Exploring and Charting the Digital Preservation Research Landscape

iPRES 2007
11 October 2007
Beijing

Photo taken Dec 03 at Sintra Museum of Modern Art

Prof Seamus Ross,
Professor of Humanities Informatics and Digital Curation and
Director, HATII (University of Glasgow)

Humanities Advanced Technology and Information Institute (HATII)

- HATII
  - http://www.hatii.arts.gla.ac.uk
- Digital Curation Centre (DCC)
  - http://www.dcc.ac.uk
- DELOS
  - http://www.dpc.delos.info
- PLANETS
  - http://www.planets-project.eu/
- CASPAR
  - http://www.casparpreserves.eu/
- DPE
  - http://www.digitalpreservationeurope.eu
- AHDS Performing Arts
  - http://www.ahds.ac.uk/performingarts/index.htm

George Service House, Humanities Advanced Technology and Information Institute (HATII)

Arts and Humanities Data Service (AHDS)

- Founded 1995
- UK organisation to support preservation of the results of digital scholarship.
- Managed by an Executive in London
- Collaboration of 5 centres of expertise: Archaeology, History, Visual Arts, Text, and Performing Arts
- Funded by JISC and AHRC

Library at Hadrian's Villa at Tivoli

Digital Repository Hardware at Swiss Federal Archives

Rescuing AHDS-PA

- The Visit from Scottish Enterprise
- Celebrating the Domesday Book
- Rescuing the BBC Domesday Book
- Rescuing data is not free, but….
  - Maintaining data costs money as well

Objectives of digital longevity

- Digital preservation aims to ensure that future users will be able discover, render, manipulate, interpret and information in the face of constar changing technology
- It involves conservation, renewal, selection, destruction, enhancing, updating, and annotating
- It is a risk management activity at all stages of the longevity pathway -- translating uncertainties into manageable risks
- Digital Preservation is an ongoing activity to ensure securing value of digital objects

Charles Dollar visits HATII, 2004

We need to make a radical change

- We need to stop 'building that agitating buzz'
- We need to deliver actual solutions that work.
- We need solutions that scale.

An Agitating Buzz about Things Digital

- We have successfully socially amplified the perception of risks associated with digital materials
- While there are 'risk amplifiers' and 'risk attenuators', preceptions of risk arise complex social and cultural processes
- Have we socially constructed our preceptions of preservation risk within a vacuum

Preservation Risk is Actual

- It is technological.
- It is social.
- It is organisational.
- And it is cultural.
- Actual risks can be assessed and measured—actual risks can be managed.

Research Roadmap

- Analysing the state of the art in Digital Preservation research and existing research agendas.
- Redefinition and refocusing of the Preservation research agenda
- Basis for development of research problem basis
- Provides a foundation for communication about research needs.

Reviewed Published Research Agendas from 1991 to 2006

- **UEI** – *Understanding Electronic Incunabula: A Framework for Research on Electronic Records* [9] by Margret Hedstrom, 1991.
- **PDI** – *Preserving Digital Information* [17], edited by John Garrett and Donald Waters, 1996.
- **DPNU** – *An Investigation into the Digital Preservation Needs of Universities and Research Funders* [15] by Denise Lievesley and Simon Jones, 1998.
- **SoDP** – *The State of Digital Preservation - An International Perspective* [5] contains articles by various authors, 2002.
- **IAT** – *It's About Time: Research Challenges in Digital Archiving and Long-term Preservation* [3] was published by the NSF in 2003.
- **I2S** – *Invest to Save* [4] was prepared for the NSF-DELOS working group on digital archiving and preservation in 2003.
- **eScience** – *e-Science Curation Report* [16] by Philip Lord and Alison Mcdonald was published in 2003.
- **Cyber** – *Revolutionizing Science and Engineering Through Cyberinfrastructure* [18] was created by the Blue-Ribbon Advisory Panel on Cyberinfrastructure of the NSF in 2003.
- **DigiCult** – *The Future Digital Heritage Space: An Expedition Report* [6] was published as a DigiCULT thematic issue in 2004.
- **Erpanet** – *Electronic Resource Preservation and Access Network*[1] was a European Commission funded project which ran from 2001 until 2004.
- **Warwick** – *Digital Curation and Preservation: Defining the research agenda for the next decade* [1] reports on the Warwick workshop held in 2005.
- **DRR** – *Digital Repositories Roadmap - Looking Forward* [29] by Rachel Heery and Andy Powell, 2006.

