G8+O5 GLOBAL RESEARCH INFRASTRUCTURE

SUB GROUP ON DATA

Draft Report, 28 October 2011

Introduction

The emergence of 'data driven science' reflects the increasing value of a range of observational, sensor, streaming and experimental data in every field of science. Information and communication technology infrastructures for scientific data are emerging world-wide, however, often these cannot be shared nor are interoperable across countries and disciplines; moreover, they are unsustainable due to lack of commonly agreed governance, legal frameworks and funding models.

The ability of scientific inquiry to address complex, multi-scale, multi-dimensional, multi-disciplinary questions means that there is a greater dependency on large, complex, widely distributed and often heterogeneous data sets at all stages of the scientific process including observation, theory, and validation. To ensure that the appropriate data are available for this type of research, it is essential to take a global approach to promote shared usage, interoperability and discoverability of scientific information resources. This approach should include accepted incentives to share and enable reuse of data and software and adequate funding mechanisms to ensure sustainability. It should help in rationalising and institutionalizing resources, building trust and widening access and use.

Essential attributes of the global infrastructure include viability, flexibility, participation, reliability, security and openness. Making it happen requires the involvement of researchers, universities, research laboratories, standardisation bodies, governments, funders, citizens and industry.

Future Scenario for Global Scientific Data Infrastructures

In 2020/2030…

> All stakeholders, from scientists, science managers, infrastructure operators and governmental authorities to the general public, are aware of the critical importance of preserving and sharing reliable data produced by a vast array of sensors and instruments during the scientific process and as a by-product of everyday life.
>
> Researchers and practitioners from any discipline are able to find, access and process the data they need in a timely manner. They are confident in their ability to use and understand data, and they can evaluate the degree to which that data can be trusted.
>
> Data are managed, shared, and preserved in a way that optimizes scientific discovery, innovation, and societal benefit. Where appropriate, producers of data benefit from opening it to broad access and routinely deposit their data in reliable repositories. A framework of repositories work to international standards, to ensure they are trustworthy.
>
> Funding bodies recognize that increased use of publically generated data can give added scientific and societal benefit. Reuse and repurposing of data across teams and disciplines is commonplace.
>
> The innovative power of industry and enterprise is leveraged by clear and efficient arrangements for exchange of data between private and public sectors allowing appropriate returns to both.

The public has access to, and can make creative use of, the huge amount of data available and they will be adequately educated and prepared to benefit from this abundance of information. All parties can contribute to the body of data stored in repositories and mechanisms are in place to attribute the source of these contributions and record their provenance.

Policy makers are able to make better informed decisions based on high-quality evidence, and can monitor the impacts of these decision.

Global governance promotes international trust and the interest and willingness to share data and support interoperability.

Challenges and gaps

Creating data

Diversity is likely to remain a dominant feature of scientific information: not only diversity of formats and types but also of the people and communities that generate and use the data. If global scientific data infrastructures are to promote multi-disciplinary science supported by reliable and high-performance infrastructures and overcome gaps and unbalances in data collection around the world, it will be necessary to work and to promote globally on incentives for data producers to benefit from interoperability between datasets and associated software at the time they are created (for acquisition, verification, annotation, rendering, etc). Without this, the whole scientific process will suffer from missed opportunities for global and interdisciplinary collaborations which are essential to the acquisition of new scientific and humanistic knowledge. Lack of coordination, interoperability and incentives to create data in shareable forms would result in unnecessary work for researchers and high costs for funders.

User communities in the different fields of science and humanities, research facilities producing data, scientific instruments and international teams of scientists should take an active role in the definition of concrete short and long term requirements for the underlying scientific data infrastructures for data. Networks of repositories, libraries and data centres should be interoperable at global level with high levels of dependability and trust, guided by international standards. They should profit and optimise the synergies with network and computing infrastructures and engage the user communities of researchers in defining the useful standards, services and platforms to be developed and maintained.

Preserving data

There is tremendous value in making data available, to use, reuse and recombine to support the creation of knowledge. In some cases, the data themselves have value and can represent such a large investment of resources that they may need to be preserved for subsequent use in the same way that unique observational data are preserved. On the other hand, in some fields, the costs and complexities of data preservation far outweigh the cost of reproduction and regeneration (eg. in high throughput analysis) and this should be considered when defining preservation policies.

