Preserving the content and the network: An innovative approach to web archiving

Amanda Spencer, Web Continuity Manager, The National Archives, Sept 2008
Introduction

• Since the early 1990s the Government has been using websites to present information: official reports, papers, transcripts of speeches, guidance, announcements, press statements, regulations and advice.
• Benefits offered by new technologies means that services are increasingly delivered online.
• Majority of all interaction between government and the public now happens online.
• Web links cited in everything from Hansard through to academic research and PR campaigns
• Government web estate, just like the rest of the web, vulnerable to technological problems
• Broken Web links are increasingly common – and frustrating for us all
Web continuity matters

• Integrity of Web links crucial to the business of government. Without it:
  • Public – impaired access
  • Westminster village – impaired access
  • The Press – impaired access
  • Academics – impaired analysis
  • Parliament – impaired scrutiny
  • Impacts on reputation of government – public and abroad
Government is taking action

- April 2007 - Raised as a serious issue by Jack Straw, as Leader of the House of Commons
- May 2007 – Hilary Armstrong commits government to a package of measures that will provide a long-lasting solution to the problem
- Working Group established in May 2007 to investigate the problem and come up with a solution
What the working group found…

• A longitudinal survey of URLs cited in response to Parliamentary Questions and recorded in Hansard revealed that 60% of URLs, cited in Hansard between 1997-2006 are broken

• Disconnect between those who view government websites as ephemeral and those who expected important information to be available in perpetuity

• Prevalence of e-only publication; issues with legal deposit in the electronic world

• Varied practice - PDF vs HTML

• Devolved system of web publication

• Concerns also about the impact of ‘Website Rationalisation’
Agreement

• Working group agreed that all web-based information should be treated as an important contribution to the body of government information

• Online information which has been cited elsewhere should remain available and accessible in its original form
Options

- Improvements to existing practices
- Encourage government departments to take more responsibility
- Use Digital Object Identifiers
- All of the above?
The National Archives and the Web Continuity solution – issues to address

1. How to capture significant levels of important Government information from possibly thousands of distributed, heterogeneous websites (including websites closing as part of the Website Rationalisation Programme);

2. How to ensure not only a greater capture of content, but also increase exposure of this content to the web harvesting crawler, from sites that vary hugely in nature;

3. How to ensure that links persist so that users will always find the last available version of the page, whether it is on a live site, or in the web archive.
Web Continuity – capture I

- Comprehensive - Whole of UK central government
- Incorporating Transformational Government Website Rationalisation Programme
- SQL Server Government Website Database developed as a single source of information for central government websites
  - allows tight control over web estate
  - gives responsibility to government
Web Continuity – capture II

- Automated seeding of crawls and capture of preservation copies via two workflows:
  - Harvesting workflow, crawls are seeded, and progress of harvesting and QA processes are recorded in the database through the exchange of a series of XML messages via FTP between TNA and the European Archive, who carry out web archiving under contract
  - Preservation workflow, enables ARC files to be ingested into The National Archives Digital Object Store, once harvesting process is complete
Web Continuity – capture II

- Promoting use of XML sitemaps
- Benefits live website users as well as helping the archiving process; ministerial backing
- Training in partnership with Third Party provider
- Guidance provided
- Installation of sitemap generation software on government web servers – requires support from senior government
- Practical implementation – will be used to augment the initial gather rather than to drive the crawl
Web Continuity – links persistence

- Configuration of open source software
- Apache and Microsoft IIS Web Servers
- Enables users to be redirected to web archive instead of receiving a 404 Page Not Found error; does not replace existing redirects on live websites
- Will need to be installed on government web servers
Web Continuity – links persistence - implications

- Brings together different audiences
- Archived websites have: historical value, current value, value in preserving integrity of network as a whole
- Introduces temporal dimension – brings longevity to the web, but also needs to be clearly signposted – new presentation and branding, including a TNA URL
Web Continuity – Presentation issues

• Several attempts at producing a banner:

  Using Iframes:

  Using CSS overlay:
Practical Implementation

- Implementation across government has required a new approach to sharing information and different technical support model
- Collaborative platform established
- Guidance for web and e-comms communities – part of Transformational Government Web standards
- Regular monitoring
- Different groups of stakeholders involved at The National Archives
Timelines

- November 2007 – April 2008 Feasibility studies
- May – November 2008 Development work at The National Archives, website archiving programme scaled up
- November 2008 Solution delivered to government, software and guidance available, monitoring in place
Any Questions?