Towards a framework for integrating digital preservation research

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

International Conference on Preservation of Digital Objects (iPres), Göttingen, Germany, 15-16 September 2005

Funded by: JISC e-Science dti

Session overview

- Some background
- Major research challenges
- Selected current initiatives
  - DELOS Network of Excellence on Digital Libraries, Digital Preservation Cluster
  - Digital Curation Centre research agenda
- Summing up

Some background

Beginnings (1)

- Task Force on the Archiving of Digital Information (1994-96)
  - Supported by the Commission on Preservation and Access and the Research Libraries Group
  - Final report published in May 1996:
    - A wide-ranging study that outlined the multiple challenges of preserving digital information (technical, legal, organisational, etc.) with recommendations for action
    - Has acted as a catalyst for research and development in the digital preservation domain

Beginnings (2)

- For example, in the United Kingdom:
  - 1st workshop on the Long Term Preservation of Electronic Materials (University of Warwick, November 1995)
  - Workshop outcomes:
    - Funding of studies into different aspects of the challenge:
      - Topics included: digital archaeology, data types and formats, preservation methods, etc.
    - Strategic priority in the 3rd phase of JISC's Electronic Libraries Programme
      - Cedars (CURL Exemplars in Digital Archives) project

Beginnings (3)

- "Warwick 2" workshop (March 1999)
- Identified co-ordination and collaboration as key issues
- Led directly to formation of:
  - Digital Preservation Coalition (from 2002)
    - Cross-sector membership organisation
- Indirectly to:
  - JISC Continuing Access and Digital Preservation Strategy, 2002-2005
  - Various JISC research programmes
  - Digital Curation Centre (from 2004)

The current situation (1)

- Increased high-level awareness of the digital preservation challenge, e.g.:
  - Unesco Charter on the Preservation of Digital Heritage
  - National initiatives like the DPC, NESTOR and NDIIPP
- Focus on the Reference Model for an Open Archival Information System (OAIS)
  - for identifying the high-level attributes of systems and information (metadata) requirements

The current situation (2)

- Much current R&D has a practical focus:
  - Topics include:
    - Appraisal, metadata and documentation, ingest and other repository workflows, authenticity and trust, audit and certification, formats and Representation Information
  - Operational service development
    - e.g., e-Depot, kopal, OCLC Digital Archive, PANDORA, Internet Archive, CDPC
  - The development and testing of software tools
    - e.g., DIAS, DSpace, Fedora, Heritrix, LOCKSS, PANDAS, Storage Resource Broker, etc.

The current situation (3)

- Good, useful, R&D work is being undertaken
- Need for co-ordination to avoid the possibility of fragmentation or duplication (but different perspectives are important)
- Need to regularly revisit research challenges

Some research challenges

Research challenges workshop (1)

- Workshop on Research Challenges in Digital Archiving and Long-Term Preservation, held in Washington, D.C. (April 2002)
  - Final report - "It's about time" (August 2003) Available: http://www.digitalpreservation.gov/
  - "… digital collections require curation and processing to ensure their longevity, protect their integrity, and enhance their value for use in the future"
  - Identified a research agenda …

Research challenges workshop (2)

- Research Agenda (key points):
  - Technical architectures
    - For dealing with extremely complex and dynamic objects, scaling to massive volumes of data (e.g., responding to the 'data deluge' in science)
  - The attributes of collections
    - Controlled curatorial processes: selection, organisation, description, quality-control, stewardship
  - Tools and technologies
    - Managing the evolution of tools, technologies and standards for ingest, identification, description, interoperability

Research challenges workshop (3)

- Research Agenda (continued):
  - Organisational, economic, policy issues
    - Need for a "deep infrastructure" of technical solutions, standards, trusted organisations, business models and skilled personnel
    - Identifying costs and benefits
    - Interaction with producers and other content creators

NSF-DELOS WG (1)

– NSF-DELOS Working Group on Digital Archiving and Preservation

- Report and recommendations - "Invest to save" (2003)
- Digital preservation encompasses organisational, legal, cultural, social and financial dimensions
- High-priority research area … need for commonly agreed research agenda to inform and co-ordinate existing work and take it forward

NSF-DELOS WG (2)