As the volume and diversity of scientific data increase, and as research becomes more multi-disciplinary and as researchers struggle to understand and correlate data, especially if from another field, reliable infrastructures for persistent identification of data (e.g.

digital object identifiers, handle systems), researchers (e.g. digital authors identifiers), and authentication, authorization and accounting systems (AAA) are required. If scientific data are to be preserved and remain usable for the long term, these technical infrastructures also need to be sustainable in the long term.

Digital preservation solutions are undoubtedly partly technical, and the tools being created will enhance digital longevity, but these solutions are also equally dependent on organisational issues. The increasing amount of native digital scientific information - be it in scientific journals, data or software, has shifted the balance, roles and responsibilities regarding digital curation and preservation. In the analogue world, the task of long term preservation was the sole responsibility of libraries, but this is not longer the case. The main issues related to the long term digital curation and preservation remain basically unresolved, as many organisational, technical, financial, and legal aspects remain open.

Data and record management and information management policies are central to this and require co-ordination of policy at a high-level. Standards and guidelines for content quality assessment, verification of authenticity and provenance, and definition of quality of services, are both critical for meaningful access to preserved research data.

Lack of appropriate financing and organisational models put the long term preservation of digital scientific material at risk. This could leave a gap in our universal scientific memory and jeopardise the use of scientific and cultural material by future generations.

Research organisations worldwide should undertake proper preservation activities and establish sound and sustainable data management plans. It is essential that decisions on selection for this preservation and curation form part of an organisational process and are not made on an ad hoc and uncoordinated basis.

Global scientific data infrastructure providers should establish "forums" to define strategies at disciplinary and cross-disciplinary levels of metadata definition and federations, data description and provenance, data integrity and privacy, persistent identifiers and discovery mechanisms and models, as well as authentication and authorisation systems.

Accessing Data

All research builds on earlier work, therefore fuller and wider access to scientific data will help avoid reinvention and accelerate innovation.

Improving access to scientific information also has potential impact beyond research. It will have a positive effect on the quality of tertiary education and lead to more science-literate citizens who can have better informed opinions on the policy decisions that need to be faced in the 21st century.

The increasing availability of primary sources of data in digital form is already positioning data as a central element in the scientific process. Researchers can today access many online sources, but this is only a small fraction of all data produced. To improve science's efficiency and productivity, researchers will need to find, access, use and reuse data in a trusted way. This should be supported by offering training modules on information literacy in the course of scientific education.

On the other hand, as data becomes immediately globally accessible by others, it is important that quality assurance mechanisms exist to keep high the levels of trust. This is especially important given the diversity of types of use that increased access will promote.

Different users bring different approaches, tools, expertise and so forth. Data needs to be not only accessible, but also understandable as well as interpretable, to citizens, students, and decision makers as well as scientists, researchers etc.

The web allows linking papers and data fostering new hypermedia narratives to present research results. Therefore, the quality evaluation systems have to be thought to include reward mechanisms for data producers, curators etc e.g. by means of data citation or other incentive schemes. Agreements and standards for metadata for data citation enabling global non discipline specific metadata exchange will be required.

Data should also be readily discoverable by all these potential users without requiring specialist expertise and prior knowledge of the existence of the data sources. Key to this is the standardisation or establishing an architecture of metadata for the data itself and for the means of aggregation and collection and indexing. These standards are the basis to provide a common look and feel to data discovery across disciplines and reducing the learning curve required to achieve productivity.

Global governance frameworks should eliminate unnecessary barriers to accessing data and promote recognition and reputation mechanisms which encourage it. The producers of data, at both individual and institutional level, should benefit from opening data to broad access so that they prefer to deposit their data with confidence in reliable repositories than to keep it in their own closed systems.

Funding bodies, research institutions and disciplinary bodies, should work on incentives and policies for sharing data and associated software.

Underlying computing infrastructures

Scientific software, models and algorithms embed valuable information and knowledge. This includes the software and models that were used for generating, processing and correlating data so that the reproducibility and accuracy of the data can be verified. Also, the software used for analysis and visualisation are needed for long-term preservation and are necessary parts of the communication process.

