Disciplinary and Institutional Perspectives on Digital Curation

Michael Day, Colin Neilson, *Alexander Ball,* Rosemary Russell

April 2, 2009
DigCCurr, Chapel Hill, NC

UKOLN is supported by: JISC MLA Museums Libraries Archives Council

UNIVERSITY OF BATH

www.ukoln.ac.uk

A centre of expertise in digital information management

Curation and context (1)

- Digital curation is largely focused on developing infrastructures and services that support the continued use and reuse of research data
  - Preserving the scientific record and documentary heritage(stewardship)
  - Essential for the validation of research
  - Enables the creation of new science and scholarship
- Understanding the *context* in which this data is generated is extremely important
  - Different disciplinary research paradigms
  - Different incentives for sharing data

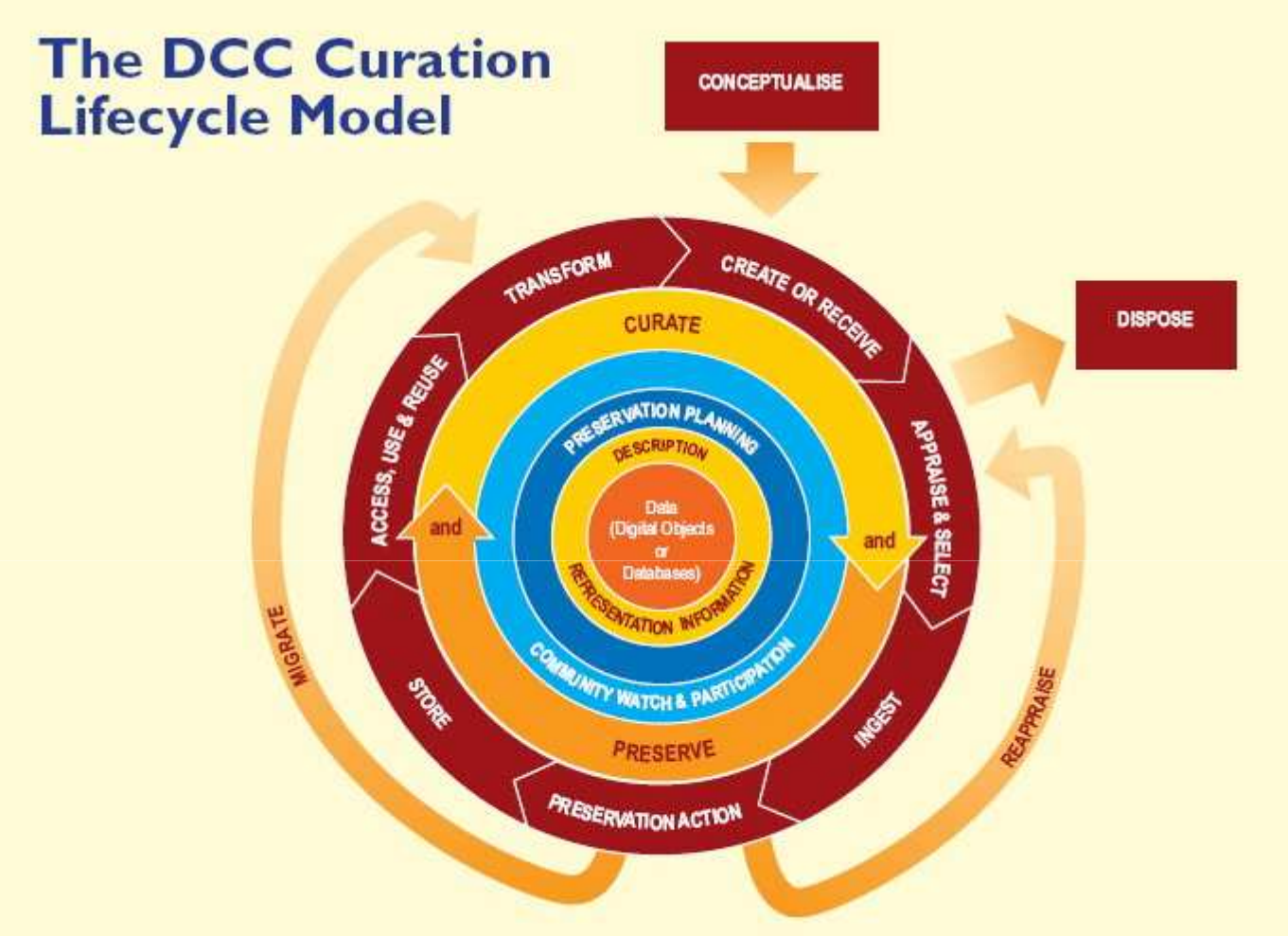
Curation and context (2)

