Towards a 10-year vision for a Global Research Data Infrastructure

www.grdi2020.eu

Fotis Karayannis

Independent,

on behalf of GRDI2020 consortium

eIRG Workshop, 4-5 April 2011, Budapest

From Global Research Libraries to ...GRDI2020

Workshop Title	Description
GRL2020 US	The 1st GRL2020 Workshop was held 30 September - 2 October 2007 and was hosted by the University of Washington. The Workshop brought together experts from Australia, Canada, China, Europe, India, Japan and the USA to define the **global research library of the future and discuss its main roles and functions.** Additionally, experts **evaluated the main implications for users and providers, and the underlying global regulatory and policy framework.**
GRL2020 Europe	Co-hosted by Microsoft Research and CNR-ISTI, the 2nd GRL2020 Workshop, 27-28 March 2008, Pisa, Italy, built on recommendations and conclusions from the 1st GRL2020 workshop, seeking also to identify the grand challenges and objectives for Global Research Libraries of the future.
GRL2020 Asia	Co-hosted by Academia SINICA and Microsoft, the third GRL2020 workshop, 24-25 February 2009, Taipei, Taiwan showcased best practices, case studies and pioneering work, with the aim of fostering innovative approaches supported by global research libraries. Focussed discussions helped evolve a GRL2020 research agenda (GRL2020 Call for Action) and chart a course that clearly defines next steps.

The complete picture

Technological Working Group

Focus on data challenges and issues like scalability, interoperability, curation, data access and financial feasibility and proposing steps towards their solution.

Roadmap on Global Research Data Infrastructures

GRDI2020 contributes to defining a future Roadmap focused on Global Research Data Infrastructures in close collaboration with the European Commission, policy groups and research user communities.

Organisational Working Group

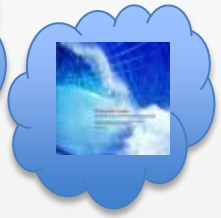
Strong support on activities related to organisational and policy topics identifying a core set of crucial open problems for the development of Data Infrastructures and suggesting possible solutions.

www.grdi2020.eu

Advisory Board

Ensuring an international perspective by facilitating the establishment of a worldwide network of relationships with relevant communities, funding agencies, and European and international organisations and supporting activities around the development of recommendations for the data research infrastructures of the future.

GRDI2020 Workshops



SDIs

Modus Operandi

Position & Briefing papers

Working Groups

Conf calls, mailing lists Wiki

Terms of Reference, Workplan

Reports

Physical meetings

GRDI2020 Project

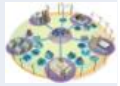
GRDI workshops

Position & Briefing papers

Consultation process

Roadmap

By the Community …
… for the Community

Target audiences	Description
Scientific Communities 	Academic, enterprise and research institutions from specific domains and including domains that are interdisciplinary by definition (e.g. Computational Biology).
Service Providers of e-Infrastructures and research Facilities 	Providers of distributed computing technologies and applications on the one hand and of large research facilities on the other. This group is one of the main targets for GRDI2020.
Technology Developers & Architects 	Technical developers closely involved in building infrastructure and digital systems, as well as people addressing issues such as data storage and management, data interoperability, data discovery and mining, data provenance and preservation and data protection.
Policy groups and funding agencies 	National and European policy groups that deal with scientific data research infrastructures, as well as their counter-organisations globally.

GRDI2020
A Vision for Global Research Data Infrastructures

First project year path...

February 2010

- GRDI2020 launch (Fact Sheet production).
- External Advisory Board set-up: 6 international experts in the field of data infrastructures to provide experts insights and advice on the project, with particular reference to the GRDI2020 Roadmap.

March 2010

- GRDI2020 website online.
- GRDI2020 Organisational & Technological Working Groups set-up: 20 international experts to identify understanding on critical organizational, policy and technical challenges related to the development of Global Data Research Infrastructures.
- eAnnouncement at OGF28 workshop: GRDI2020 presentation by Erwin Laure, KTH.
- Press Release on GRDI2020 launch.

May 2010

- GRDI2020 Working Groups Workshop: 7 Position Papers – "Analysis of Information Infrastructure Policy Reports" document discussion.
- Article on "Global Information Infrastructures for Science & Cultural Heritage: The Interoperability Challenge", Costantino Thanos, CNR-ISTI.

July 2010

- GRDI2020 Web-channel online (www.grdi2020.eu).

September 2010

- GRDI2020 at ICT2010:
  - Information booth "Scientific Data Infrastructures – Empowering research in the data-driven Age" in collaboration with 7 Scientific Data Infrastructure Projects (D4Science-II, DRIVER II, HELIO, IMPACT, METAFOR, OpenAIRE, SEALS) New synergies with SDIs, Video interviews, 50 new contacts.
  - Networking Session co-ordinated by DL-org and co-hosted by GRDI2020 and DC-Net "Global Information Infrastructures for Science & Cultural Heritage: The Interoperability Challenge" - Synergies.

October 2010

- GRDI2020 Workshop in collaboration with GRL2020 Africa, co-located with CODATA22, Cape Town, South Africa: 13 position papers, 37 participants from 13 countries, technical contribution to the Roadmap, Synergies with CODATA, ICSU, WDS.
- GRDI2020 Wiki as a collaborative tool to follow-up the discussion & Contribution to the WP2 report.
- 22nd International CODATA Conference (High Level & Parallel Session) New synergies with CODATA, ICSU and ICSU-World Data.
- GRDI2020 on the e-IRG website (Workshop Summary Report).
- Support to the High Level Group on Scientific Data event where the Report "Riding the wave: How Europe can gain from the rising tide of scientific data" was presented to Commissioner Kroes.

