DPC Special Interest Group on WEB ARCHIVING

Deborah Woodyard
Digital Preservation Coordinator, BL
DPC SIG WA Chair

DPC Forum, London
12 March 2003

OVERVIEW

- About the SIG on Web Archiving
- First meeting topics
  - PANDORA / PANDAS
  - Internet Archive
  - European Web Archive proposal
  - OCLC archive
  - BBC / PRESTO experience
  - Edinburgh University Library
  - Other work – NLW, PRO
- Future plans

THE WEB ARCHIVING SIG

- Open to all members of the Digital Preservation Coalition who wish to participate.

- The Special Interest Group will report on its activities to the membership via the chair and to the board via regular reports to the Secretary and Coordinator.

SIG TERMS OF REFERENCE

Mission

To promote practical web-archiving activity among members.

To foster collaboration and co-ordination of web-archiving activity within the UK and where relevant internationally.

SIG AIMS

- advise on and promote collaborative web-archiving projects involving DPC members
- provide a forum for discussion and agreement on co-ordination of members interests, responsibilities, and activities in web-archiving
- share expertise, research and experience in web-archiving between the membership
- advise the DPC board on international activity and DPC participation in this (eg web-archiving activity within the EU Framework Programme 6).
- coordinating DPC member collaboration to achieve national strategies for web-archiving in the UK

PANDORA /PANDAS

- PANDORA = Preserving and Accessing Networked Documentary Resources of Australia

  http://pandora.nla.gov.au/

- PANDAS = PANDORA Digital Archiving System

PANDORA /PANDAS

- Selective, quality capture
- Low cost
- Supports distributed archive capture
- In use
- Provides immediate access to archive
- Weakness: resource intensive,
  therefore ~600 new sites per FTE per year

INTERNET ARCHIVE CONSORTIUM PROPOSAL

- Involves countries (National Libraries) such as : Iceland, Finland, Switzerland, USA, France, Canada, UK, and more
- To create new crawler based on library specifications
- IA collecting roughly 10 Tb per month
- $2.5 million for the whole project
- Weaknesses: Contractual and Legal issues

EUROPEAN WEB ARCHIVE PROPOSAL

- European 6th Framework expression of interest planning to combine with another proposal to form an Integrated Project proposal
- EOI involved many European national libraries, university libraries and IT companies, including DPC
- Proposed merger with DLM - CUBE (Cultural heritage UmBrella for REsearch in the Archiving Domain)

OCLC ARCHIVING

- Archive came into production in May 2002
- Limited file types
- Including bit stream preservation
- Selective collection
- ~$12k per year for the harvester + storage costs

BBC & PRESTO

- PRESTO : ‘Preservation Technology’
  Survey of 10 public broadcasting archives
  2/3 objects cannot be easily used : obsolete
  1/3 objects showing deterioration

- Project successful because it was run by the users

“Effective preservation needs a process, not just good individual items of technology”

Preservation alone will not inspire funding, need to couple with access etc.

Edinburgh University Library

- Have just purchased Endeavour Encompass Library System
- Aiming at selective archiving
- Proposed project “Archiving the Islamic Web”

OTHERS

National Library of Wales

- Positioned to start a selective 100 site pilot

Public Records Office

- Test bed focussing on intranets

FUTURE PLANS

- Low interest in European Web Archive proposal
- DPC and Wellcome to continue negotiations for PANDAS software
- High interest in PANDAS from many participants
- June Forum – sample sessions for PANDAS and DSpace