Archiving the UK Web - What I Wish I Knew Before I Started

Helen Hockx-Yu
Head of Web Archiving
British Library

January 2015
Web Archiving at the British Library

- Collect UK digital heritage and provide continued access to archived web resources
- Started web archiving in 2003
- Selective, topical collections and key sites
- Based on websites owners’ permissions
- Open UK Web Archive now contains 68,000 point-in-time snapshots of over 15,000 websites, 21.6TB
- Archiving UK Web for non-print Legal Deposit since April 2013: Legal Deposit UK Web Archive
- Comprehensive national archive with on-site access only
- Joint responsibility of six Legal Deposit Libraries (LDLs)
The UK Web Domain

- 4th TLD after .com, .de and .net
- Over 10 million .uk registered domain
- UK organisations also use non .uk domain names (eg .com or .org) – scale unknown
- Non-print Legal Deposit applies to the open (freely available) web: .uk and other UK-published web resources
“Hockx-Yu is now on the front line of the most ambitious expansion of the British Library’s archiving capability in more than 300 years. At the stroke of midnight on April 5, 2013, legislation known as the Legal Deposit regulations came into force, charging the library with capturing the contents of the entire U.K. web domain—every site carrying the .uk suffix—preserving the material and making it publicly accessible.”
Collecting strategy for websites

Domain Crawl

Events
- Domain crawl:
  - Broad sweep of UK domain
  - Once or twice a year

Key sites
- Events & key sites and news:
  - Events of UK interest
  - High value, high impact sites
  - National & regional news

News
- Special Collection:
  - Focused, thematic collections
  - Support priority subjects
UK Domain Crawl

2013 domain crawl stats
- 3.86 million seeds
- 1.9 billion URLs (web pages, docs, images)
- ~31TB
- Duration: 70days

2014 domain crawl
- 90 million seeds (starting URLs)
- 19th June – 24th December
- ~ 56TB (incl. 4.4GB of viral & 3TB homepage screenshots)
- Over 2 million non .uk domains
Archived website as historical document
Data – information – Knowledge - Wisdom

- Data collection driven by research paradigm
- Contextualise data through analysis
- Discover patterns or recurring behaviour, more analysis
- Interpreting findings, develop/support/refute a theoretical model as explanation
Advance the scholarly use of web archives

http://netpreserve.org/web-archiving/videos
My Job

“...a hybrid of tasks including technical, legal, managerial, communication and curatorial activities.”
My Team

Roles:
Head of Web Archiving
Technical Lead, Crawl Engineer, Service Engineer
Web Archivists (2x)
Engagement and Liaison Manager (IIPC) Programme and Communications Officer

“the British Library’s web archiving operation, [is] rated by experts in the field as among the best in the business.”
- Newsweek

Skills profile: cross-disciplinary expertise
- IT
- Collection management, digital curation
- Management
- Communications
- Web Archiving
A busy start of 2015

Tasks

- Legal Deposit Implementation Group Meeting
- Meeting with Head of IT
- ITIL Continual Service Management training and exam
- Track director for RESAW Conference
- OpenWayback project meeting
- IIPC Technical Training Workshop
- New members of staff starting
- DPC student conference

Emails!
School runs!

Nature of activities

- Management
- High variation
- Action oriented
- Communicative
- Planning / focus on future
- Collaborative

Skills

- Managerial skills
- Communications / intern-personal skills
- Technical understanding of domain / tool
- Ability to create and use (high-level) concept
- Ability to handle complex or uncertain situations
- Education and experience
What I Wish I Knew Before I Started

- The fast evolving web
- Our jobs do not just cover a single field. Require general understanding of multiple areas but still helpful to specialise in one area
- Everyday is a school day! Continuous learning / training / professional development
- Cannot be a perfectionist
- Have to make trade-offs
- Think and plan ahead
- Work life balance