Digital Preservation Research: A review of the challenges

Kevin Ashley
Head Of Digital Archives
K.Ashley@ulcc.ac.uk

Background

- Some past documents defining research areas:
  - “Invest To Save” - DELOS/NSF 2002/2003 (I2S)
  - “It’s About Time” - NSF/Library Of Congress 2002 (IAT)
  - Liz Lyon - “Dealing with Data” (2007)
  - Warwick statement (2005) and others on European agenda
- Common themes - some common authors

Invest to Save

- Small team of authors
- 21 (25) research areas; 7 supporting issues (legal, organisational, etc.)
- Explicitly trans-continental

It's About Time

- 51-member workshop
- Small editorial team to draft outputs
- 64 research issues
- USA participants and funders only

What's long-term preservation?

"A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing user community, on the information held in the repository. This period extends into the indefinite future."

OAIS definition.

It's not always forever!

Where is the research now

- Some has been done
- Some is being done
- Some has yet to be started
- There may also be new challenges not yet identified

Work that is done

- Salvage and rescue (digital material) (I2S)
  - Perhaps we can learn to make it easier or cheaper
- Repository models (Both)
  - But Dspace etc. lack preservation architecture
- Format repositories (Both)
- Distributed storage (Both)
  - but what about torrent-type models
- Understanding media (Both)

Work that is being done

- cost modelling and process modelling (Both)
- collection completeness (I2S); anomaly detection (Both)
- Complex entities - web archiving (Both)
- Scalability (Both)
- Preservable metadata: PREMIS is a big step (Both)
- Audio and video preservation (I2S)

Work being done 2

- Certification + trustworthiness: (Both)
  - Of repositories and content
- Effective refreshing of media (IAT)
- Automated acquisition and description (Both)
- Representation Information Registries (I2S)

Work being done 3

- Automated policy application
- Workflow preservation and sharing
- Distributed content (e.g. comments/data)
- Disciplines or places ?
- Generic database preservation

Work that remains

- Newer formats: virtual worlds, musical scores, XML (I2S)
- Accelerated ageing of systems + software (I2S)
- Software repositories: (Both)
  - Architectures
  - Classification schemes and content models

Work that remains 2

- Multilingual entities (I2S)
  - preservation/migration, not creation
- Anomaly detection:
  - At ingest; migration; of a collection (IAT)
- Automated provenance generation (IAT)
- Migration of authentication information (IAT)
- Defining designated communities

Work that remains 3

- Self-aware objects (I2S)
- Scaling down (I2S)
- Market analysis: customer base and needs (IAT)
- Formal models for selection (Both)
- Exchange of content and services between repositories (Both)
- Migrating ontologies, schemas, etc (IAT)

What type of research?

- I2S requires pragmatic and theoretical research

"Practice informs research in a fashion similar to the ways in which research informs practice."

- Need good pathways in both directions

Final observations

- General problem: tension between big abstracts and specifics.
  - Sometimes research is too specific
  - Some of the problems are defined too generally
- Good research can be done by those outside the field
- Those within it must define the problems better