Archiving the UK Web

Learning on the Job

Helen Hockx-Yu
Head of Web Archiving
The British Library

July 2011
The UK context

- 9 million .uk domain names registered with Nominet in December 2011
- ~ another 1 million using other domain names, e.g. .com
- Growing between 12%-17% per year

- 2 Web Archives in the UK – covering ~1% of the UK domain
  - UK Web Archive provided by the British Library
  - UK Government Web Archive provided by the National Archives

- Legislation for non-print Legal Deposit not yet in place
  - Currently BL has to obtain permissions prior to archiving
  - LD Expected April 2012
  - Access expected to be restricted to Legal Deposit Libraries’ reading rooms only – 6 LDLs in the UK
10 years of web archiving – a learning process

**Explore**
- Domain.UK project
- No public access

2001-2002

**Collaborate**
- Establish Web Archiving Programme
- Define collection policy
- Led UK Web Archiving Consortium
- Hands-on selective archiving

2003-2007

**Build capacity**
- Moving operations in-house
- Data and system migration
- Tools development
- Access
- Ingest
- Integrated search
- Domain crawl preparation

2008-2011

**BAU**
- Programme disbanded – business as usual
- Implementing non-print Legal Deposit in 2012
- Domain harvesting
- Value-added selective archiving
- Supporting other business areas

2011
The UK Web Archive

- 10,027 websites, 37,879 instances by June 2011
- Streaming media solution (eg One and Other)
- Earliest archived website instance in the world: British Library website archived on 18 April 1995
- Comprehensive search and browse functions including full-text search
  - full-text search at archive, collection and website levels
  - browse by subject, special collection, title
  - Alternative access methods, eg tags clouds, 3D wall and N-gram search
- Increasing usage of the archive
  - 167,726 visits and 683,966 page views between 1\textsuperscript{st} April 2010 - 31\textsuperscript{st} March 2011 (~50% increase)
And much more

- Capability and know-how
- Unified file format (including migration from legacy format)
- Tools development
  - Web Curator Tool (WCT)
  - Selection and Permission Tool (SPT)
- Integration with Primo for resource discovery
- Ingest of web archive into Digital Library System for long term preservation
- Preparations for non-print Legal Deposit
  - Domain crawl
  - Deposited websites
- Develop access and understand researcher needs
  - UK Web Archive and Researchers Project
Building web archiving capability

- Capability is people, systems and processes
- Hire good people with wider set of skills (eg industrial experience)
- Make sure good people remain good – training
  - Linux, Java, Hadoop, SOLR
  - Visits to other web archiving teams
- Bridging the “digital divide“
- Know how things work - in possession of core expertise
- Workflows, policies & guidelines, roles and responsibilities
- Not something one-off, need to respond to the evolving web continuously

In 2020, the British Library will be a leading hub in the global information network, advancing knowledge through our collections, expertise and partnerships, for the benefit of the economy and society and the enrichment of cultural life.
More than crawling the web…
Placing web archive in the Library context
Remember who we are doing this for

- (Academic) researchers are the (targeted or perceived) main user group of the UK Web Archive
- Researchers and the UK Web Archive Project
  - 8 selected researchers take part in the project, all from UK HEIs
- Objectives
  - Engage researchers in use of the UK Web Archive as a resource for research
  - Develop the UK web archive in line with the requirements of its key user group
  - Develop special collections in collaboration with the researchers
  - Apply data analysis tools to support research
  - Disseminate findings of the project nationally and internationally
  - Work with JISC and IA on the next phase, involving wider range of disciplines, eg web science, digital humanities
Web archives as historical documents for reference
Showing the value of haystacks

- Web archives as datasets for analytics
- Focus on entirety of web archive collection (or subset) instead of individual websites
- Unlock embedded patterns, trends, context and relationships
- Use technology developed for the live web, eg visualisation
- Develop machine to machine access
Content visualisation

UK Web Archive Subject Hierarchy Visualisation
N-Gram search
Postcode-based access
Vision – UK Web Archiving in 2015

- The British Library will be the primary collector and provider of a web archive which is representative of the UK domain.
- The Web Archive will be used for scholarly research in a range of disciplines.
- Researchers will be able to search and use the web archive as a part of the Library’s overall digital collections.
- The UK Web Archive will be known as the place where researchers and general public look for inactive and or historical UK websites.
- The Web Archive will be stored in the Digital Library System and preserved into the future.
List of URLs

- The UK Web Archive:  
  http://www.webarchive.org.uk/ukwa/

- UK Web Archive data-based visual tools data

- Content Visualisation  
  http://www.webarchive.org.uk/ukwa2/seadragon.html

- N-gram search  
  http://www.webarchive.org.uk/ukwa/ngram/

- Demo Post-code based access (Google Maps replaced with OpenStreet Map)  
  http://www.webarchive.org.uk/ukwa2/map.html