UKWAC - the first two years

Philip Beresford
Web Archiving Project Manager
British Library

DPC Forum on Web Archiving, 12th June 2006
UKWAC’s origins

- JISC / Wellcome joint report 2002
- UKWAC pilot - for three years
  - agreement signed October 2003
- Consortium members:
  British Library,
  JISC,
  The National Archives,
  National Library of Scotland,
  National Library of Wales,
  Wellcome Trust
UKWAC’s Aims & Objectives

- **Aim:** To develop collaborative approaches to web archiving within the UK on a shared infrastructure

- **Objectives:**
  - To award a contract to provide the common infrastructure for the pilot project.
  - To test and develop PANDAS (Pandora Digital Archiving System) software from the National Library of Australia.
  - To evaluate the development of the collaborative infrastructure for web archiving.
  - To work collaboratively in: selection; obtaining permissions; collection management and other curatorial issues and standards; digital preservation.
  - To provide access to web sites collected as part of the project.
Key dates

- Oct 2003 UKWAC consortium agreement signed
- May 2004 Magus contracted
- Sept 2004 Pandas installation & trials completed
- May 2005 UKWAC archive goes live
- Apr 2006 Consortium Agreement extended to Sept 2007
- June 2006 UKWAC Evaluation Report completed
- Sept - Dec 2006 Alternative software evaluations
- ? Sept 2007 migration onto new infrastructure
Deliverables

UKWAC has delivered:

- A set of collection development policies
- A shared permission request/licensing process & FAQs
- Shared infrastructure & support from Magus Research
- An archive of nearly 1400 websites (4000 instances) freely accessible at [http://www.webarchive.org.uk/](http://www.webarchive.org.uk/)
- Special ‘collections’ e.g. Tsunami, Election, Women’s
- A regular listing of archived sites no longer available on the live web
Constraints

- Need to seek (and obtain) permission to archive
  - low response rate
- Resource-intensive operation (to gather, process and QA)
- PANDAS 2 software problems and limitations
- Metadata limitations
- Technology dependent
  - archivists need technical skills
Common archiving problems

- Difficulty of setting filters
  - to gather only those parts of a site for which we have permission

- Impossible to gather some content types:
  - e.g. Flash, Javascript, streaming media

- Limits to number of links / file sizes

- Problems in following robots.txt directions

- Problems handling links in CSS files

- Problems with HTTrack log files
Workarounds

- Copy log files into Excel in order to sort and analyse error reports
- Use HTTrack v3.33 outside of Pandas
  - avoids Pandas limitations and HTTrack bugs
- Obtain published content on CD and upload into Pandas archive
- Edit archive files to correct links
- Manually fetch and restore missing content files
Issues for UKWAC

- Scaling up a cottage craft to an industry for legal deposit
- Integrating selective and large-scale web archiving
- Still immature technology
- Need a storage solution
- Active preservation processes required for long-term
- Human resourcing
- Metadata
New Infrastructure Options

- **Web Curator Tool:**
  - IIPC project led by NLNZ and BL
  - Scope of Release 1 now agreed
    - Will include multi-agency capability
    - Will exclude access
  - Due to be completed (for testing) summer 2006

- **Pandas 3:**
  - Re-write
  - In internal testing now
  - Release expected (for trials) also summer 2006
Next Steps for UKWAC

- Migrate to new toolset
- Increase throughput
- Determine its role beyond legal deposit
- Further develop collaborative approaches to selective web archiving
Contact:

philip.beresford@bl.uk