The role of the International Science Council and CODATA in Enabling Global Transdisciplinary Integration of Data in Support of New Research Horizons.

Lesley Wyborn, Simon Cox, Simon Hodson, Xiaogang Ma

**Summary**

Scientific data are being generated at an ever increasing rate, whilst citizen science is proliferating and providing much valuable data. Open source software makes tools readily available at no cost. Cloud computing is ubiquitous: no longer does a researcher/research institute need to invest in substantial storage and processing infrastructure. The internet provides unprecedented online access to data, tools and compute resources and there is now an unrivaled opportunity for global research challenges to be addressed by people from anywhere in the globe: participation is no longer the purview of the high income countries. The stumbling block for this vision is that data from multiple sources and diverse communities are hard to bring together because of differences in standards, vocabularies and data structures. In order to be able to share data, information and services, standards related to the retention, discovery and access to these data will need to be developed. Many, if not most, of the decisions about what to store, what standards to apply and what are the minimum required metadata lie within the relevant peak international science and social sciences unions: they have a new role in the establishment of standards that enable data to be born connected across natural and social science boundaries and beyond. But it is not just about promoting the development and endorsing the standards, the unions also have a role in long term governance to ensure that as new major research challenges appear and technologies change, the required standards evolve to support them.

# **Abstract**

The internet of today gives researchers unprecedented online access to data, tools and compute resources. Researchers from anywhere can participate including those from low to middle income countries in what is becoming an unrivaled opportunity for global research challenges to be addressed using distributed data and compute resources. Citizen scientists are becoming an increasingly important participant in these global research networks: theoretically data from anyone and anywhere can contribute. Further, much of modern scientific research can no longer be undertaken in isolation from the community and humanity - projects need to be studied within the context of the real world and any potential social impacts evaluated. The research community is rapidly moving towards an era of transdisciplinary data intensive research.

However, the reality is that varying and incompatible data standards used across the different disciplines, and inadequate definition of scientific vocabularies needed to categorise observed phenomena across multiple languages, are making scientific data integration impossible. At best, integration of diverse data can generally only be achieved within small communities.

In recent years International Standards bodies such as ISO (International Organization for Standardization), OGC (Open Geospatial Consortium) and W3C (World Wide Web Consortium) have been developing technical standards that are applicable for interchange of scientific data such as GML (Geography Markup Language), Observation and Measurement Standard, Spatial Coordinate Systems, Metadata Standards etc. However, what is now needed is agreement as to who will endorse the domain-specific standards for much of the data collected as part of scientific research. The OGC and RDA (Research Data Alliance) have been supporting some communities that wish to develop the more domain specific standards, but to be authoritative, the science and social science unions need to be involved. Some unions are already getting involved in this work, but their participation is patchy, and the results are inconsistent.

Formalisation of the discipline-specific vocabularies is an essential prerequisite for integration of data from different disciplines. There is thus an urgent need to develop a unified approach to standards development in order to reduce replication of effort and a proliferation of incompatible practices. The need to reduce the number of standards required and coordinating their development is essential if at a basic level we are to discover and access data from across the various science disciplines.

There is a clear lack of infrastructure and governance not only to develop the required standards but also to store, maintain and extend these standards over time. There is also no formal mechanism to harmonise decisions made been the various international and national scientific bodies to avoid unwanted overlap. For example, many of fundamental data interchange standards related chemistry and physics data could also apply to geochemical and geophysical data respectively.

Several members of the international science unions have already set up specific commissions on data and information. For example, the International Union of Geodesy and Geophysics has a Commission for Data and Information, the International Union for Geological Sciences has a Commission for the Management and Applications of Geoscience Information, and the International Astronomical Union has a Working Group on Astronomical Data. Nevertheless, the development of these commissions/specialist groups has been on an ad hoc basis and there is no coordination between these various commissions.

What is needed is for the newly formed International Science Council, created through the merger of the International Social Science Council (ISSC) and International Council for Science (ICSU), to take on a leadership role in coordinating standards development by these groups and minimising duplication of effort and to assist each of the International Unions to establish a specific Commission on data and information. Once the endorsed standards are identified, it is hoped that CODATA can use a website and social media to take a leadership role in raising awareness of standards endorsed by and/or being development by the individual unions to assist in promoting the authoritative standards and minimise duplication of effort and to also provide a web-accessible page that provides links to repositories for data models, information standards, vocabularies, ontologies, etc., for each of the unions.

To ensure ‘fitness for purpose’, data users and software developers also require guides as to the useability of the standards and an assessment of the overall maturity of their standards within the specific community: this will require the development of a broad ‘maturity model’ for scientific standards. Developers of the standards will also benefit from provision of best practice examples for the development and application of the required standards and guidance on developing governance frameworks for the maintenance and revision of these standards, preferably by assisting linkages to key groups such as the international community of RDA, as well as national efforts such as the Australian National Data Service, the US Earth Science Information Partners (ESIP) and European Union Horizon 2020 projects.

Above all, the International Science Council and CODATA could also help to provide support to the scientific community in this transition to transdisciplinary data intensive research by emphasising the need to participate in the development of the required standards and promote the benefits of adherence to standards to increase discoverability and accessibility to data.