The Digital Curation Centre

Michael Day
Digital Curation Centre
UKOLN, University of Bath
http://www.ukoln.ac.uk/

MRC Human Genetics Unit, Western General Hospital, Edinburgh, 14 June 2005

Funded by:

Presentation overview

- Digital curation and its importance
- The Digital Curation Centre:
  - Structure
  - Overview of activities
- Some current issues:
  - Metadata
  - Institutional repositories and open access
  - Trust

What is digital curation?

- New(ish) term, from science data world (e.g. bioinformatics)
- Reflects those extra things that need to be done to facilitate access and reuse
- "... managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and reuse" - Philip Lord, *et al*. (2004)
- "Maintaining and adding value to a trusted body of information for current and future use" -- DCC presentation at CNI (2005)

What is digital preservation?

- Dealing with the potential technical problems that impede continued access to all types of digital resource
- No longer possible to place physical artefact on a shelf and ignore for 100+ years
- Sometimes seen as focused on the maintenance of specific object over time (e.g., a facet of curation)
- But older definitions emphasise that it is not just a technical problem:
  - "... The planning, resource allocation, and application of preservation methods and technologies to ensure that digital information of continuing value remains accessible and usable" - Margaret Hedstrom (1998)

Why is it a problem? (1)

- An increasing flood of 'born-digital' data
  - The World Wide Web
    - Comprises billions of pages + "deep Web"
    - Internet Archive = >1 petabyte, and growing @ 20 Tb. per month (http://www.archive.org/)
  - Data deluge in science and engineering
    - Petabytes generated by high throughput instruments, streamed from sensors and satellites, etc.
    - Data-driven science, e-science, cyberinfrastructure, ...
  - 5 exabytes of *new* information created in 2002:
    - http://www.sims.berkeley.edu/research/projects/how-much-info-2003/

Why is it a problem? (2)

- Need for (open) access to this data
  - Results in added scientific value
  - New analytic techniques
  - 2004 - OECD member states endorsed the principle that publicly funded research data should be openly available to the maximum extent possible
- Interoperability
  - Technical and cultural

Digital Curation Centre (1)

- Funded from 2004 for three years by the JISC and the e-Science Core Programme
- Main aim: "continuing improvement in the quality of data curation and digital preservation"
- Will focus on all aspects of the research process, e.g. from data creation to publication and beyond, also on the work of repositories and data archives
- Not itself a digital repository, but offering outreach and practical services to assist those who curate data …

Digital Curation Centre (2)

- Some organisational basics:
  - Director: Chris Rusbridge (University of Edinburgh)
  - *Research* team led by Professor Peter Buneman (School of Informatics, University of Edinburgh)
  - *Development* team led by Dr David Giaretta (Astronomical Software and Services, CCLRC)
  - *Advisory services* team led by Professor Seamus Ross (HATII, University of Glasgow)
  - *Outreach* team led by Dr Liz Lyon (UKOLN, University of Bath)

DCC requirements analysis

- Commissioned from Leona Carpenter
- Desk research, focus groups and interviews
- Taxonomy of stakeholders (the creators, curators and users of digital information)

