Curation of Scientific Data:

Challenges for Institutions & their Repositories

Chris Rusbridge

The Adaptable Repository

3 May 2007, Sydney

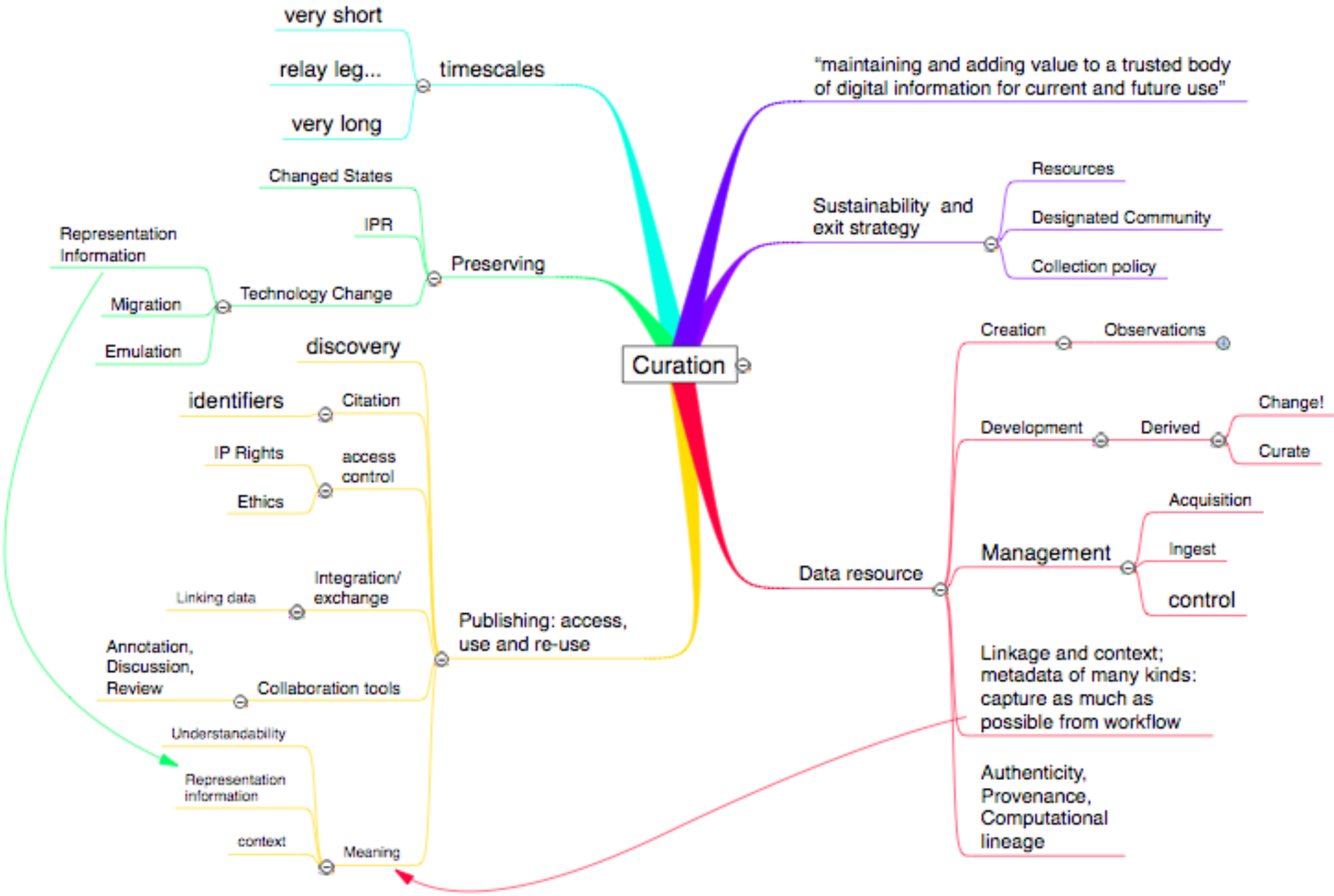
Funded by:
JISC
e-Science

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"