- Research Agenda (preliminary)
  - To "make possible new theoretical approaches, viable tools and methodologies needed to respond to the array of challenges created by technology evolution"
  - Urgency for practical solutions that can be applied now, but "must be balanced by the need to avoid quick fixes that defer, without resolving, the fundamental requirement to carry digital materials forward in a coherent, consistent, appropriate, authentic and affordable manner" (p. 8)
  - Research agenda deliberately excludes important policy, organisational and educational issues

NSF-DELOS WG (3)

- Research Agenda (topics)
  - Preservation strategies
    - » The roles of repositories and registries, physical media, rescue, documenting functionality and behaviour, ...
  - Re-engineering preservation processes
    - » Modelling of processes, automation, scalability, ...
  - Preservation of systems and technology
    - » Formats, complex entities, automated metadata capture, ...

Some existing work

DELOS Network of Excellence

- DELOS: well-established project in digital library research field
  - Thematic workshops and other events
  - Joint Working Groups with NSF's Digital Library Initiative
- DELOS NoE funded by the European Union
  - Sixth Framework Programme, from 2004
  - Network of Excellence - intended primarily for the *integration* of research
  - >40 research partners in first phase (JPA1)

DELOS aims and objectives

- *Overall aim: To carry out joint activities integrating research activities of European teams working in Digital Library related areas*
- To define unifying frameworks for the life-cycle of Digital Library information
- To build interoperable services and integrated content management
- To provide a forum where researchers, practitioners, and representatives of interested applications and industries can exchange ideas and experiences
- To promote an exchange programme to improve international cooperation in Digital Library research areas

Main DELOS research clusters

1. Digital library architectures
2. Information access and personalization
3. Audio-visual and non-traditional objects
4. User interface and visualization
5. Knowledge extraction and semantic interoperability
6. Digital preservation
7. Evaluation

Preservation cluster partners

- **Department of Software Technology and Interactive Systems, Vienna University of Technology, Austria**
- **Goettingen State and University Library, Germany (from JPA2)**
- **Historisch-Kulturwissenchaftliche Informationsverarbeitung, University at Cologne, Germany**
- **Humanities Advanced Technology and Information Institute (HATII), University of Glasgow, UK (cluster co-ordinator)**
- **Istituto di Studi per la Tutela dei Beni Archivistici e Librari, Università degli studi di Urbino "Carlo Bo," Italy**
- **Nationaal Archief, Netherlands**
- **Österreichische Nationalbibliothek, Austria (from JPA2)**
- **Phonogrammarchiv, Austrian Academy of Sciences, Austria**
- **UKOLN, University of Bath, UK**

Strategic goals

- To eliminate duplication of effort between research activities; the co-ordination and promotion of research, the sharing of knowledge and expertise
- Guidelines, methods and tools that enable the construction of preservation functionality within digital library activities
- The establishment of testbeds and valuation metrics, e.g. for testing preservation strategies
- To relate the digital preservation research agenda more directly with the development of exploitable product links

Phase 1 tasks, 2004-2005 (1)

- Testbed design for the testing and validation of different digital preservation strategies
  - Based on testbed framework developed by the Nationaal Archief in the Netherlands
  - Report on cluster Web site: http://www.dpc.delos.info/
- An evaluation of repository models
  - [ongoing, relates to work in WP5]
- Investigating the interaction of file format types and preservation strategies
  - In the context of registries of format information
  - Initial report on cluster Web site

Phase 1 tasks, 2004-2005 (2)

- Developing frameworks for documenting behaviour and functionality (using utility analysis)
  - Innovative work from the Vienna University of Technology team
  - Integrates into the testbed framework developed by the Nationaal Archief
  - Metrics to test preservation approaches
  - Paper available on cluster Web site: http://www.dpc.delos.info/
- The integration of preservation functionality into digital library design

Phase 1 tasks, 2004-2005 (3)

- Events:
  - Workshop on Digital repositories: interoperability and common services (jointly organised with the DELOS Knowledge Extraction and Semantic Interoperability cluster), Heraklion, Crete, 11-13 May 2005: http://www.ukoln.ac.uk/events/delos-rep-workshop/
  - Summer School on Digital Preservation in Digital Libraries, Sophia-Antipolis, France, 5-11 June 2005 http://www.dpc.delos.info/

Phase 2 tasks, 2005-2006 (1)

