Curation of Scientific Data: Challenges for Institutions & their Repositories

Chris Rusbridge
The Adaptable Repository
3 May 2007, Sydney
Contents

- Science and digital curation
- Why are data important?
- What kinds of data?
- What to do with data: frontiers of practice
- Repository challenges
- Changing practice
Digital Curation Centre Mission

“The over-riding purpose of the DCC is to support and promote continuing improvement in the quality of data curation, and of associated digital preservation”
A centre of expertise in data curation and preservation

APSR 2007
Records of science

• Data increasingly important as evidence
  • Key part of the scholarly record (public good)
    • Unrepeatable observations & experiments
  • Experimental verifiability (the basis of science)
    • Would Chang retractions have been reduced if his first data were available?
• Allows additional interpretations
• Legal and compliance
  • See APSR/AERES report for good examples

What kinds of data?

- Observations
  - eg UARS (Upper Atmosphere) Level 0: telemetry
  - UARS Level 1: measured physical parameters (post calibration?)
- Derived data
  - UARS Level 2: calculated geophysical? profiles
  - UARS level 3: gridded, interpolated?
- Combined data
- Crafted data
  - Eg annotated gene/protein databases
- Descriptive (meta)data
Retaining research data means…

- Data secure against loss (within group)
- Communal repository (secure data store)
- Re-usable, sharable information
- As above, plus active curation (eg bio-informatics)
- Long term preservation of information

- Be clear what you are trying to do!
... or the data trajectory is...

- Hard drive → lost (crash)
- Hard drive → DVD → Cardboard box → Loft → Skip/dumpster → lost

- Sometimes this is a very bad thing
- Sometimes these are the right options!
Long term bit storage…

• A solved problem? Just requires well-understood good data management practices?

• Wrong! For very large datasets over very long time, there are significant problems…

How Well Must We Preserve?

Keep a petabyte for a century
- With 50% chance of remaining completely undamaged

Consider each bit decaying independently
- Analogy with radioactive decay

That's a bit half-life of $10^{18}$ years
- One hundred million times the age of the universe

That's a very demanding requirement
- Hard to measure
- Even very unlikely faults will matter a lot
What to do about curation

• Build curation/reusability into science workflow
  • Curation begins before creation
  • What’s easy at first becomes (impossibly) hard later
  • Describe data (metadata schemas, “representation info”, etc)
  • Keep experimental parameters (technical, who, what, when, where)
  • Keep ability to process
  • Keep data!
What to do about curation - 2

• Use standard/agreed formats for data
• Make ownership & restrictions clear, & explain how to cite data
• Offer for deposit in institutional or discipline repository
  • Appraisal and selection essential
  • Possible time-limited embargos
• “Publish” data in support of articles
Internet Archaeology: publication with data
Database as book...

- Buneman (early pilot) work on IUPHAR database
- MySQL to XML database
  - Historic to logical schema
- XML via XSLT to LaTeX
The StORe vision

- Seamless transport from research data to research publications and vice versa
- Bi-directional links proven in social science e-research but capable of export to other disciplines

http://jiscstore.jot.com/WikiHome/
What are the reusability issues?

- Data not neutral to hypothesis
- Hard to know the risks & pitfalls of a particular dataset
- Data not self-describing: hard to find appropriate data (but see Murray-Rust on Googling InChI etc)
- Hard to “understand” data once found
  - Really need information, not data!
- Hard to use data once understood
Context

• Data meaningless without context
  • Metadata of many kinds
  • Representation information… from data to information
  • Linkage and connection between datasets

• Provenance
  • Authenticity/integrity
  • Computational lineage
Access and re-use

- Ethics and rights control access
  - Weak in expressing this long-term
- Collaboration tools
  - Annotation, discussion, review (see DART…)
  - Re-use leading to change and development
- “Publication”
  - Not just in “print”
  - Underlying data should be “published”, too
Data citation issues…

• Citation for human readers and machine use cases
• Granularity: database, record, item
• Citation of changing objects
  • Version change (eg W3C practice: no version = latest, vs bibliographic: no version = first)
  • An efficient way to reference and access “archived” past states of more rapidly changing dataset, eg Genomics… datasets that result from the combined work of curators, or contain opinions or facts likely to change (work in progress, Buneman et al)
• Standards conflict and immature (NLM best?)

• Citation ESSENTIAL for motivating quality academic work on data management and curation
Who does data curation?

- Individuals
- Departments or groups
- Institutions, often through libraries
- Communities
- Disciplines
- Publishers
- National services
- Other 3rd parties...
Who are the curation players?
Repository challenges

• Data are different: you’ll need some domain knowledge
• Appraisal/selection harder
• Broader range of formats
  • Appropriate “standards” for longevity? XML-based?
• What metadata are needed?
  • Descriptive, to find the dataset
  • Context and background
  • Provenance
  • “Representation information” to connect data to information (whatever gives meaning to data)
<table>
<thead>
<tr>
<th>Vote #</th>
<th>Vote Date</th>
<th>Democrats</th>
<th>Republicans</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>1</td>
<td>January 25, 1989, 5:31 PM</td>
<td>55</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>January 25, 1989, 5:49 PM</td>
<td>55</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>January 25, 1989, 6:02 PM</td>
<td>55</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>January 31, 1989, 5:00 PM</td>
<td>55</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>January 31, 1989, 5:31 PM</td>
<td>55</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>January 31, 1989, 5:43 PM</td>
<td>55</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>February 2, 1989, 4:46 PM</td>
<td>55</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>February 2, 1989, 5:46 PM</td>
<td>55</td>
<td>0</td>
</tr>
</tbody>
</table>

Codebook file for Senate number 101, session 1.

- Column Number: 1-2
  - State Code
  - Column Number: 3-4
  - District Code
  - Column Number: 5-8

Data from MIT DSpace Political Science
Repository challenges - 2

• May distort your repository
  • Size
  • Number of objects
  • Rate of deposit
  • Nature of use

• Databases may be dynamic
• Databases may need to be accessed in situ
• Rights and ethical limitations hard to describe and enforce
• Need to build links to publications (cf StORe)
• Need to build discipline links across repositories…
Cultural change

• If we build it, will they come? NO!!
• Outreach important: communication with scientists and researchers is hard graft
• Cultural change to new approach requires more:
  • Incentives, rewards and mandates
  • Successful exemplars (well publicised)
  • Discipline-oriented approach (one size does not fit all)
Australian context?

- In the emerging context of the Research Quality Framework, and the expected National Collaborative Research Infrastructure Strategy, curation can only increase in importance!
Thank you
c.rusbridge@ed.ac.uk