Curation
timescales
very short
relay leg...
very long
Preserving
Changed States
IPR
Technology Change
Representation Information
Migration
Emulation
Publishing: access, use and re-use
discovery
identifiers
Citation
access control
IP Rights
Ethics
Integration/ exchange
Linking data
Collaboration tools
Annotation, Discussion, Review
Meaning
Understandability
Representation information
context
"maintaining and adding value to a trusted body of digital information for current and future use"
Sustainability and exit strategy
Resources
Designated Community
Collection policy
Data resource
Creation
Observations
Development
Derived
Change!
Curate
Management
Acquisition
Ingest
control
Linkage and context; metadata of many kinds: capture as much as possible from workflow
Authenticity, Provenance, Computational lineage

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

CHANG, G., ROTH, C. B., REYES, C. L., PORNILLOS, O., CHEN, Y.-J. & CHEN, A. P. (2006) Retraction of Pornillos et al., Science 310 (5756) 1950-1953. Retraction of Reyes and Chang, Science 308 (5724) 1028-1031. Retraction of Chang and Roth, Science 293 (5536) 1793-1800. Science Magazine, 314. http://www.sciencemag.org/cgi/content/full/314/5807/1875b

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 → los

- 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…

BAKER, M., SHAH, M., ROSENTHAL, D. S. H., ROUSSOPOLOUS, M., MANIATIS, P., GIULI, T. J. & BUNGALE, P. (2006) A Fresh Look at the Reliability of Long-term Digital Storage. EuroSys '06. Leuven, Belgium, ACM.

