Web Archiving Services in The British Library: An Update

Web Archiving Panel, DLF Fall Forum 2006

November 8th 2006, Boston

John Tuck

Head of British Collections, The British Library

Web Archiving Services in the British Library: an Update

The presentation will cover the following areas:

- Legal deposit and progress relating to UK websites
- UK Web Archiving Consortium developments
- Thematic collections
- BL and International Internet Preservation Consortium developments

Legal Deposit: a Reminder

- Legal Deposit Libraries Act 2003 and extension of legal deposit to non-print
- Enabling legislation now moving to secondary legislation stage through the work of the Legal Deposit Advisory Panel
- Triads set up to look at specific areas, e.g. e-journals, websites etc.
- Work being carried out on mapping the universe of electronic publications potentially eligible for legal deposit

Legal Deposit and Websites: Progress

- Legal Deposit Advisory Panel being reorganised into three new groupings: offline; e-journals; and websites.
- LDAP websites group has worked closely with UKWAC (UK Web Archiving Consortium) and is proposing working towards a Regulation for public websites
- Web archiving seen as important for LDAP in view of significant barriers in current environment, i.e. voluntary deposit onerous on deposit library and depositor; agreements and copyright clearances required; low permissions leading to imbalanced collections; labour-intensive nature of harvesting; small scale and unsustainable to continue with voluntary deposit

Legal Deposit and Websites: Progress

Proposal being considered by LDAP as follows:

- Regulation to be put in place to enable legal deposit from the public domain web space of the UK through a combination of domain-level harvesting and selective web archiving
- Retention of such content by the Legal Deposit Libraries to be enabled
- Access to the content within the Legal Deposit Libraries infrastructure to be enabled and not hindered by legal barriers such as FOI
- Technical mechanisms to be put in place to make deposit easier and more effective

Aims:

- Regulation to come into force as soon as is practicable
- Impact assessment and consultation to be wide enough to include all relevant publisher groupings

Realistic timescale:

- Not likely to be implemented before 2008

UKWAC: a Reminder

- UKWAC comprises six institutions: British Library, National Library of Scotland, National Library of Wales, The National Archives, JISC, and the Wellcome Trust.
- Initial two-year project to develop and evaluate a collaborative infrastructure for web archiving in the UK
- Has achieved its objectives and has agreed extension until at least September 2007
- UKWAC operates on website owner rights clearance basis: less than 25% successful return rate on permissions; although few outright refusals
- Live archive of web sites, fully accessible, free of charge is in place http://www.webarchive.org.uk. Over 1675 sites available

UKWAC: Collecting

Significant thematic collections strand: seven examples:

- General election UK 2005 (93 sites)
- Personal experiences of illness (27 sites)
- Tsunami disaster 2004 (23 sites)
- Avian and pandemic influenza (11 sites)
- Digital lives (24 sites)
- Terrorist attacks, London, July 7th 2005 (49 sites)
- Women’s issues (117 sites)

More underway, e.g. Olympics 2012

UKWAC: Thematic Collections: Women's Websites and UK General Election 2005

Archiving women's websites: collaborative project between the British Library and the Women's Library:

- Aim was to create a new resource for future researchers and to ensure that valuable information currently on the web about women is not lost
- Project started in autumn 2005 with objective of archiving 100 sites
- Joint identification of categories for selection including sites with research content; women's organisations and campaigns; women's networks; personal sites of women including blogs and samples of women's e-zines
- Sites are archived every 6 months

UKWAC: Thematic Collections: Women’s Websites and UK General Election 2005 contd

UK General Election 2005:

- Involved British Library, National Library of Scotland and National Library of Wales
- Particular challenges were: tight timescale as many sites only appeared in second or third week of campaign; in case of BL about a third of the sample of 300 selected sites gave permission; more than half in case of NLS.
- Technical challenges; measures had to be put in place to avoid crashing of the PANDAS software, i.e. limits on working drive space; limits on processing per partner; stalled gathers etc.

International Developments with IIPC: Web Curator Tool

The IIPC Web Curator Tool (WCT): an open source solution for selective web harvesting

- Joint project of National Library of New Zealand and the British Library
- WCT is a tool for managing the selective web harvesting process
- Designed for use in libraries by non-technical users
- Heritrix is used to download web material
- WCT supports: harvest authorisation (getting permission); selection, scoping and scheduling; description; harvesting; quality review; and submitting the harvest results to a digital archive
- WCT is not an access tool
- WCT has been tested and released (http://webcurator.solurceforge.net)
- BL is testing WCT in own environment together with linked access module as part of UKWAC infrastructure evaluation and development

International Developments with IIPC: Automated Smart Crawler

Automated Smart Crawler

- Joint project involving British Library (project lead), Library of Congress, Bibliotheque nationale de France and Internet Archive

Scope includes developing Heritrix crawler so that it has ability to:

- recognize when resources have not changed from last time they were gathered
- Prioritise the order in which resources are visited and gathered
- Recognize when resources are changing more frequently and to visit them more frequently
- Ensure that new capabilities will scale to crawls of at least 100 million resources
- Ensure that web archiving community can take advantage of these new capabilities

Working with Publishers

Automated Content Access Protocol (ACAP)

Led by Rightscom, aim of this pilot project, working with publisher bodies, is to:

- Develop a specification which will allow the publisher of a website or any piece of content to attach extra data, in a standardised form, to specify what uses of that piece of content or of the website are permissible
- British Library intends to participate in the pilot project in view of potential benefits in connection with legal deposit legislation

Tricky Web Content

Meanwhile, still some tricky technical issues:

- Database content accessed via web forms, e.g. Collect Britain search forms
- Interactive web applications, e.g. Electoral Calculus (`Make your prediction’ poll predictor)
- Streaming multimedia files, e.g Respect Coalition
- Macromedia flash files that are linked from other Flash files e.g. J.K.Rowling
- URLs controlled by JavaScript e.g. Elcano Royal Institute