Data generated through computer simulations are increasingly important in a variety of fields. Data generated entirely by computation can in principle be regenerated, assuming that enough is known about the hardware, software, and inputs used in the computation. However, each of these three components of a computation may be so complex or indeterminate that preservation can be more cost-effective than recomputation.

The infrastructure must be able to manage the expected scale of the future data resources. Some disciplines in particular will “push the envelope” with respect to what is technologically possible. The provenance of scientific data should also be managed through the data infrastructure that should support traceability and reproducibility by recording the derivation history of data.

Creating federated scientific data infrastructures building on already existing resources is economically efficient; for example, as the capability and reach of data centers and clouds continues to rapidly expand, cost per unit both in terms of storage as well as service will also continued to decrease. It also allows for a shorter "time-to-market". The main benefit for users is easy access to far wider resources than they could otherwise access through only a few content providers. It would allow also for an easy aggregation of new technologies that provide additional value (discoverability, data mining, cross discipline research).

Developers of advanced applications, tools and services should have better incentives to innovate. In the future, this may be even more important as increasingly complex data will not be understandable without access to and use of yet-to-be-developed analysis tools, visualizations, decision-making support, models, simulations, , etc…

Cultural Aspects

Science will experience even more major changes in the way it is performed. Researchers are already facing unprecedented levels of complexity to tackle scientific challenges with global societal impact.

Efficient support to global e-science and research communities requires the development of world-class e-Infrastructures capable of new "participative" collaboration paradigms. There are also social and behavioural aspects to be addressed like the clash of cultures between different disciplines, legacy frameworks and the need to rethink organisational models.

There is a need to train a new generation of data scientists requiring new knowledge about how researchers use and re-use information in different disciplines and countries. Data management and governance considerations should be included in the secondary and higher education curricula.

An inclusive infrastructure should enable adequate education and preparation of citizens to benefit from the abundance of technical and scientific information. They should support reliable methods to assess quality and impact of data collections.

Citizens, in particular students, should have adequate access to data and could contribute to it. They should be educated and prepared to benefit from the abundance of information.

The young would be inspired by an ambition for new discoveries and the creation of new businesses and industry, and join the ranks of scientists, engineers and entrepreneurs in far-greater numbers.

Governments, education and research funders and deliverers, and infrastructure service providers should work together to raise the general level of understanding of data and ensure that there is an adequate cohort of data specialists are available.

International Coordination and Governance

It will not be possible to achieve this integrated data infrastructure if each country or region acts alone. Interoperability, for example, requires that there be reciprocal agreements between governments.

Considerable effort is put into data acquisition, building data collections, curating data and producing value through annotations. With the internet and the emergence of more complex data environments including the emergence of data clouds, the task of “organising the data space” will become a crucial step of the whole scientific process.

Legal provisions on aspects such as privacy protection, intellectual property (copyright, licensing, etc.) have a strong impact on the development of global scientific data infrastructures. Policies for access, preservation, security, have to be compliant with legal frameworks that are quite heterogeneous. There are differences from country to country and from discipline to discipline.

Basic principles need to be agreed at global level as a pre-condition for further transparency and harmonisation. Global coordination will also be required to harmonise the changes to research culture which will lead to increased motivations for collaboration and data sharing.

There should be a group of international representatives who could meet regularly to discuss the global governance of scientific data infrastructures including policy, legal cultural and technical aspects.

Final remarks

Science is already experiencing major changes in the way it is performed. Researchers are facing unprecedented levels of complexity to tackle scientific challenges with global societal impact. Bringing or combining knowledge of different fields of science will be essential. Innovation in methods, processes and infrastructures is necessary to realize the vision of Global Scientific Data Infrastructures.

The engagement of all relevant stakeholders is essential and research communities from different disciplines have an active role to play setting the requirements for data infrastructures and services. Global infrastructure providers have to respond increasing the capacity and functionality so that researchers enjoy leading-edge capabilities and services for networking and computing, and seamless and wide access to data intensive science environments and global data resources.

The members of the working group agreed that this report is a good basis to set a vision and identify challenges and gaps and are looking forward to the feedback from the GSO.

The group felt the need to progress further with the work and discussions to elaborate an action plan with concrete steps to realize the vision.