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

•Slide from David Rosenthal, LOCKSS

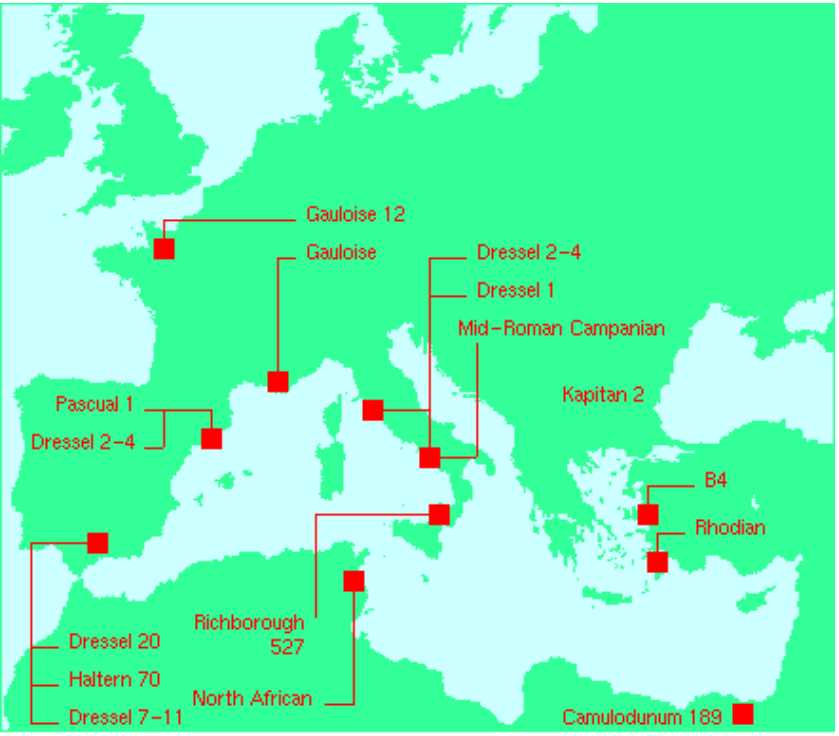
What to do about curation

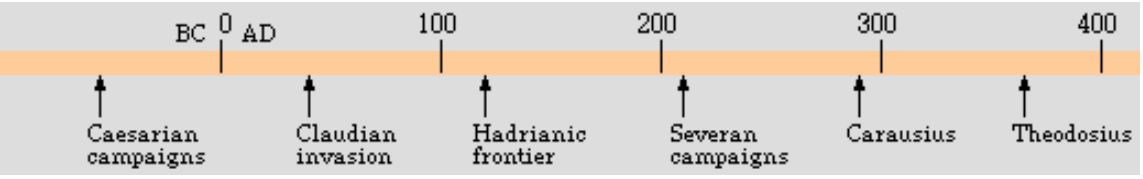
- Build curation/reusability into science workflow
  - Curation begins before creation
  - What's easy at first becomes (impossibly) hard later
  - Describe data (metadata schemas, "representation info", etc)
  - Keep experimental parameters (technical, who, what, when, where)
  - Keep ability to process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

