Global Scientific Data Infrastructures: The Big Data Challenges

Capri, 12 – 13 May, 2011

Data-Intensive Science

Science is, currently, facing from a hundred – to a thousand-fold increase in volumes of data compared to the volumes generated only a decade ago.

The availability of huge datasets is a big opportunity, at the same time, a big challenge for scientists.

Scientific Data Infrastructure

A new type of e-infrastructure, the Scientific Data Infrastructure, must be developed, optimized for supporting the full life cycle of scientific data, its movement across scientific disciplines, and its integration with published literature.

Scientific Data Infrastructure (II)

Scientific Data Infrastructures can be defined as managed digital scientific data-networked environments consisting of services and tools that support:

- The full life cycle of scientific data (capture, collection, curation, documentation, analysis, visualization, preservation, and publication
- The movement of scientific data across scientific disciplines
- The creation of open linked data spaces by connecting data sets from diverse disciplines
- The interoperation between scientific data and literature

Scientific Data Infrastructure (III)

The next generation of scientific data infrastructures is facing two main challenges:

- To effectively and efficiently support data-intensive Science
- To effectively and efficiently support multidisciplinary/interdisciplinary Science

Data-Intensive Science

It is characterized by:

- Increasing volumes and sources of data
- Complexity of data and data queries
- Complexity of data processing
- High dynamicity of data
- High demand for data
- Complexity of the interaction between researchers and data, and
- Importance of data for a large range of end-user tasks.

Multidisciplinary/Interdisciplinary Science

Barriers to be overcome:

- A number of technological barriers must be overcome when moving data, information, and knowledge between disciplines
- The integration of activities that are taking place on different ontological foundations.

Data Challenges

To make this happen several breakthroughs must be achieved in the fields of scientific data modelling and management.

They include:

- Data modelling challenges (data description, context, provenance, quality, etc.)
- Data management challenges (data acquisition, organization, archiving, access, discovery, protection, privacy, authentication, curation, linking, integration, sharing, preservation, etc.)
- Data sevice/tool challenges (data analysis, data visualization, data mining, etc.)

Data Modeling Challenges

There is a need for data models and query languages that:

- More closely match the data representation needs of the several scientific disciplines
- Describe discipline-specific aspects (metadata models)
- Represent and query data provenance information
- Represent and query data contextual information
- Represent and manage data uncertainty
- Represent and query data quality information

Data Management Challenges

There is a need for advanced scientific data management capabilities that:

- Provide reliable, long-term, cost-effective preservation and access at appropriate quality
- Ensure high-confidence protection of privacy, confidentiality, security and property rights
- Ensure transparent search and discovery across a wide range of resources and data types
- Create open linked data spaces by connecting data sets from diverse domains

Data Service/Tool Challenges

Currently, the available data tools and services for most scientific disciplines are not adequate.

It is essential to build better tools and services in order to make scientists more productive.

Tools helping them to capture, curate, analyse and then visualize their data.

In essence, we need tools and services that support the whole research cycle and enable scientists to follow new paths, try new techniques, build new models and test them in new ways that facilitate innovative multidisciplinary/interdisciplinary activities are required

GRDI2020 Project

A Roadmap Report for

Global Research Data Infrastructures