White Paper: 5 Principles for an Open Data Infrastructure

Draft v2.0, 21 May 2013

(editors: Alan Blatecky, Juan Bicarregui, Carlos Morais Pires)

The effective management of digital data is indispensable in today's research and innovation activities and scientific communities have an important role to play in verifying and reviewing the data to ensure that science is self-correcting and to promote the scientific dialogue which improves the quality of research results and increases the levels of trust.

Research infrastructures and facilities have demanding requirements for Information and Communication Technologies supporting advanced data, computation and connectivity and the percentage investment in these technologies is growing. Research *data* infrastructures are made of environments supporting advanced data acquisition, connectivity, storage, curation, management, integration, mining and visualisation and computing and information processing services and communication networks beyond the scope of a single institution.

For these technologies and infrastructures, interoperability at the level of storage, connectivity and computation is well established, but achieving interoperability at the level of data interoperability is a far harder challenge. The paragraphs below describe five principles for establishing a global interoperable data infrastructure with cost-effective solutions for widening access and ensuring long-term preservation.

1. Discoverable

Scientific data sets, even though valuable and available, will see little or no use if they cannot be easily found using conventional search method. These data sets must be accessible to web crawlers and have adequate description information to allow the searcher to understand the content, specialized formats, or languages. Making data easier to find and more readily discoverable will enable efficient use of the data sets, reduce costs for research and development, and increase innovation. Examples of approaches that can improve discoverability include implementation of appropriate persistent identifier frameworks, adoption of descriptive metadata standards, and the use of appropriate data formats and taxonomies.

2. Accessible

Publicly funded scientific research data should be made openly available with as few restrictions as possible. Ethical, legal or commercial constraints may be imposed on the use of research data to ensure that the research process is not damaged by inappropriate release of data. Normal research practice regarding the citing of sources should be observed—namely, that secondary users of research data must acknowledge the source of the data and have to give credit for any intellectual contribution made in

collecting and organising the data. Users of data may be required to sign up to terms of use or licenses associated with particular data, for example, regarding consent or the protection of privacy of human subjects, preservation of intellectual property, or agreeing to an embargo approach. Once alternative systems are in place to adequately acknowledge the contribution of data providers, it should be possible to reduce or remove embargo periods and other restrictions to accelerate data driven knowledge creation and innovation.

3. Understandable

Scientific data sets must be understandable in order to be effectively used. A set of numbers, texts, pictures or even videos alone cannot be understandable without additional context, semantics, data analysis tools, and algorithms. Observational and sensor data must be accompanied with the documentation about the location, time, and method of observation. A set of rules must be established to ensure the integrity and provenance of the data including data quality, curation and preservation. Although English has become the de facto common language in many fields of research, translation among data collections is essential and translation between disciplines will become ever more important as inter-disciplinary research efforts and international collaboration continues to grow.

4. Manageable

In order for research data to be managed in an efficient and effective manner, data management policies and plans must exist for all data at both project and institutional levels. Making research data available in a form in which can be effectively used by others requires considerable and continued efforts over and above those which are necessary to undertake the primary research itself. Data management policies and plans must make it clear who is responsible for maintaining the availability of data and how the associated costs are to be met including issues associated with curation, storage and services. Plans and processes must be in place that will take into consideration the full range of potential uses for the data in a cross disciplinary context as well as the requirement for data with acknowledged long term value to be preserved and remain usable for future research. Coordination of provision of technology and data services so as to make effective use of common technology and avoid duplication of effort will be instrumental in ensuring an efficient data infrastructure.

5. People

A global approach to research data infrastructure requires a highly skilled and adaptable workforce and culture that is able to capture the available data and make it available to those that are able to use it appropriately. There is a need for specialised data custodians who can work across complex data sets and with diverse data protocols. Some of the specialist skills required are not yet fully reflected in existing qualifications and training programs and there is a need to develop suitable programs of training to meet this need. There is also a need to change the culture of research data management within the research community. In addition to the specialised skills

required, there will need to be broad-based training provided as part of research training programs to ensure that researchers understand the need for, and benefits of, data sharing and the principles by which this will happen.

Metrics

Today, many nations and institutions are investing in digital scientific data and data infrastructure recognizing the potential value in catalyzing discovery, innovation, education, and entrepreneurship. Responsible management requires the ability reliably to measure the impact of those investments. Many of the standards and technologies that promote access also enhance capabilities for measuring impact. As an example, persistent identifier frameworks not only enable discovery and use of data, but facilitate tracking the applications of a data set over time as it is used in many different settings for a wide variety of purposes. Facilitating the ability to measure return-on-investment, including implementing incentive schemes for researchers, is among the goals of this effort.

Recommendation that these 5+ principles be integrated into and form the basis of national policies