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

Lesley Wyborn
National Committee for Data in Science and Geoscience Australia

Outline

1. Introduce the National Committee for Data in Science
2. Drivers for the increasing need for standardisation
3. The Vision: No Boundaries
5. Can CODATA be the WHO?
National Committee for Data in Science (NCDS)

Members

Rhys Francis (Chair)  Australian e-Research Infrastructure Council (Melbourne)
Kim Finney  Department of Environment and Heritage (Hobart)
Alex Held  CSIRO Div. of Marine and Atmospheric Research (Canberra)
Jane Hunter  University of Queensland (Brisbane)
Tim Littlejohn  IBM Australia (Sydney)
Ray Norris  CSIRO Australian National Telescope Facility (Sydney)
Karen Wilson  Royal Botanic Gardens (Sydney)
Lesley Wyborn  Geoscience Australia (Canberra)

National Committee for Data in Science

- Established in 2008 by the Australian Academy of Science to provide an interdisciplinary focus for scientific data management

- The NCDS aims to
  1. Promote and facilitate data use in science across all disciplines of science
  2. Provide a national voice that can represent Australia at international fora related to Data in Science.
  3. Hold regular workshops promoting the development of data management policies and protocols, and promote the adoption of standards for data exchange.
  4. Represent Australian interests on the international interdisciplinary Committee on Data For Science and Technology (CODATA)
Drivers (1): The Data Deluge (Tsunami?)

- Scientific data are being generated at an ever increasing rate
- Existing volumes of data can no longer be effectively processed by humans
- Efficient and timely processing by computers, particularly at petascale, requires development of standardised machine readable formats and interfaces

Source: http://www.tsunami.org/images/student/art/hokusai.jpg

Drivers (2): Our changing science - the move from reductionism

Source: Office of Integrative Activities NSF
Drivers (3): the change in how we do our science

- There is a growing need to share data, information and services across multiple disciplines
- Increasingly:
  - digital data collections are being re-used and re-purposed by much broader communities
  - data are being accessed by scientists who do not necessarily have the same level of discipline expertise as the originator
  - solving major challenges facing our planet will require that data will be integrated from globally distributed sources

Drivers (4): NCRIS 2008 Strategic Roadmap

- Creating cross capability linkages was recognised as one of the 5 key lessons in the 2008 NCRIS Strategic Roadmap for Australian Research Infrastructure
- In particular, the need for developing the collaborative tools, networks and mechanisms to facilitate the sharing of data
- Nearly every one of the NCRIS capabilities has an informatics component
Standards: enabling The Vision

- Internationally relevant cross-disciplinary research requires standards, particularly if it involves large amounts of data.
- It can only occur if there is some coordination across the relevant disciplines in the development of standards related to the retention, discovery and access.
- But what, who and where?

Source: [http://www.mgmyers.com/assets/world_in_hands.jpg](http://www.mgmyers.com/assets/world_in_hands.jpg)
Standards: what and who?

• But What
  • do we store?
  • standards should we follow?
  • is the minimum metadata to allow data to be recreated, reused and repurposed across disciplines?

• And Who?
  • We cannot solve this on a national or single discipline based approach
  • To undertake global challenges the standards must be international-based and where possible cross-disciplinary

Standards: where do we find them?

• There is no complete one-stop shop: standards have to be modularised into common components

• Already international standards bodies (ISO, OGC, W3C) have developed critical generic standards, e.g.,
  • GML (Geography Markup Language),
  • Spatial Coordinate Systems
  • Metadata Standards (e.g., ISO 19115, 19139, 2146)
  • Observation and Measurement Standard
  • SWE (Sensor Web Enablement)

• These enable us to leverage common components across many scientific disciplines which will facilitate cross discipline data integration

• Think Lego!
But what about discipline specific standards?

- The path for developing the remaining discipline specific and discipline independent standards is less coordinated and there is widespread confusion.

- They need to be developed at an international level, but who should be developing these??

- Should these standards be free or should we pay?

Source: [http://bintiafrica.files.wordpress.com/2009/05/confusion.jpg](http://bintiafrica.files.wordpress.com/2009/05/confusion.jpg)
What do we actually need?

We need internationally endorsed data transfer standards for

- Geology
- Physics
- Chemistry
- Astronomy
- Biology
- Materials science
- etc

What about the fringe dwellers?

- Geochemistry
- Geophysics
- Geobiology

We need to coordinate the development of scientific standards to avoid a plethora of incompatible data transfer standards and the uncontrolled growth of YAMLs (Yet Another Markup Language), vocabularies, ontologies, etc.

We need governance!

- Having developed the standards governance is required
  - for the storage, maintenance and extension of standards over time within a discipline
  - to provide a formal mechanism to harmonise decisions made by the various scientific bodies to avoid overlap
  - a web accessible persistent repository of these standards in machine readable formats (not .pdf’s)

- Can CODATA help?
Committee on Data For Science and Technology

- 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.

- Traditionally CODATA provides 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.

- In the digital age, could CODATA be the one to provide much needed governance on digital scientific data standards in the rapidly expanding data deluge?

Current Scientific members of CODATA

1. International Astronomical Union (IAU)
   - Working Group on Astronomical Data
2. International Union of Pure and Applied Chemistry (IUPAC)
4. International Union of Biological Sciences (IUBS)
5. International Geographical Union (IGU)
6. International Union of Crystallography (IUCr)
7. International Union of Biochemistry and Molecular Biology (IUBMB)
8. International Union of Geological Sciences (IUGS)
   - Commission for the Management and Application of Geoscience Information
9. International Union of Psychological Science (IUPsyS)
10. International Union of Pure and Applied Biophysics (IUPAB)
11. International Union of Nutritional Sciences (IUNS)
12. International Union of Basic and Clinical Pharmacology (IUPHAR)
13. International Union of Immunological Societies (IUIS)
15. International Union of Soil Science (IUSS)
16. International Union of Geodesy and Geophysics (IUGG)
   - Commission for Data and Information

Data Groups/Commissions
NCDS Proposal to CODATA

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

Watch this space for further developments

National Committee for Data in Science

Any Questions?