- Data sharing not enough
  - Data need to be made available in ways that facilitate high-throughput reuse
  - Open Science agenda
- How do we capture the context(s) of research?
  - Not just papers and data, but Web-sites, annotation services, blogs, wikis, etc.
  - Importance of recording provenance

Curation and context (3)

- Need for frameworks that support the understanding curation in a generic way
  - Identifying core tasks and functions
- Need to be flexible enough to reflect the diversity of disciplinary practice
  - Disciplines (and subdisciplines) evolve, and there are big differences in data management practices and data sharing motivations
- DCC Lifecycle Model: http://www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf
  - Generic tool for supporting curation planning, identifying risks and strategies

DCC



UKOLN

UNIVERSITY OF BATH

www.ukoln.ac.uk

A centre of expertise in digital information management

Curation and context (4)

- It is important for curators to be able to work directly with researchers and research groups
  - Deeper understanding of research motivations and processes
  - Knowledge of data sharing norms within disciplines

Responsibility for curation (1)

- Many potential stakeholders
  - Dealing with Data report (2007) identified: scientists, institutions, data centres, the users of data, funding bodies and publishers
  - Also emerging group of 'data scientists'
- The potential for duplication of effort and confusion is high
- All of these probably have some kind of role ... so how should we co-ordinate?
- Specifically, should curation be the responsibility of the discipline or the institution?

Responsibility for curation (2)

- Institutional role in stewardship
  - Institutional Repository paradigm
  - Management of institutional assets
- Disciplinary roles
  - Keeping Research Data Safe (2008) report notes that, in practice, data is more often dealt with by discipline-based consortia
- Bottom-up approaches to curation work well in some domains – but not in all
  - Need to understand domain differences
  - DCC SCARP studies reveal much complexity

Responsibility for curation (3)

- A role for institutions?
  - They have traditionally had a very important role (e.g., research libraries)
  - Currently are major supporters (and hosts) of Institutional Repositories
  - Potential skills gap WRT data:
    - We need to think about the status and skills of data curators
    - Digital Curation Education, DCC Curation 101, WePreserve (Europe)

DCC SCARP Project

- Investigating disciplinary attitudes and approaches to aspects of data management
  - **S**haring, **C**uration **A**nd **R**e-use, **P**reservation
- Using a range of methods:
  - Surveys
  - Literature reviews
  - Ten immersive case studies in selected discipline areas

