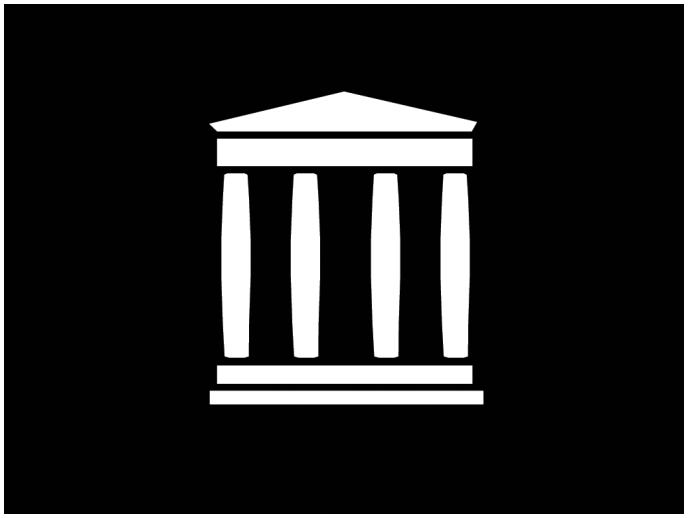

Rethink Web Archiving

Helen Hockx-Yu, Director of Global Web Services
Internet Archive

DPC Students Conference January 2016

About Me

- Digital preservation / Web Archiving
- Project / Programme / Operation/Service management
- IT related
  - 2003-2007: Programme Manager, Digital Preservation and Shared Services, JISC
  - 2007-2008: Planets Project Manager, British Library
  - 2008 – 2015: Web Archiving Programme Manager & Head of Web Archiving, British Library
  - September 2015 – Present: Director of Global Web Services, Internet Archive

20 years of Web Archiving

- Started by the Internet Archive in 1996
- Increased awareness
- Legal issues much better understood
- Growing community
  - 68 initiatives across 33 countries
  - 534 billions of web-archived files since 1996 (17 PB)
- Scholarly use of web archives
- Many challenges

Internet Archive

- A not-for-profit digital library founded in 1996 by Brewster Kahle
- Contains 24+PB of data and is growing
  - Digitised books, manuscripts and other texts
  - Movies & music
  - TV news archive: https://archive.org/details/tv
  - Software
  - Archived webpages
- Over 2 million registered users https://archive.org/about/stats.php

INTERNET ARCHIVE
WayBackMachine

- https://archive.org/web/
- Started web archiving in 1996. Wayback released in 2001
- Largest publicly available web archive in existence
  - 450+ Billion URLs, 100+ million websites
  - content in 40+ Languages
  - 600,000 visit / day
- We collect a broad snapshot of the web every 60 days, +1billion ULRs/week
- Also crawl wikipedia, news, RSS feeds, YouTube etc

Archive-IT

- Subscription service launched in February 2006
- Fully hosted web application for creating, managing accessing and storing web archive collections
- Tools for selection, scoping and capturing web resources at different frequencies
- Support for cataloguing and metadata
- Browse archived content 24 hours after capture is complete; full-text search available within 7 days
- Private access option
- Archive-IT Release 5.0 https://blog.archive.org/2014/10/27/archive-it-crawling-the-web-together/

National Library Service

- Bespoke services for national libraries / archives
- Crawling of national domains or subsets of national domains
- Based on partners' requirements
- Also includes metadata, index, reports and datasets, e.g.
  - WAT files
  - YouTube video report listing all YouTube files captured during each week of crawl
- Developing new aspects

Rethink Web Archiving

- Scope new services
- Consulting stakeholders incl. national libraries, archives, researchers
- Understand requirements
- Propose new service and obtain feedback
- Decide on priorities and plan next steps for development
- Principles:
  - **Collaborative collection development**
  - **distributed preservation;**
  - **Global & local access**

National Libraries

- 246 national libraries
- Public bodies responsible for preserving national heritage
- Legal mandate (including copyright exemption) to collect publications, access to collection often restricted
- Mandate now includes digital publication such as websites
- Archiving process needs to comply with legal requirements
- Tradition of collaborating through various federations, consortia

Key Challenges

- Need to improve quality and comprehensiveness of collections
  - e.g. social media; rich media
  - In-scope content outside national TLD
- Selective and domain harvesting as separate processes
- Effort required to integrate web archives with institutional infrastructure and workflow
- Duplication – web archives contain content streams libraries collect separately
- Access
  - Reading room access only; lack of use
  - Scholarly use not well understood / supported

An End to End Service (1)

- A end to end service offering lifecycle support
- Content collection
  - All crawls in one place
  - Manage crawls regardless of collecting tool; in-house or outsourced
  - pool of tools to choose from
- Processing, integrated with institutional environment and workflow
  - automatically generated catalogue records
  - SIPs for digital storage / preservation system
  - Metadata records for resource discovery
  - Indexes

An End to End Service (2)

- Access – various scenarios depending on institutional requirements
  - Local copy for local access
  - Hosted (private or restricted) access
  - Statistics, analytics, metadata, datasets for researchers
  - Global access via Wayback Machine and archive.org
    - TLD statistics and visualization
    - Search in WBM restricted to TLDs
- Preservation
  - Format analysis
  - Old web, real time emulation
  - Crawl + provenance data

I Wish I Knew

- How hard it is to prioritise
- How things stop working when the scale increases
- How you cannot just do things because they make sense
- How you cannot be a perfectionist
- How important it is to keep learning
- How significant organisational culture is
- How challenging yet how rewarding it is to reach consensus / find a solution