Review Conducted Segmented by

- Digital Object Level
- Collection Level
- Repository Level
- Process Level
- Organisational Environment

The Landscape

- lack of common understanding
- loss of focus
- lack of practical experience
- fragmentation
- frictional losses
- lack of training

Refocused Landscape

- Restoration
- Conservation
- Management
- Risk
- Significant Properties of Digital Objects
- Interoperability
- Automation
- Context
- Storage
- Experimentation

Out of Scope, but not out of mind

- Research also needed in areas of
  - policy and procedures,
  - organisational structure and communication,
  - education,
  - business case development, or
  - legal arena.

Preservation of Systems and Technology

- Preservation of Systems & Technology
- Managing Complex and Dynamic Digital Entities
- Automated Metadata Creation
- Long-term Metadata Viability
- Multilingual Entities and Technology
- Acceptable Loss
  - Authenticity
  - Renderability
- Repurposing

Digital Repository Infrastructure, Swiss Federal Archives, Berne, October 2004

Process Planning

- Different formats require different kinds of strategic approaches to ensure that they can be accessed in the future.
- Problems with formats are exacerbated by the fact that archival collections, which need to be managed as a whole, generally contain entities in multiple formats; these formats have different rates of obsolescence.
- E.g. we need predictive measures to enable developers to assess the preservation impact of attributes of formats in advance of their completed development or use.

Automated Metadata Creation

- Preservation metadata is an essential part of the information infrastructure necessary to support all the processes in digital preservation.
- automatic or semi-automated creation and authoring of the technical, descriptive, structural, and contextual metadata are a crucial issue.
- Need for creation of metadata supporting the discover, use and understandability of digital objects.

Automation (or semi-automation)

- Huge quantities of materials to ingest and manage – human effort does not scale
  - selection, validation, description, assigning unique persistent identifiers, data management, migration, and selection and appraisal
- Automation of workflows allow integration of independent services
- Standardized logging/record creation
- Reduce human intervention
  - Cheaper and faster
  - Less error prone
  - Enables higher level of security and reliability
- Enables intensive test and verification mechanisms
- Automated Metadata Extraction

Hans Hofman (Dutch National Archives) and Charles Dollar at ICA2004 Wien.

binary code translation technologies

- automatically translating a binary executable program from one machine (M1) running a particular operating system (OS1) and using a particular file format or data representation (R1) (i.e. platform (M1,OS1, R1) to
- another platform (M2) running a different operation system (OS2) and using a different file format or data representation (R2) (i.e. platform (M2,OS2, R2).

Self-describing & monitoring objects

- Digital objects that know what they are
- Digital objects that know something about their semantics
- Digital objects that can observe the state of other objects (e.g. observe decline in numbers of similar classes of objects)
- Digital objects that know where they are
- Digital objects that know where their metadata are
- Digital objects that can notify their originator/manager if they need to be protected, migrated, secured

Safe-Harbour Seeking Objects

- Embed Trust mechanims in the objects themselves
- Make objects active
- Exploit grid and peer-to-peer technologies
- Exploit existing know-how and reasoning
- Enable objects to look after themselves

Experimental Testbeds

- integrate, automate, and evaluate a framework for digital entity preservation by integrating and combining the testbed framework and evaluation metrics
- tools to automate selected steps of the preservation process, such as ingest validation, preservation experiment set-up and control, preservation criteria definition, and verification of formal transformation, to support semi-automatic alternative evaluation.
- to investigate the potential metrics for measuring the effectiveness of different preservation strategies in the context of complex digital objects
- integration of software tools to support the digital preservation testbed framework.

Thank You

- Contact:
  - s.ross@hatii.arts.gla.ac.uk