The Role of Scientific Data in e-Science: How Do We Preserve All Necessary Data So They are Useful

John Rumble
Technical Director
Information International Associates
Oak Ridge TN USA

CODATA 2006 Beijing - E-Science Session

Data and e-Science

Data express the ***quantitative results*** of scientific research

- Experiments on nature
  - Testing an isolated and controlled part of the natural world
- Observations of nature
  - Making measurements on the natural world as it is found but not controlled
- Calculations on models of nature
  - Creating a virtual world containing some, but not all, factors that control natural phenomenon
- ***Today fostered and facilitated by e-Science***

Data and e-Science

- Virtually all data are generated, collected and preserved digitally by scientists in an e-Science environment
- Preserved data collections allow others to reuse past measurements rather than generate new measurements to develop new ideas and knowledge
- ***Today’s large scale data collections are a new source of scientific discoveries***

- Both an extension of traditional scientific method and a new discovery mechanism

Challenges to the Preservation and Reuse of Data in e-Science

- The gap between an ideal experiment, observation and calculation and reality
- ***Large number of independent variables***

- The evolution of scientific knowledge and language
- ***Changing scientific language***

- The multi-center nature of scientific research
- ***Science is a team effort – across disciplines, places and time***

The gap between an ideal experiment, observation and calculation and reality

Real systems are very complex

- ***Large number of independent variables***

The Challenge

- ***Real experiments, observations and calculations do not control, capture or record all independent variables***

- ***The reporting of independent variables changes over time as scientific knowledge increases***

Time and Independent Variables

- Independent variables are the quantitative mechanism for expressing our knowledge about how and why a phenomenon occurs
- Capturing complete knowledge of independent variables requires a large or (perhaps) even an impossible amount of data
- One goal of research is to understand which variables are important and why
- And which variables to report!
- ***Our knowledge clearly evolves over time***

Time and Independent Variables

Major challenge of data standards is to capture evolution of knowledge of independent variables

- The set of variables we must report today is not adequate tomorrow
- Standards must allow for growth of knowledge
- Yet must also enable compatibility of data generated at different times

- Let’s work through a quick example of the complexity

Time and Independent Variables

Brain imaging

- Recording techniques evolve and improve over time
  - X-ray, CT, MRI, PET, next?
- Each technology individually evolves, as do the types of signals collected, their association with brain activity and region
- Monitoring reactions to stimulus: pain, visual, auditory, tactile, etc.
- Details of independent variables must be defined and recorded

Time and Independent Variables

Consider brain history

- If we imagine the details necessary to describe this, the number of independent variables expands rapidly
  - – Stimuli history, physiological history, developmental history, environmental exposures, education, more

- As with the development of unifying theories of the large-scale physical world – motion, evolution, chemistry, genetics - the details are necessary to find the dominant factors

What are the most important independent variables for recording brain history? Still an open question and will change over time!

Time and Scientific Language

How do languages evolve?

John McWhorter – The Power of Babel

- Contractions of words
- Reordering sentences
- Borrowing words
- Dropping and adding of word beginnings and endings
- Differentiation of concepts
- Evolution of concepts

These are powerful change factors that cannot be ignored in preserving data

Time and Scientific Language

- ***Data preservation efforts must recognize evolution of scientific language***
- Not just independent variables and metadata – the scientific language itself

- As concepts change, the definition of the word(s) defining the concept change
- As different disciplines begin to overlap, slightly different concepts merge into a unified concept, different from each original, often using the same word

Time and Scientific Language

Examples of evolving concepts and language

- Atomic nucleus
- Nonsense DNA
- Chemical bonding
- Dam

- The tools to locate and determine ***the equivalency of scientific and technical words and language over time*** are just beginning to be developed

Data and e-Science

Yesterday

- Collections managed by a small number of people
- Collections readable by one scientist
- Collections interpretable by one person
- Discoveries made by thinking, with analysis by one person

Today and the Future

➔ Collections managed by groups

➔ Collections not readable by any individual

➔ Collections interpretable only with aid of software

➔ Discoveries made by computers, with verification by people

Data in e-Science

- One goal of e-Science is ***to create large data sets*** consisting of all measurements relevant to a system, phenomena, filed of scientific discourse
- This can be done only be ***combining data generated over time, by different groups, looking at different features (independent variables) and using different language***
- To support discovery, the ***aggregation of these different data sets must be legitimate***

- ***The role of data is central to e-Science and these challenges must be met***