

# Standards: a critical enabler for cross-disciplinary scientific research....but who?

Lesley Wyborn<sup>1,2</sup>, Ray Norris<sup>1,3</sup>, Rhys Francis<sup>1,4</sup>, Kim Finney<sup>1,5</sup>, Alex Held<sup>1,6</sup>, Jane Hunter<sup>1,7</sup>, Tim Littlejohn<sup>1,8</sup>, Karen Wilson<sup>1,9</sup>

<sup>1</sup>Australian Academy of Science, National Committee for Data in Science, Canberra, Australia

<sup>2</sup>Geoscience Australia, Canberra, Australia, [lesley.wyborn@ga.gov.au](mailto:lesley.wyborn@ga.gov.au)

<sup>3</sup>CSIRO Australian National Telescope Facility, Sydney, Australia, [ray.norris@csiro.au](mailto:ray.norris@csiro.au)

<sup>4</sup>Australian eResearch Infrastructure Council, Melbourne, Australia, [rhys@pfc.org.au](mailto:rhys@pfc.org.au)

<sup>5</sup>Department of Environment and Heritage, Hobart, Australia, [Kim.Finney@aad.gov.au](mailto:Kim.Finney@aad.gov.au)

<sup>6</sup>CSIRO Division of Marine and Atmospheric Research, Canberra, Australia, [alex.held@csiro.au](mailto:alex.held@csiro.au)

<sup>7</sup>University of Queensland, Brisbane, Australia, [jane@itee.uq.edu.au](mailto:jane@itee.uq.edu.au)

<sup>8</sup>IBM Australia, Sydney, Australia, [tgilittle@au1.ibm.com](mailto:tgilittle@au1.ibm.com)

<sup>9</sup>Royal Botanic Gardens, Sydney, Australia, [karen.wilson@rbgsyd.nsw.gov.au](mailto:karen.wilson@rbgsyd.nsw.gov.au)

Scientific data are being generated at an ever increasing rate. Existing volumes of data can no longer be effectively processed by humans, and efficient and timely processing by computers requires development of standardised machine readable formats and interfaces. There is also a growing need to share data, information and services across multiple disciplines. Increasingly, digital data collections are being re-used and re-purposed, often by scientists who do not necessarily have the same level of discipline expertise as the originator of the data.

Current key challenges in earth and space sciences, such as climate change, hazard prediction and sustainable development of our resources require a cross-disciplinary approach, and in some cases, data will need to be integrated from globally distributed sources. There are many other examples of the need for cross-discipline research. For example naturally known anomalous concentrations of deleterious elements recorded in some geoscience data sets have relevance to medical research, whilst social data collections can underpin research into social sciences and impacts of natural hazard events. Such cross-disciplinary research, particularly if it involves large amounts of data, can only occur if there is some coordination across the relevant disciplines in the development of standards related to the retention, discovery and access to these data.

Creating cross capability linkages was recognised as one of the 5 key lessons for future program implementation in the NCRIS Strategic Roadmap for Australian Research Infrastructure released in August 2008[1], in particular, developing the collaborative tools, networks and mechanisms to facilitate the sharing of data. Nearly every one of the NCRIS capabilities has an informatics component and most of these are adopting a standards based, open access approach to making their data and information accessible via community agreed standards.

However, at the discipline level, many, if not most, of the decisions about what to store, what standards to apply and what are the minimum required metadata are being made within the individual discipline. It is increasingly apparent that the existing standards are becoming 'stove-piped' on a discipline by discipline basis. When groups from different communities do try combine data across the discipline boundaries much time is spent reformatting and reorganizing the data and it is conservatively estimated that this can take 80% of a project's time and resources.

For data to be interoperable across multiple domains action needs to be taken urgently, particularly as the potential of the semantic web now is being realised. To enable efficient cross-disciplinary research a more modular approach to standards development is required so that common components (eg location, units of measure, geometric shape, instrument type etc) can be identified and standardised across all disciplines.

Already international standards bodies such as ISO and OGC (Open Geospatial Consortium) are well advanced in developing technical standards that are applicable for interchange of some of these common components such as GML (Geography Markup Language), Observation and Measurement Standard, SensorML, Spatial Coordinate Systems, Metadata Standards, etc.

However the path for developing the remaining discipline specific standards is less coordinated and there is some confusion over who should actually be developing these. Should they be developed at an institution, national or international level? In some disciplines there is also conflict between some groups developing standards that are free and openly available vs commercial standards that can only be accessed through subscription and/or at times prohibitively high fees.

There is thus a pressing 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, even 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 for the development of the required standards but also for storage, maintenance and extension of these standards over time. There is also no formal mechanism to harmonise decisions made by the various international and national scientific bodies to avoid unwanted overlap. For example, many

of fundamental data interchange standards related to chemistry and physics data could also apply to geochemical and geophysical data respectively.

In Australia, the National Committee for Data in Science (NCDS)[2] was established in 2008 by the Australian Academy of Science to provide an interdisciplinary focus for scientific data management. The NCDS aims to promote and facilitate data use in science across all disciplines of science and to provide a national voice that can represent Australia at international forums related to Data in Science. To achieve this, the committee will hold regular workshops promoting the development of data management policies and protocols, and promote the adoption of standards for data exchange.

The NCDS will also represent Australian interests on the international interdisciplinary Committee on Data For Science and Technology (CODATA)[3]. CODATA is sponsored by the International Council For Science (ICSU) and is concerned with improving the quality, reliability, management, and accessibility of data of importance to all fields of science and technology. They provide for example, the scientific and technological communities with a self-consistent set of internationally recommended values of the basic constants and conversion factors of physics and chemistry [4].

The NCDS has noted that 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 Application of Geoscience Information, and the International Astronomical Union has a Working Group on Astronomical Data. However, their development has been on an ad hoc basis and there is no coordination between the various commissions.

The NCDS has recognised that CODATA could possibly provide the much needed governance and infrastructure to coordinate scientific standards development. The NCDS put an informal request to CODATA at their 2008 General Assembly to suggest that CODATA could take on a new role which could include

1. Assisting each of the International Unions to establish a specific Commission on data and information
2. Taking a leadership role in coordinating standards development by these groups and minimising duplication of effort
3. Providing a web-accessible international standards repository for data models, standards, ontologies, and vocabularies
4. Providing best practice examples for the development of the required standards
5. Providing a governance framework for the revision and updating of these standards
6. Promoting the benefits of adherence to metadata standards to increase discovery and accessibility to data.
7. Providing guidelines to the scientific community on the need to adhere to these standards

## REFERENCES

1. Department of Innovation, Industry, Science and Research, 2008. Strategic Roadmap for Australian Research Infrastructure. Available from <http://www.innovation.gov.au/ScienceAndResearch/Documents/Strategic%20Roadmap%20Aug%202008.pdf>
2. National Committee for Data in Science <http://www.atnf.csiro.au/people/rnorris/NCDS/about/index.html>
3. CODATA accessed from <http://www.codata.org/>
4. CODATA recommended values for fundamental physical constants. Available from <http://physics.nist.gov/cuu/Constants/index.html>