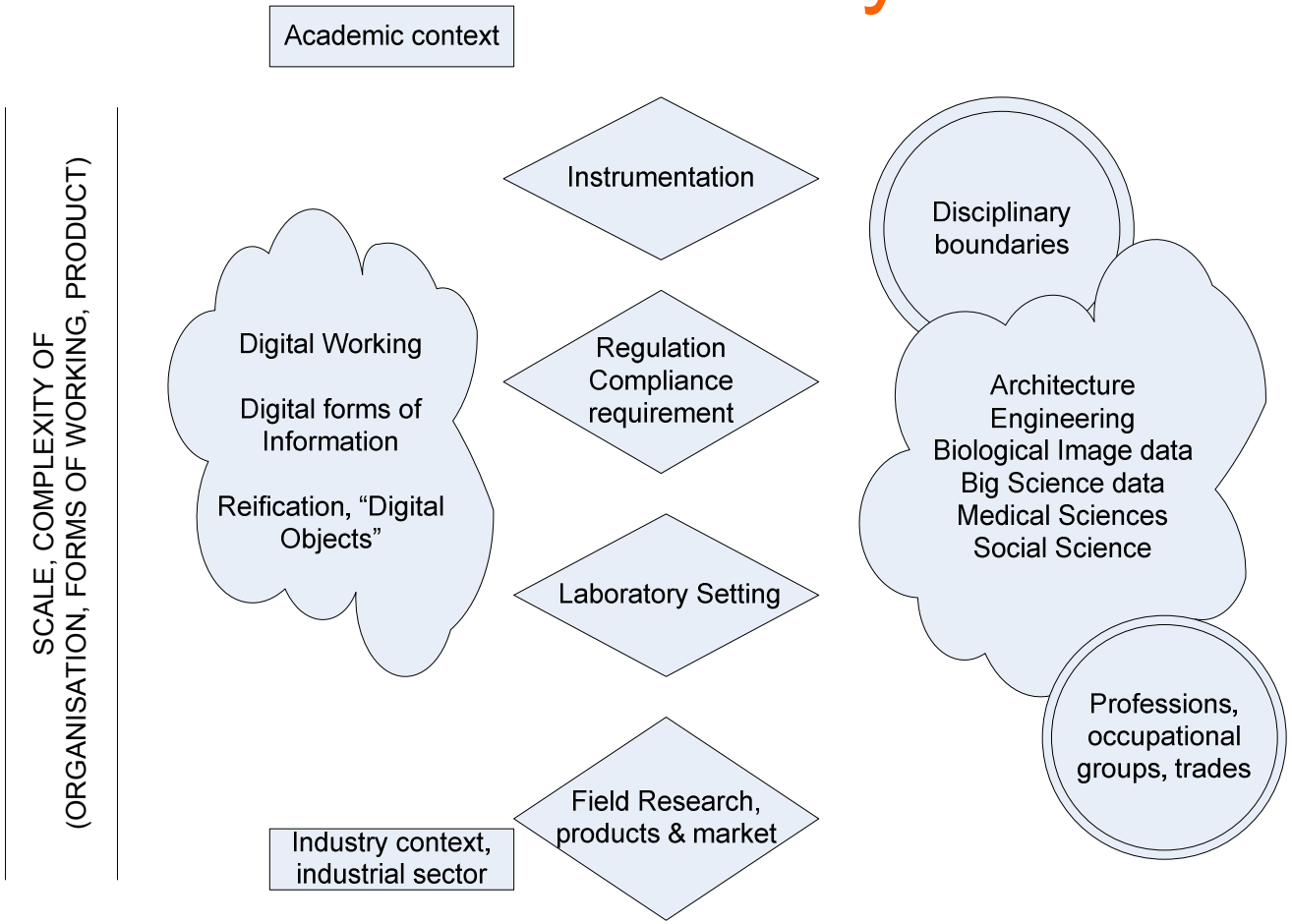
SCARP Project: Case Studies

1. Architecture (Colin Neilson)
2. Engineering (Colin Neilson)
3. Neuroimaging in Psychiatry (Angus Whyte)
4. Video data in Social Sciences (Angus Whyte)
5. Public engagement in the life sciences – curating the data and translating the methods, Social Sciences (Angus Whyte)
6. Earth Observation Data (Esther Conway)
7. Atmospheric Sciences (Esther Conway)
8. Astronomy data reuse survey (Malcolm Currie)
9. Astronomy – amateur astronomy data (Malcolm Currie)
10. Biology – Edinburgh Mouse Atlas Project (Elizabeth Fairly)
11. Community medicine – Data Curation Lifecycle in the Context of Telecare

Main Challenges

- Promoting digital curation in different cultures
- Identifying any common factors so that the differences between disciplines are better understood

Some factors looked at by SCARP



Trends observed so far

- Need to find the right level of genericity
  - Tools such as DRAMBORA (Digital Repository Audit Method Based on Risk Assessment) need to be adapted for scale
  - DCC Digital Curation Lifecycle Model and Preservation Analysis seem to be widely applicable
  - Differences at research team level just as important as disciplinary differences

Application Profiles

- In the beginning there was Dublin Core
- Then came the Scholarly Works Application Profile
  - Entity model to permit efficient data management and complex user interaction
  - (Still not used widely)

Application Profiles

- Repositories with more diverse content – more application profiles
  - Still images
  - Time based media
  - Geospatial data
  - Learning objects?
  - Scientific data?

Scientific Data Application Profile Scoping Study

- Broad consideration of data
  - Generated from applying *instrument* to *system*
  - Generated from applying *process* to *existing data*
- Focus on discovery to delivery
  - Interdisciplinary cross-searching

Contexts

- Culture of collaboration versus culture of independence
- Centralized versus distributed
- Stability versus rapid evolution