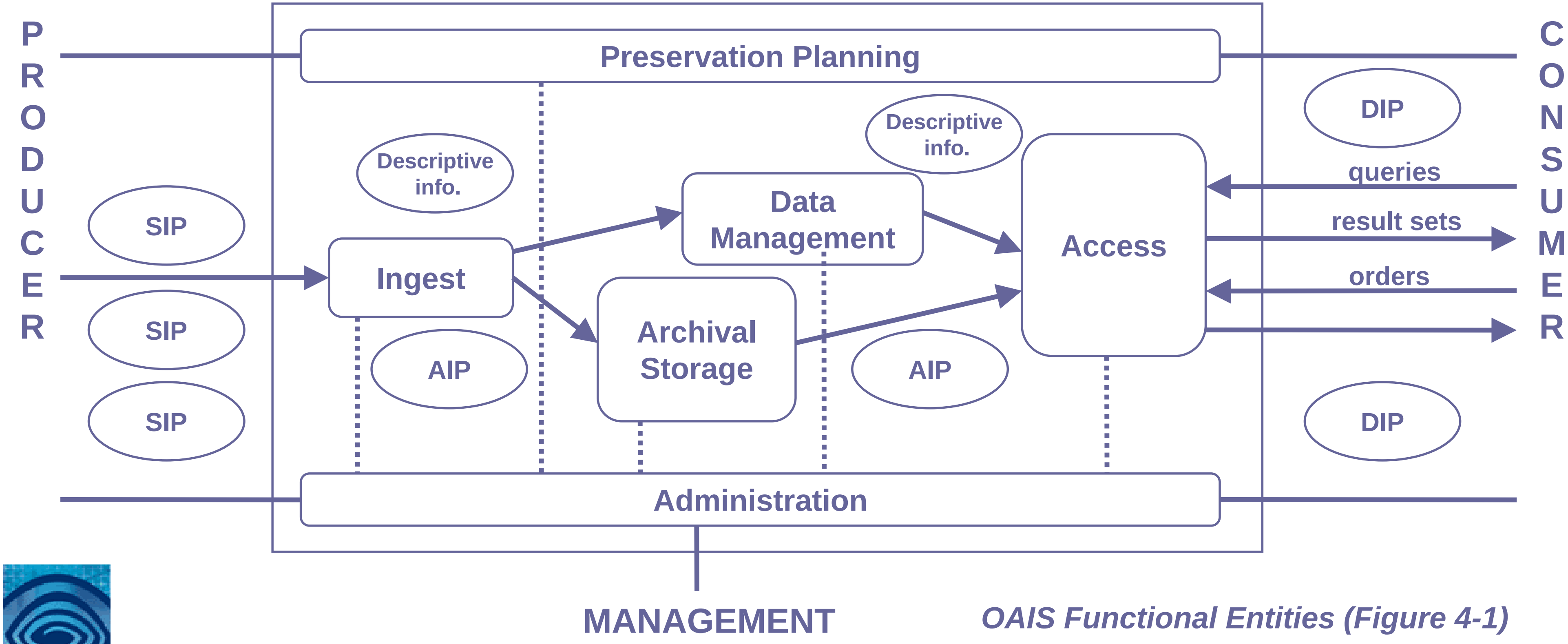
DCC research

- Curating databases
  - Publishing and integrating data
  - Database provenance and annotation
  - Preserving past states of volatile databases
  - Citing data
- Automated extraction of metadata
- Cost-benefit analysis of curation
- Networks of repositories
- Rights and responsibilities

DCC development (1)

- Work based on concepts outlined by the Reference Model for an Open Archival Information System (OAIS)
- Current focus on Representation Information
  - The information required (metadata, documentation, community knowledge) to render objects
  - Trusted repositories of Representation Information (link with file format registries like GDFR) - pilot
  - Persistent identifiers

DCC development (2)



OAIS Functional Entities (Figure 4-1)

DCC services

- Advisory services
  - Queries to HELPDESK@dcc.ac.uk
  - Site visits
- Information
  - Briefing papers (e.g. FOI by Mags McGinley)
  - Curation manual (invited authors, peer-reviewed)
- Events
  - For example, Workshop on Institutional Repositories, Cambridge, 6 July 2005

DCC outreach

- Examples:
  - Web site: http://www.dcc.ac.uk/
  - International Journal of Digital Curation (IJDC): http://www.ijdc.net/
  - 1st DCC Conference, Bath, 29-30 September 2005
  - Network of Associates

Some issues (1)

- Metadata:
  - Vitally important for recording the characteristics and behaviour of objects, agents and processes
    - Descriptive metadata
    - Technical metadata (hardware and software environments, information about formats, etc.)
    - Structural metadata
    - Administrative metadata
  - Wide range of initiatives in this area (PREMIS DD)
  - DCC curation manual chapter, scientific metadata model (CCLRC data portal), ...

Some issues (2)

- Institutional repositories:
  - The impact of new RCUK policies
  - Research outputs, including data (?)
  - Role in preservation
    - e.g., National Institutes of Health policy requesting grantees to submit papers to PubMed Central
  - Disaggregated model
    - Not all repositories will have preservation responsibilities
    - Possible need for mechanisms for transferring content to third parties, e.g. national libraries

Some issues (3)

- Trusted repositories:
  - Attributes and responsibilities of 'trusted repositories' defined by RLG and OCLC working group (2002)
    - Builds on 1996 Task Force report and OAIS model
    - Attributes include the viability and financial sustainability of the organisation, and the need for accountability
    - Question whether these (and other criteria) could be used as a basis for certification is being explored by the Task Force on Digital Repository Certification, supported by RLG and the National Archives and Records Administration (NARA)

Acknowledgements

The Digital Curation Centre is funded by the Joint Information Systems Committee (JISC) of the UK higher and further education funding councils and the e-Science Core Programme of the UK research councils. The consortium comprises the University of Edinburgh (lead partner), the University of Glasgow, the Council for the Central Laboratory of the Research Councils, and UKOLN:

http://www.dcc.ac.uk/

UKOLN is funded by the Council for Museums, Libraries and Archives (MLA) and the JISC, 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/