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

Same job, just bigger, more complex

> “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.”

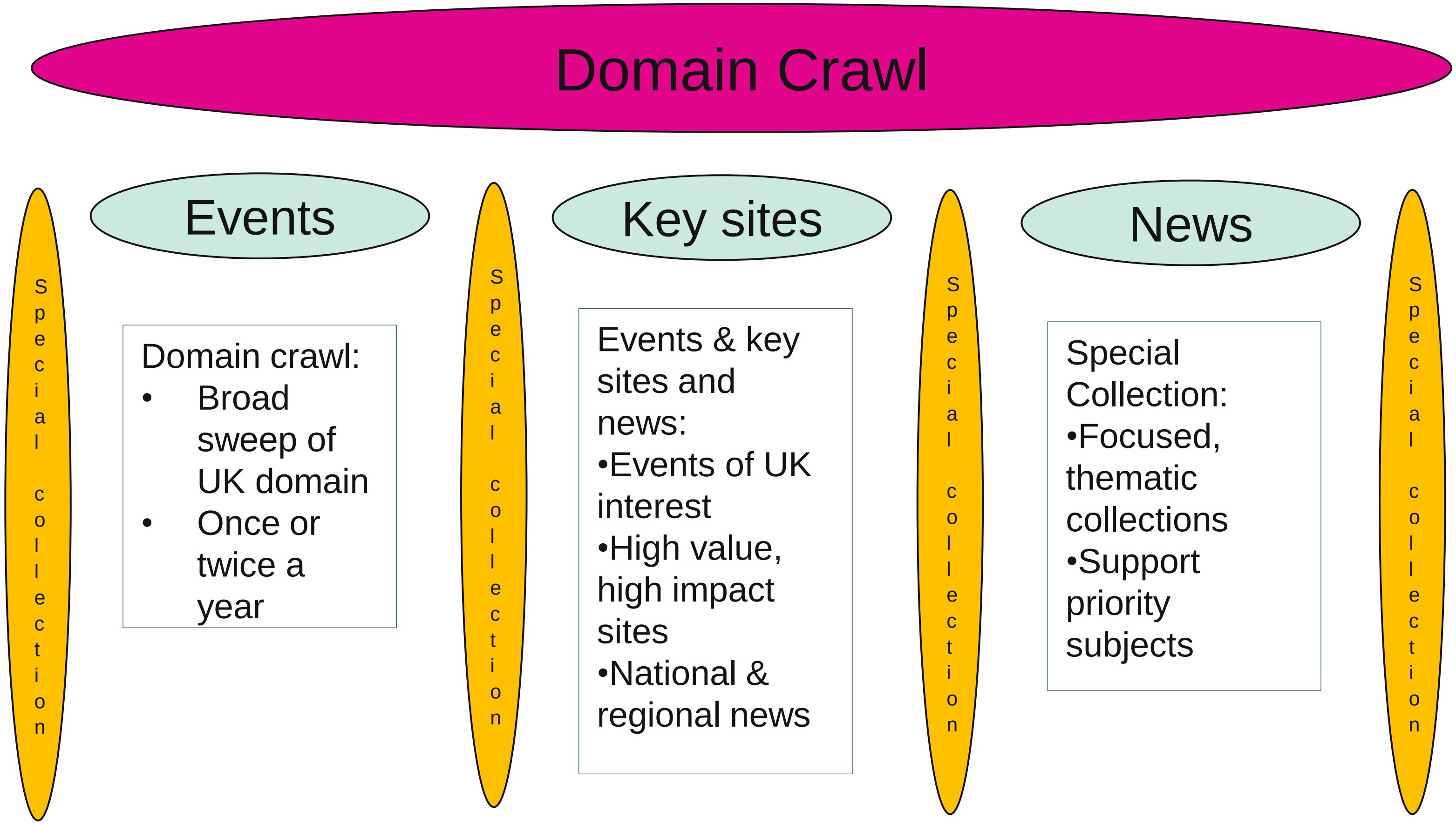
TECH & SCIENCE

Inside the Struggle to Preserve the World's Data

BY PHILIP JACOBSON / JULY 2, 2014 12:21 PM EDT

Newsweek

Collecting strategy for websites



Domain Crawl

Special collection

Events

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

Special collection

Key sites

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

Special collection

News

Special Collection:
•Focused, thematic collections
•Support priority subjects

Special collection

UK Domain Crawl

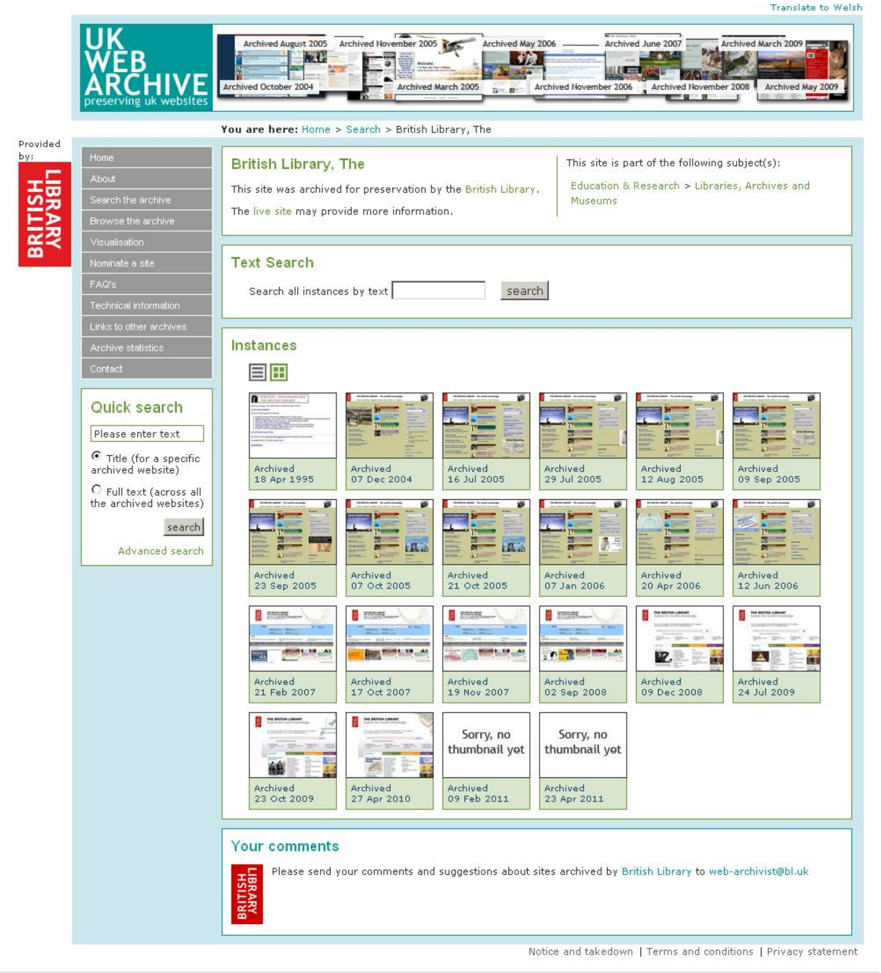
2013 domain crawl stats

- 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



Translate to Welsh

UK WEB ARCHIVE preserving uk websites

Archived August 2005 Archived November 2005 Archived May 2006 Archived June 2007 Archived March 2009
Archived October 2004 Archived March 2005 Archived November 2006 Archived November 2008 Archived May 2009

You are here: Home > Search > British Library, The

Provided by:

- Home
- About
- Search the archive
- Browse the archive
- Visualisation
- Nominate a site
- FAQ's
- Technical information
- Links to other archives
- Archive statistics
- Contact

Quick search

