter term land- and sea-based deployments, provide key observations essential to tackling Grand Challenges. In the arena of seismic data considerable value is placed in open data access and real-time data collection, which have been embraced by much of the international community. New opportunities are now presenting themselves for the geosciences thanks to the US National Science Foundation’s DataNet Program. The programme fosters new community-driven stepwise approaches interweaving R&D costs, close out, education and the standards process to transform the way we do research. Exploring the innovation potential that can be unleashed by addressing the grand challenges in environmental monitoring, the GRDI2020 Workshop will connect a number of European and international initiatives that will help drive forward GRDI.

Technological Perspectives

Science has long entered the fourth paradigm, where scientific discovery is data-intensive requiring a large distributed and powerful computing infrastructure to collect, analyse and distribute ever increasing amounts of scientific data. Cloud computing is making the revolutionary collaborative models of today’s European e-Infrastructures more broadly accessible and applicable. The GRDI2020 Workshop also shines the spotlight on the potential impact of Cloud computing as we ride the data wave, by empowering researchers with user-friendly tools to focus on core work for the benefit of cross-discipline, cross-border co-operation. According to Fabrizio Gagliardi, EMEA Director at Microsoft Research Connections, “Cloud has the potential to seed scientific discovery across a spectrum of disciplines, solve complex problems and tackle data assimilation and real-time response by a much wider user community than was previously possible. But it does not stop here. Cloud also has the potential to create innovation clusters for e-Infrastructures in sectors as diverse as architecture, civil protection and drug discovery, paving the ground for new ‘Cloud-active’ European start-ups”.

Research infrastructures are also being shaped by supercomputing and advanced research facilities. DEISA (Distributed European Infrastructure Supercomputing Applications) has illustrated the central role of supercomputing in enabling computational scientists and European advances in chiefly driven by fundamental research that are now being taken forward by PRACE (Partnership for Advanced Computing in Europe) to compete globally. In the view of Kimmo Koski, Managing Director CSC & EUDAT Coordinator, “Securing Europe’s competitiveness over several supercomputer generations requires smart financial support coupled with the drive towards sustainable structures. PRACE marks a significant step forward in improving the competitiveness of European research by benefiting large and small countries. Alongside these objectives are the challenges surrounding data and the increasingly important field of data management. The goal of the EUDAT project is to create long-term collaboration and sustainable data services as an important legacy as we move towards Horizon 2020”.

Organisational and Policy Perspectives

Key strategic issues on data and information for science have been high on the agenda of the International Council for Science (ICSU) for the last decade. The four pillars cover long-term ICSU strategic leadership, professional data management in science, universal and equitable access and the important question of who pays for the data and information. Ray Harris, University College London, takes us on ICSU's journey to improve data and information management for science as we rise to the big data challenges. "The Strategic Coordinating Committee for Information and Data (SCCID) recommendations highlight the need to adopt a best practice guide to data management, clearly define 'open access' to data in science, explore the merits of regarding data as a publication and provide support on data management for science", explains Ray Harris. The CODATA and World Data System biennial conferences offer important forums to focus debate on data management and infrastructure issues.

On the policy front, the e-Infrastructure Reflection Group (e-IRG) and European Strategy Forum for Research Infrastructures (ESFRI) play a central role in shaping the current and future European landscape in a global context. As Gudmund Høst, e-IRG Chair highlights, "leading edge users are setting the pace for innovation in e-Infrastructures and the development of new services. e-Infrastructure providers should encourage these leading edge users to help define strategic directions". Reviewing regulations to allow equal treatment of public and private users to facilitate innovation and fostering funding through the budgets of users to enable alternative commercial provisioning of services will be important milestones moving forward. Speaking on behalf of the e-IRG, Høst remarks that "s model in which the use of e-Infrastructures is increasingly paid through user budgets will improve user participation and influence, bringing more options for choice and leading to greater user involvement in governance".

About GRDI2020

GRDI2020 is a Coordinated Action funded by the GÉANT and e-Infrastructures Unit of DG INSFO, European Commission. GRDI2020 is producing a roadmap to facilitate the development of next-generation global data research infrastructures that are needed to support a broad set of communities for enhanced knowledge generation and to address grand global challenges more effectively. It does this by bringing together experts with a proven track record in developing complex information infrastructures. This cross-fertilisation ensures research data infrastructures bring the best of both worlds to enhance data intensive and interdisciplinary research and international co-operation.