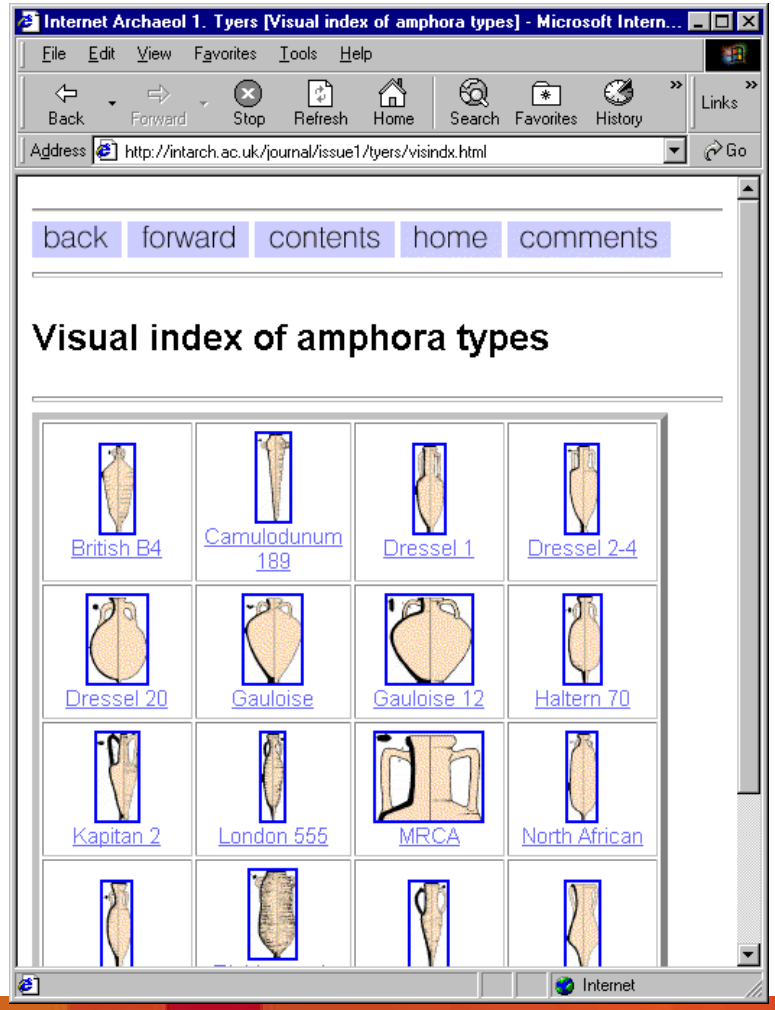
What to do about curation - 2

- 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

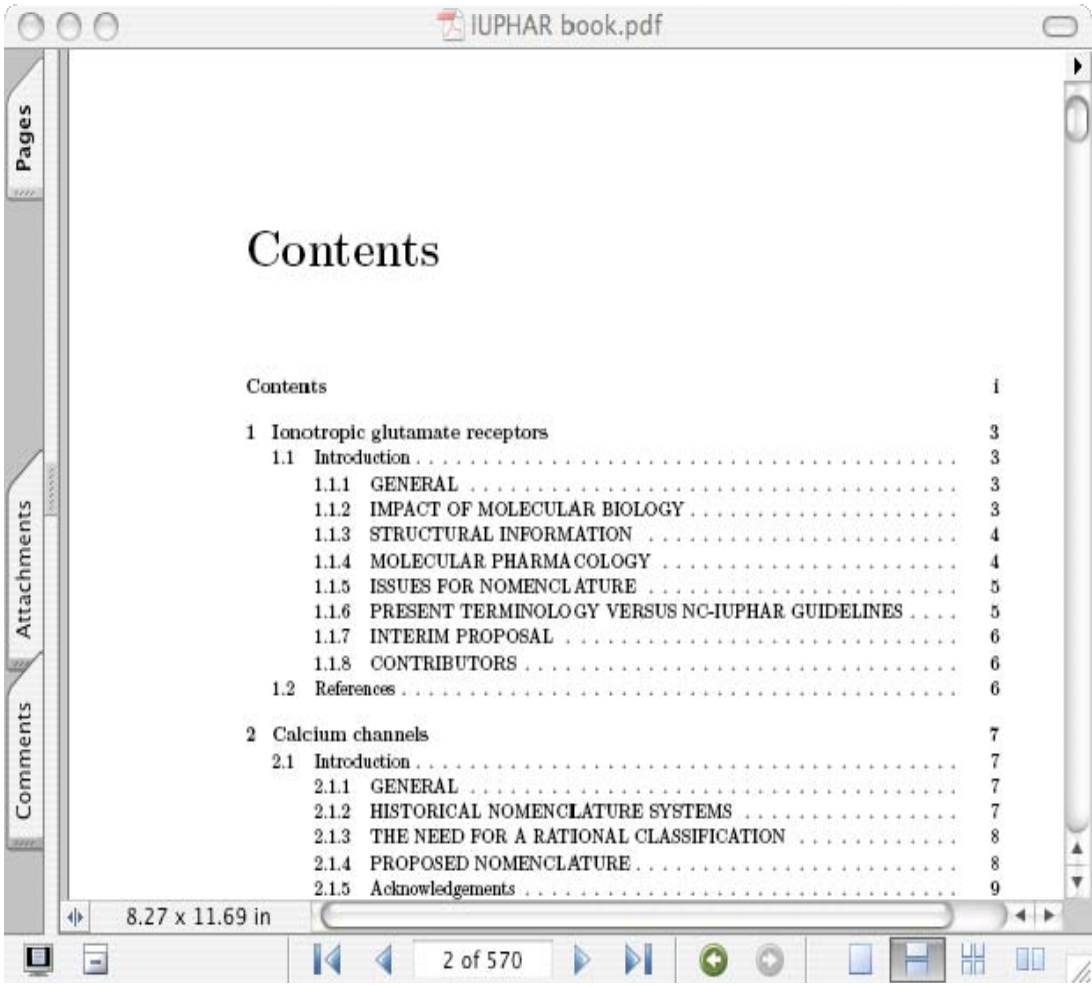






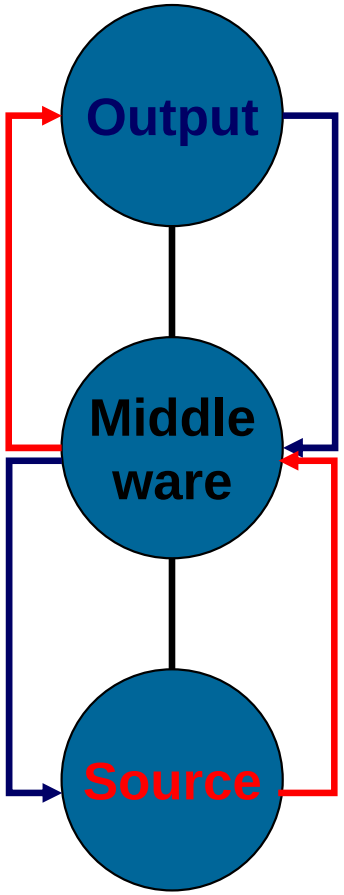
Database as book…

- 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…

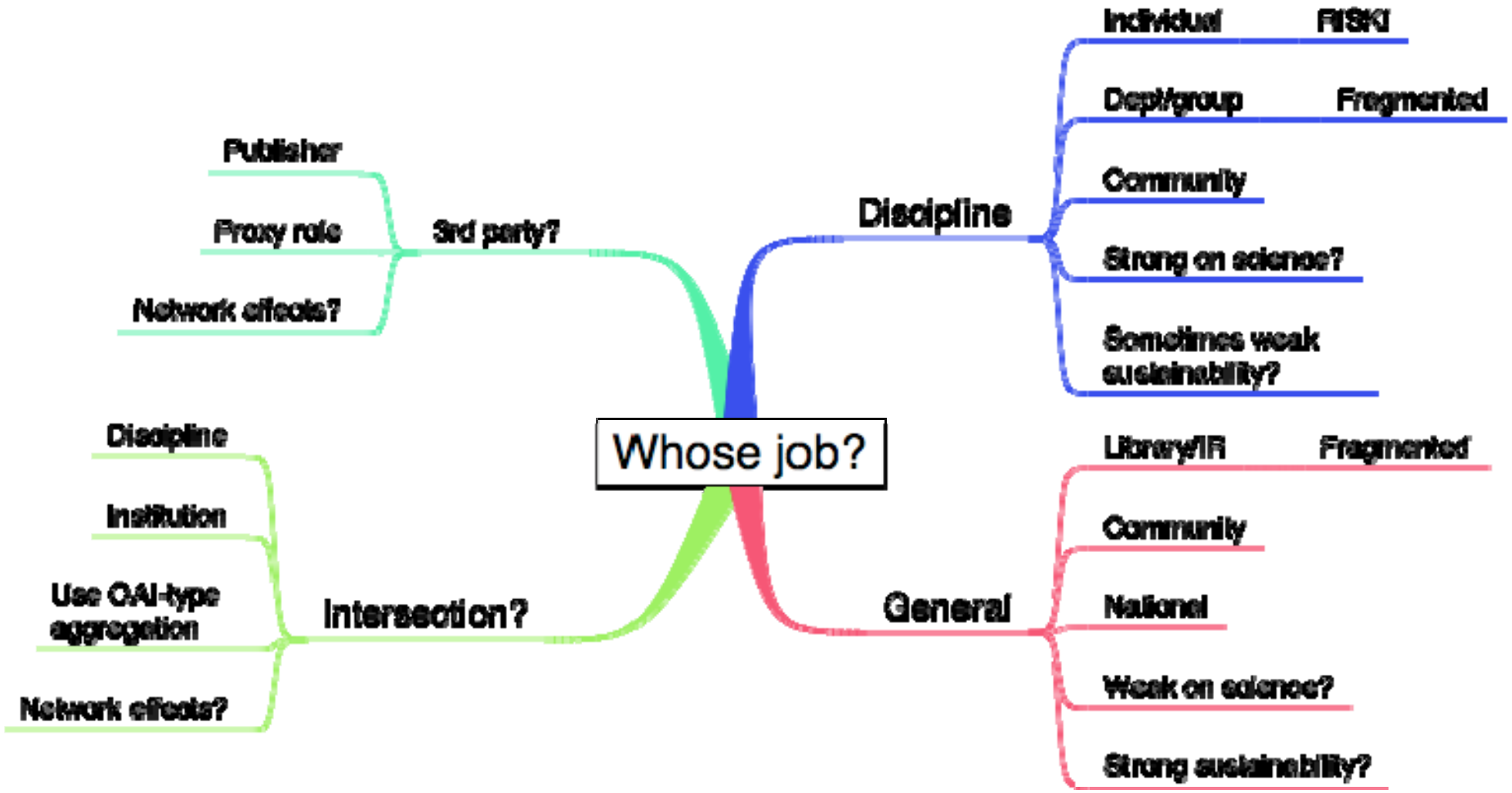
- Citation for human readers and machine use cases
- Granularity: database, record, item
- Citation of changing objects
  - Version change (eg W3C practice: no version = latest, vs bibliographic: no version = first)
  - An efficient way to reference and access “archived” past states of more rapidly changing 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?