- Digital Preservation Testbed and Evaluation Framework
  - Building on Phase 1 tasks (testbed framework & the documentation of functionality and behaviour)
  - Tools to support the evaluation of files at ingest
  - Case studies with Phonogrammarchiv, Austrian National Library, Goettingen State and University Library, and the Nationaal Archief

Phase 2 tasks, 2005-2006 (2)

- Automated Ingest and Appraisal Metadata
  - Extraction of both technical and semantic metadata
  - Evaluation of existing metadata capture tools
  - Experiments with limited classes of document types (e.g., PDF files)
  - Integrated with Digital Curation Centre research agenda (a small example of research integration)

Digital Curation Centre (DCC)

- Basics (to recap from yesterday):
  - DCC is UK initiative funded by the Joint Information Systems Committee and the e-Science Core Programme of the UK Research Councils
  - Combines cultural heritage and e-science perspectives
  - Has a major research component, led by Professor Peter Buneman, School of Informatics, University of Edinburgh

DCC research goals

- To draw together the various functions of curation, from the traditional archival functions to the maintenance and publication of evolving knowledge as seen in scientific databases
- To conduct research in areas already identified by the partners as crucial to digital curation
- To identify through direct research collaboration, and through interaction with the service arm of DCC, the key projects in which research is needed
- To institute two-way conduits between research and service in which practical issues can be drawn to the attention of researchers and the products of research can be tested in practice

DCC research agenda (1)

- Data integration and publication
  - Review of techniques
  - Publishing data that conforms to a given format or schema
- Performance and optimisation
  - Safe data analysis environments within data centres
    - Initial testbed based on sky survey databases (in collaboration with the Wide Field Astronomy Unit and AstroGrid)

DCC research agenda (2)

- Performance and optimisation (continued)
  - Automated metadata extraction and generation
    - Essential for testing the scalability of metadata-based preservation strategies
    - Review of tools, assessment of text mining techniques
- Annotation
  - Survey of the forms of annotation
  - Annotation and provenance
    - A model for data transformations that maintains annotation and provenance

DCC research agenda (3)

- Appraisal and long-term preservation
  - Appraisal techniques
    - Investigating the applicability and scalability of traditional appraisal techniques in 'data-intensive' contexts
    - Dynamic databases
    - Preservation techniques for evolving metadata and databases

DCC research agenda (4)

- Socio-economic and legal contexts
  - Networks of trusted repositories
    - Varying preservation role for repositories
    - Roles for co-operation, exchange formats, replication, etc.
  - Economic cost-benefit analysis of curation processes
    - Quantifying costs and benefits
    - Testing economic viability of curation processes

DCC research agenda (5)

- Socio-economic and legal contexts (continued)
  - Rights and responsibilities
    - The legal contexts of curation, e.g. impacts of the EU Database Directive
    - Complexity of rights held in databases, impacts on aggregation and reuse of data

Other UK research (1)

- JISC research programmes:
  - Supporting Digital Preservation and Asset Management in Institutions
    - Relatively small-scale projects: assessment tools, training, user guides, etc.
  - Digital Repositories
    - Building on earlier Focus on Access to Institutional Resources (FAIR) programme
    - >20 projects
  - http://www.jisc.ac.uk/

Other UK research (2)

- Digital Preservation Coalition:
  - UK Digital Preservation Needs Assessment
  - Current state-of-the-art survey
  - Being undertaken by Tessella, due to report in October 2005

Summing up

Some conclusions

- A need for digital preservation research to be:
  - *Incremental* - building on what exists already (e.g. learning from mistakes)
  - *Innovative* - not afraid of new thinking
  - *Implementable* - must feed into development activity, have relevance at policy level
  - *Integrated* - collaborative, combining different viewpoints wherever possible (as in DCC and DELOS)
- Also, *Infrastructures* are important

Further information:

- DELOS Digital Preservation cluster: http://www.dpc.delos.info/
- Digital Curation Centre (DCC): http://www.dcc.ac.uk/
- Digital Preservation Coalition http://www.dpconline.org/
- National Digital Information Infrastructure and Preservation Program (NDIIPP): http://www.digitalpreservation.gov/

Acknowledgements

The Digital Curation Centre is an initiative of the the Joint Information Systems Committee (JISC) of the UK higher and further education funding bodies, and the e-Science Core Programme of the UK research councils. The consortium is led by the University of Edinburgh and also includes the University of Glasgow (HATII), the Council for the Central Laboratory of the Research Councils, and the University of Bath (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/