Please enter text

● Title (for a specific archived website)

○ Full text (across all the archived websites)

search

Advanced search

British Library, The

This site was archived for preservation by the British Library.

The live site may provide more information.

This site is part of the following subject(s):

Education & Research > Libraries, Archives and Museums

Text Search

Search all instances by text [ ] search

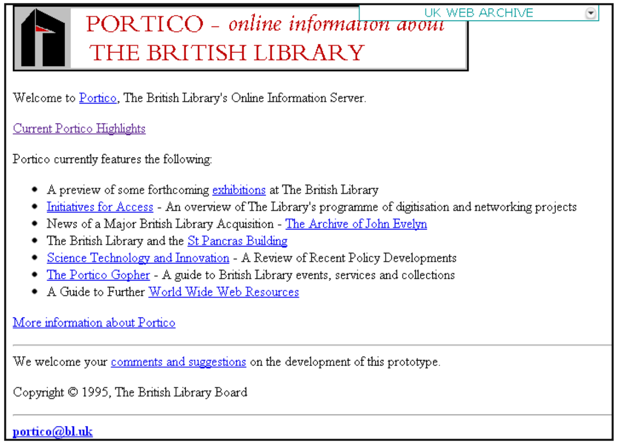
Instances

Archived 18 Apr 1995
Archived 07 Dec 2004
Archived 16 Jul 2005
Archived 29 Jul 2005
Archived 12 Aug 2005
Archived 09 Sep 2005
Archived 23 Sep 2005
Archived 07 Oct 2005
Archived 21 Oct 2005
Archived 07 Jan 2006
Archived 20 Apr 2006
Archived 12 Jun 2006
Archived 21 Feb 2007
Archived 17 Oct 2007
Archived 19 Nov 2007
Archived 02 Sep 2008
Archived 09 Dec 2008
Archived 24 Jul 2009
Archived 23 Oct 2009
Archived 27 Apr 2010
Sorry, no thumbnail yet Archived 09 Feb 2011
Sorry, no thumbnail yet Archived 23 Apr 2011

Your comments

Please send your comments and suggestions about sites archived by British Library to web-archivist@bl.uk

Notice and takedown | Terms and conditions | Privacy statement



UK WEB ARCHIVE

PORTICO – *online information about*
THE BRITISH LIBRARY

Welcome to Portico, The British Library's Online Information Server.