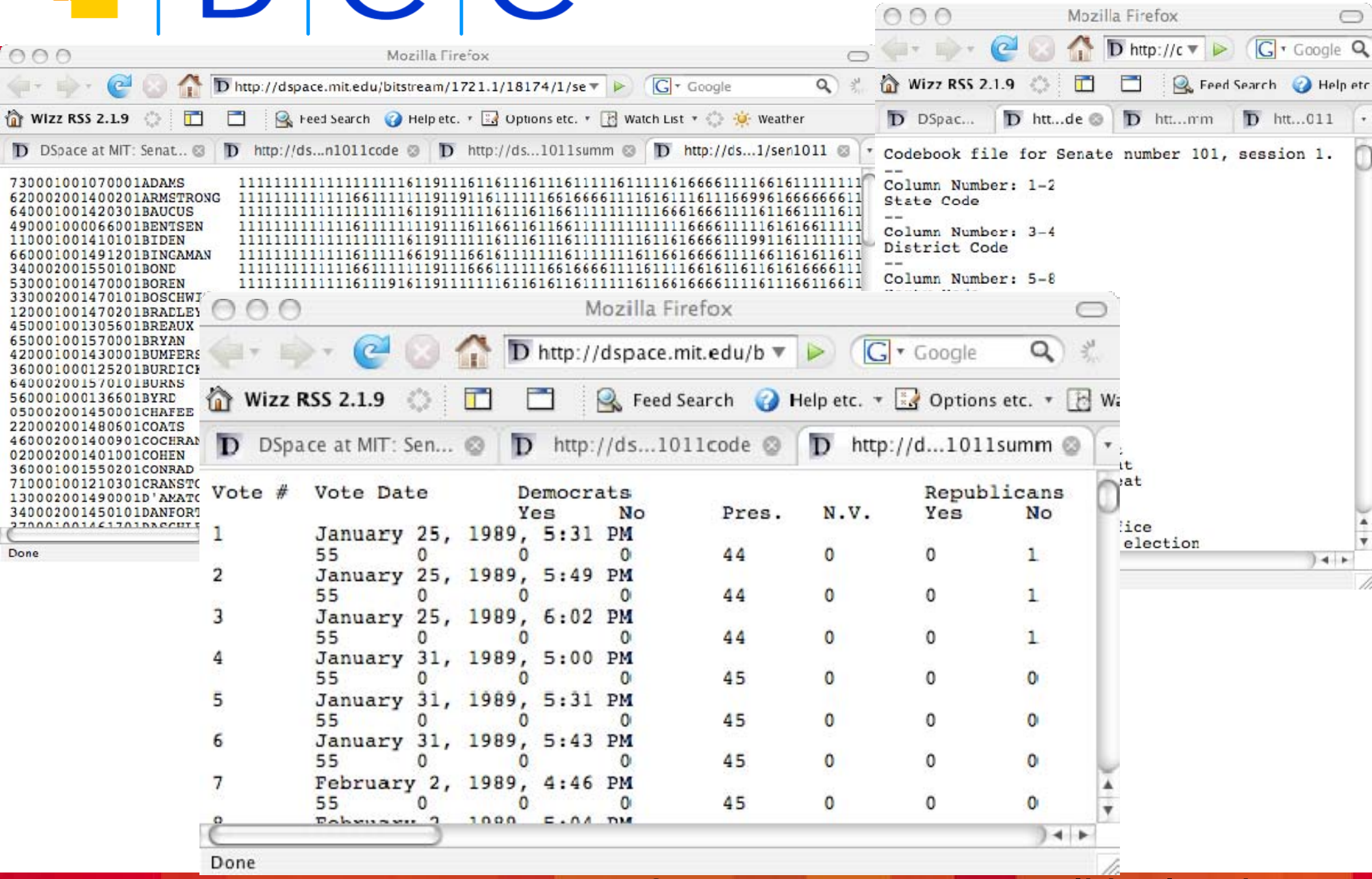


Repository challenges

- 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)

DCC

a centre of expertise in data curation and preservation



APSR 2007

•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