Trends observed

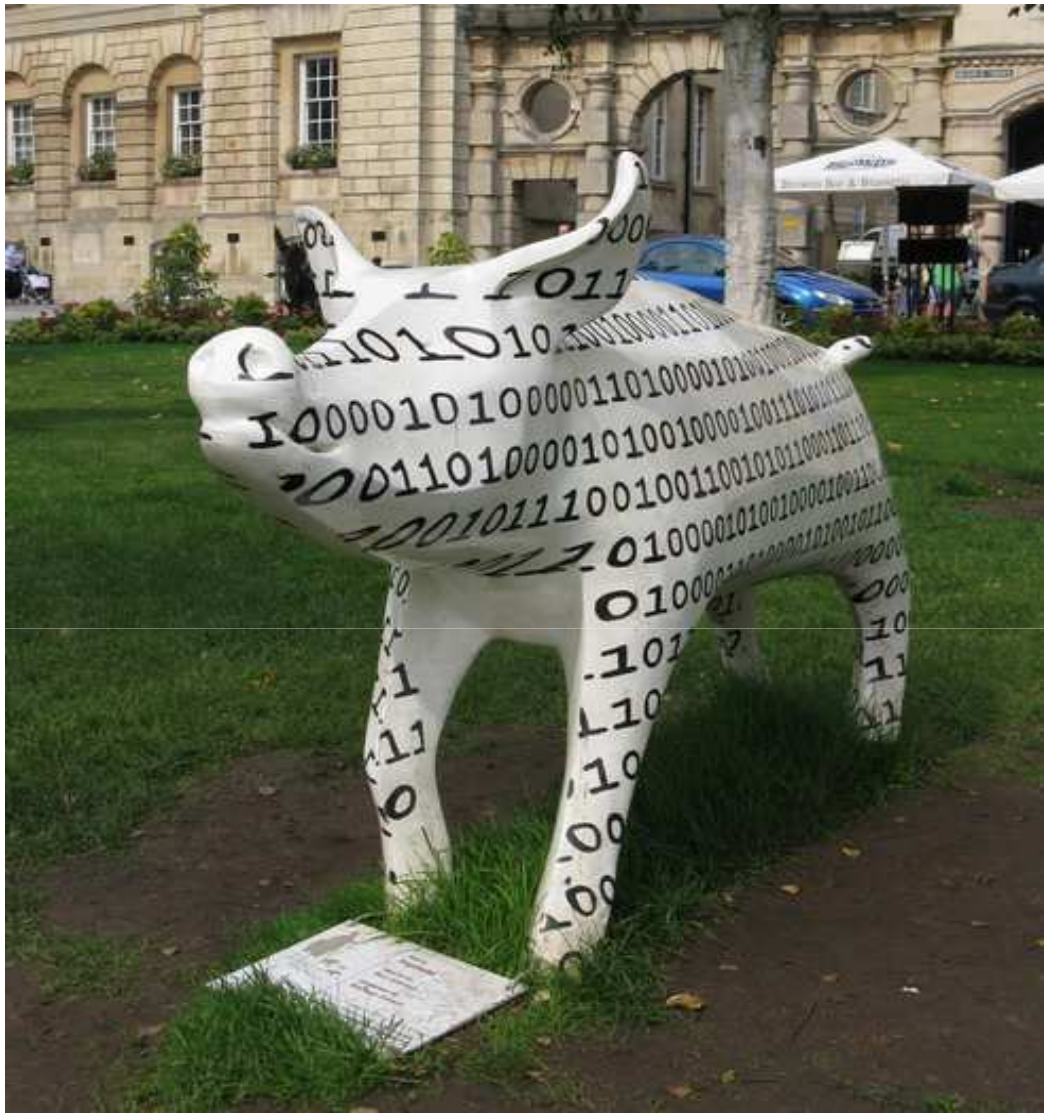
- Common needs for discovery to delivery
  - Identifiers, versioning
  - Provenance, responsibility
  - What the data are about
  - How the data were generated
  - How one might access the data

Trends observed

- This isn't enough for all applications
  - Combining different datasets to form time series data, e.g. census data
  - Data mining across many datasets
- Still a need for detailed, discipline specific metadata

Final thoughts

- Current scientific reward structures do not support either data curation or open science
  - Funding bodies can 'mandate' (and in some cases fund) Principal Investigators to maintain data and make it available
  - Without sustainable infrastructures and available expertise, however, this can only be a short term solution
- Some progress:
  - DCC SCARP project
  - Scientific Data Application Profile study



DCC

Thank-you for your attention!

"Pigabyte"

From: King Bladud's Pigs in Bath (public art project), Summer 2008

UNIVERSITY OF BATH

UKOLN

Acknowledgments

- *UKOLN* is funded by the Joint Information Systems Committee (JISC) of the UK higher and further education funding councils, the Museums, Libraries and Archives Council (MLA) , as well as by project funding from the JISC, the European Union, and other sources. UKOLN also receives support from the University of Bath, where it is based.
- http://www.ukoln.ac.uk/
- The *Digital Curation Centre* is supported by the JISC and the UK e-Science Core Programme.
- http://www.dcc.ac.uk/