November 2010

- GRDI2020 First International Workshop – Post event Report.

December 2010

- GRDI2020 eNewsletter.
- "Analysis of Information Infrastructure Policy Reports" document available on GRDI2020 website.

January 2011

- **GRDI2020 Roadmap Preliminary Version**

GRDI2020
A Vision for Global Research Data Infrastructures

Second year path…

May 2011
- Workshop on "Global Scientific Data Infrastructures: The Big Data Challenges", 12-13 May'11 Capri, Italy, GRDI2020 Roadmap distribution.
- Support the European Commission and the Scientific Data Infrastructure projects "Co-ordination Workshop on Open Access to Scientific Information", 4 May 2011, Brussels.

June 2011
- Extremely Large Databases XLDB Workshop, 9-10 June 2011, Edinburgh: GRDI2020 Roadmap distribution.

September 2011
- The 1st ICSU World Data System Conference – Global Data for Global Science, 3-6 September 2011, Kyoto, Japan; GRDI2020 Workshop.
- EGI Technical Forum, 19-23 September, Lyon: GRDI2020 presentation.
- GRL2020 Workshop, Latin America, GRDI2020 Presentation & roadmap distribution

October 2011
- 2nd GRDI2020 Event, Brussels.
- e-IRG Workshop, 18-19 October, Poznan: GRDI2020 presentation.

December 2011
- GRDI2020 Pre-Final Roadmap Release.
- eScience 2011, 5 December 2011, Stockholm: GRDI2020 Workshop

January 2012
- **GRDI2020 Final Roadmap**
- **GRDI2020 Final event**

List of topics

Technical WG Topics	Organisational/Policy WG Topics
1. Data Security	**1. Data Security**
2. Data interoperability	2. Data interoperability
3. Data discovery	3. Data discovery
4. Data Provenance and trust	**4. Funding, sustainability and governance**
5. Data curation and preservation	**5. Data Policy –Open access**
6. Data Use - Virtual Research Environments	6. Data Provenance and trust
7. Data Storage & Management	7. Data curation and preservation
8. Training and education	**8. Data Use - Virtual Research Environments**
	9. Training and education

WGs Recommendations

Major Organisational and Policy Recommendations

Libraries and data centers

- Maximise synergies and minimise function overlaps

Translator Role

- For cross-disciplinary use of data and e-Infrastructure adaptation

Link data management and the **grant process**

- As a prerequisite for funding

Collate, categorise and learn from **best practices** for sustainable RDIs

- ICSU WDS (International Council of Science – World Data System) certifying data centers

Make the **costs** of data infrastructure transparent

- Identifying costs (including hidden ones!) will bring value and ensure conscious use of RDIs (otherwise laptops will still be used and data will be lost!)

Citations to data

- Incentives for researchers to publish their datasets / citation system to acknowledge them

GRDI2020 Roadmap

GOAL

- To **identify** and **address** the open research problems, both **technical** and **organizational**, which currently hinder the development of theoretically founded Global Research Data Infrastructures.
- To indicate **new research directions** necessary in order to address these problems.

Methodology

presenting a **complementary** vision, with respect to other roadmap reports:

- investigating and describing the technical/organizational challenges that must be tackled **from a computer science point of view**
- target audience are **policy-makers** & **scientists**

Roadmap Vision (I)

- In the next future several **Science Ecosystems**, composed of **Digital Data Libraries** and **Archives** (for primary datasets), and **Digital Research Libraries** (for research publications/literature) will be established
- A Global Research Data Infrastructure will act as the **enabler** of an **Interoperable Science Ecosystem**
  - By "**Interoperable Science Ecosystem**" we mean that all the components of the Ecosystem are able to exchange data and information without semantic distortions within a framework of shared policies

Preliminary Recommendations (I)

- GRDIs must be based on **scientifically sound** foundations
  - Current DB technology not adequate to support e-Science
- Formal **models and query languages** for data, metadata, provenance, context, uncertainty and quality must be defined and implemented
- New **advanced** data tools must be developed
  - Supporting the capture, curation, analysis and visualisation of data
- Future Research Data Infrastructures must support **open linked data spaces**
  - connecting datasets from diverse disciplines, regions and nations
- Future Research Data Infrastructures must support **interoperation** between science data and literature
  - On-line data sets and literature being interoperable

Preliminary Recommendations (II)

- Future GRDIs must support **Data-Intensive Research**
  - 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
- Future GRDIs must support **Multi-/ Inter-disciplinary** Research
  - Must support the "data/information/knowledge flow" across disciplines
    - i.e. crossing syntactic and semantic boundaries without distortions
  - An "**integrated mediation framework**" capable of providing means to handle all kinds of heterogeneities and inconsistencies
    - Such a mediation software will be a key component of the GRDIs
- A **new international research community** must be created
  - Composed of both end-users and computer scientists/engineers
- New **Professional Profiles** must be created
  - Promote the career path of mixed profiles (end-user+computer scientists)

Have Your Say …

A roadmap … By the community … For the community

- **Online public consultation:**
  - Access WGs report – GRDI2020 Channel (http://pakag.it/wgreport)
  - Access roadmap – GRDI2020 Channel (http://pakag.it/roadmap)
- **Final Roadmap – January 2012**

Contact: info@grdi2020.eu