CREATING WEB ARCHIVING SERVICES IN THE BRITISH LIBRARY

John Tuck
Head of British Collections, British Library

Web Archiving Services Panel
DLF Fall Forum 2005
November 9th 2005

Creating Web Archiving Services in the British Library

The presentation will cover the following areas:

- Brief history of web archiving in the British Library
- Relationship with UK Web Archiving Consortium and International Internet Preservation Consortium
- Legal Deposit in UK and impact on non-print: latest developments
- Challenges: legal, technical, operational

British Library Web Archiving Programme:
Where are we now?

- Web archiving began in BL in 2001/2002 with experimental domain.uk project

- This led to creation of BL Web Archiving Programme which comprises three main elements: Underpinning collection development policy, UK Web Archiving Consortium (UKWAC), and International Internet Preservation Consortium (IIPC)

British Library Web Archiving Team

- Seven full time staff dedicated to web archiving
- Web Archiving Programme manager
- Curator, Web Archiving
- Each of the above has a team

UKWAC

- UKWAC comprises six institutions: British Library, National Library of Scotland, National Library of Wales, The National Archives, JISC, and Wellcome Trust

- Initial two-year project to develop and evaluate a collaborative infrastructure for web archiving within the UK consisting of both shared costs and shared objectives, and exploring the nature of collaboration between partners

Issues in Providing a Service:
Legal

- UKWAC operates on website owner rights clearance basis
- Less than 25% successful return rate on permissions; although few outright refusals
- Considering ‘notice and takedown’ approach which has been tested in connection with events-based collecting, e.g. General Election 2005; 7/7 London bombings

UKWAC: Achievements / Challenges

- Fulfilled goals of testing and using National Library of Australia's PANDAS software
- Live archive of web sites, fully accessible, free of charge is in place http://www.webarchive.org.uk/
- Permissions-based approach has worked but has raised challenges
- Key issues in terms of software and its functionality especially in a collaborative environment e.g. green-light red-light approach, downtime etc
- Challenges in formulating agreed collecting procedures
- Challenges in different remits of the member institutions

Issues in Providing a Service: Legal:
Legal Deposit

- Legal Deposit Libraries Act 2003 and extension of legal deposit to non-print
- Enabling legislation now moving to secondary legislation stage through work of the Legal Deposit Advisory Panel (LDAP)
- Triads set up to look at specific areas, e.g. e-journals, websites. Great interest expressed in work of UKWAC

Issues in Managing a Service:
Technical

There was full agreement by UKWAC to use PANDAS but we have encountered a number of problems, including:

- Robot exclusion
- Content filtering
- Maximum URL limit
- Style sheets
- JavaScript
- Failed downloads
- Missing gathers

UKWAC: Unexpected Issues

- National library archiving websites implies validation of that site
- Imbalance of collection, e.g. if one political party agrees to have site archived, others not
- Website owners fear confusion over archived site and current site
- Bringing down sites.

IIPC: Whole domain searching and tools

- Approximately 5 million UK websites
- Stated aim of seeking a snapshot on a regular basis, i.e. in line with legal deposit
- Procurement in place (British Library and Bibliotheque nationale de France) for smart crawler (prototype planned in 2006/2007; possibility of whole domain searching in 2008)
- Issues of territoriality; resourcing; access in respect of Legal Deposit Libraries Act

Addressing the Curator Interface Issues

Options appraisal planned:

- PANDAS 3
- IIPC tools tested within BL and appraised by UKWAC partners
- IIPC Curator Tool: project under auspices of IIPC (National Library of New Zealand, Library of Congress, British Library)

Next steps

- Defining the future of UKWAC – current pilot ends in June 2006
- Future of IIPC
- Need to work with and learn from others to achieve web archiving!

THANK YOU

…. and concluding with a recent quotation from an RLG report ‘But web archiving tools and services require a level of local commitment and resources that often exceed in-house capacity’