Scientific Data Curation and the Grid

David Boyd
CLRC e-Science Centre
http://www.e-science.clrc.ac.uk/
d.r.s.boyd@rl.ac.uk

Outline

- Some perspectives on scientific data
- Requirements of scientific data curation
- How the Grid can help
- Some current work at CLRC

The problem is growing . . .

- we have a rapidly increasing capability to generate data in many different formats in the physical and life sciences
  - environmental science, astronomy, particle physics, genomics, proteomics, clinical trials, . . .
- in all these areas data volumes are tending towards petabytes
- the facilities used to generate data are increasingly expensive
  - satellites, synchrotrons, supercomputers, telescopes, . . .
- some data is irreplaceable
  - measurements, surveys, . . .
- much future potential will lie in analyses which combine data across traditional disciplinary boundaries
- access to data will increasingly be needed on a global scale

Where is the value in data?

- in the **information** it conveys, not just in the bits
- in context and relationships as well as content
- **metadata** provides a key to this information - enhances value
- this needs to be recorded and preserved with the data
- if not captured digitally in (near) real time then it may never be recovered - need skills to do this
- need **ontologies** and **thesauri** to define and relate concepts
- well documented **data formats** help to maintain value
- use of globally accepted **standards** aids accessibility
- value also in **models** and **theories** needed to interpret data
- these are encoded in **software**
- how to preserve software as well as data?

OAIS Reference Model . . .

"it is harder to preserve working software than to preserve information in digital or hardcopy forms"

Whose data is it anyway?

- data ownership often (usually?) unclear
  - with ownership goes responsibility
    - so responsibility for data curation usually unclear
      - so it often doesn't get done
- one problem is the initial and ongoing cost of curating data
  - though often less than the cost of (re)generating data
    - but may not be funded as part of a grant
- another problem is lack of a culture of data curation
  - benefit not necessarily to the original producer
- needs a fundamental change of attitude by "sponsors"
  - recognition that data is a significant long term asset
  - willingness to fund long term data curation activities

Requirements of scientific data curation

- traditionally “keeping a collection”
- management
  - security - keeping data safe from threats
  - integrity - ensuring authenticity and completeness of the data
  - preservation - over time and technological change
  - acceptance of responsibility for data
- accessibility
  - exercising access control where necessary
  - providing knowledge of existence
  - enabling exploration of related information (metadata)
  - enabling retrieval of content (data)
  - preserving ability to read (physically) and understand (logically)

How the Grid can help

- offers a mechanism for controlling access
  - authentication and authorisation
- makes existence and location of data resources more visible
  - hierarchical directory service, global visibility
- provides easier access to data
  - single sign-on, location transparency, controlled replication
- integrates processes and data to generate enhanced value data
  - recalibration, filtering, transforming, post-processing . . .
- moves data to application or application to data
- facilitates distributed collaborative working by sharing data
- . . .

Some current work at CLRC

- Developing a Web/Grid portal for accessing scientific data
  - common top level metadata schema for scientific data
  - links to existing application specific metadata
  - ability to search distributed metadata catalogues
  - transparent access to distributed data resources
  - GUI access for people and API access for software/agents
- Operating a large data archive (~PB scale)
  - Grid-accessible primary data storage resource
  - provides data security, integrity, preservation
  - offers data archiving service to other primary data resources
- Participating in global standards activities
  - Global Grid Forum, WWW Consortium, ISO, etc, . . .