Current Portico Highlights

Portico currently features the following:

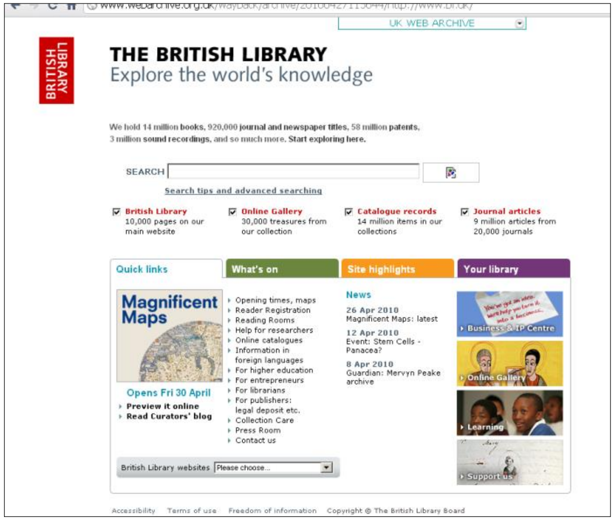
- A preview of some forthcoming exhibitions at The British Library
- Initiatives for Access - An overview of The Library's programme of digitisation and networking projects
- News of a Major British Library Acquisition - The Archive of John Evelyn
- The British Library and the St Pancras Building
- Science Technology and Innovation - A Review of Recent Policy Developments
- The Portico Gopher - A guide to British Library events, services and collections
- A Guide to Further World Wide Web Resources

More information about Portico

We welcome your comments and suggestions on the development of this prototype.

Copyright © 1995, The British Library Board

portico@bl.uk



www.webarchive.org.uk/wayback/archive/20100427115044/http://www.bl.uk/

UK WEB ARCHIVE

THE BRITISH LIBRARY
Explore the world's knowledge

We hold 14 million books, 920,000 journal and newspaper titles, 58 million patents, 3 million sound recordings, and so much more. Start exploring here.

SEARCH [ ]

Search tips and advanced searching

☑ British Library 10,000 pages on our main website
☑ Online Gallery 30,000 treasures from our collection
☑ Catalogue records 14 million items in our collections
☑ Journal articles 9 million articles from 20,000 journals

Quick links | What's on | Site highlights | Your library

Magnificent Maps

Opens Fri 30 April

- Preview it online
- Read Curators' blog

- Opening times, maps
- Reader Registration
- Reading Rooms
- Help for researchers
- Online catalogues
- Information in foreign languages
- For higher education
- For entrepreneurs
- For librarians
- For publishers: legal deposit etc.
- Collection Care
- Press Room
- Contact us

News

26 Apr 2010
Magnificent Maps: latest

12 Apr 2010
Event: Stem Cells - Panacea?

8 Apr 2010
Guardian: Mervyn Peake archive

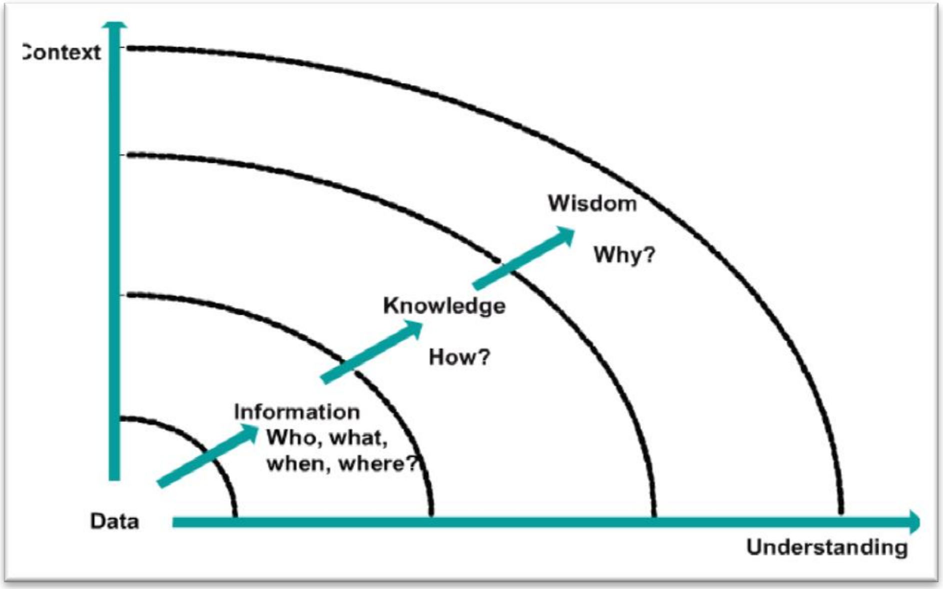
- Business & IP Centre
- Online Gallery
- Learning
- Support us

British Library websites Please choose...

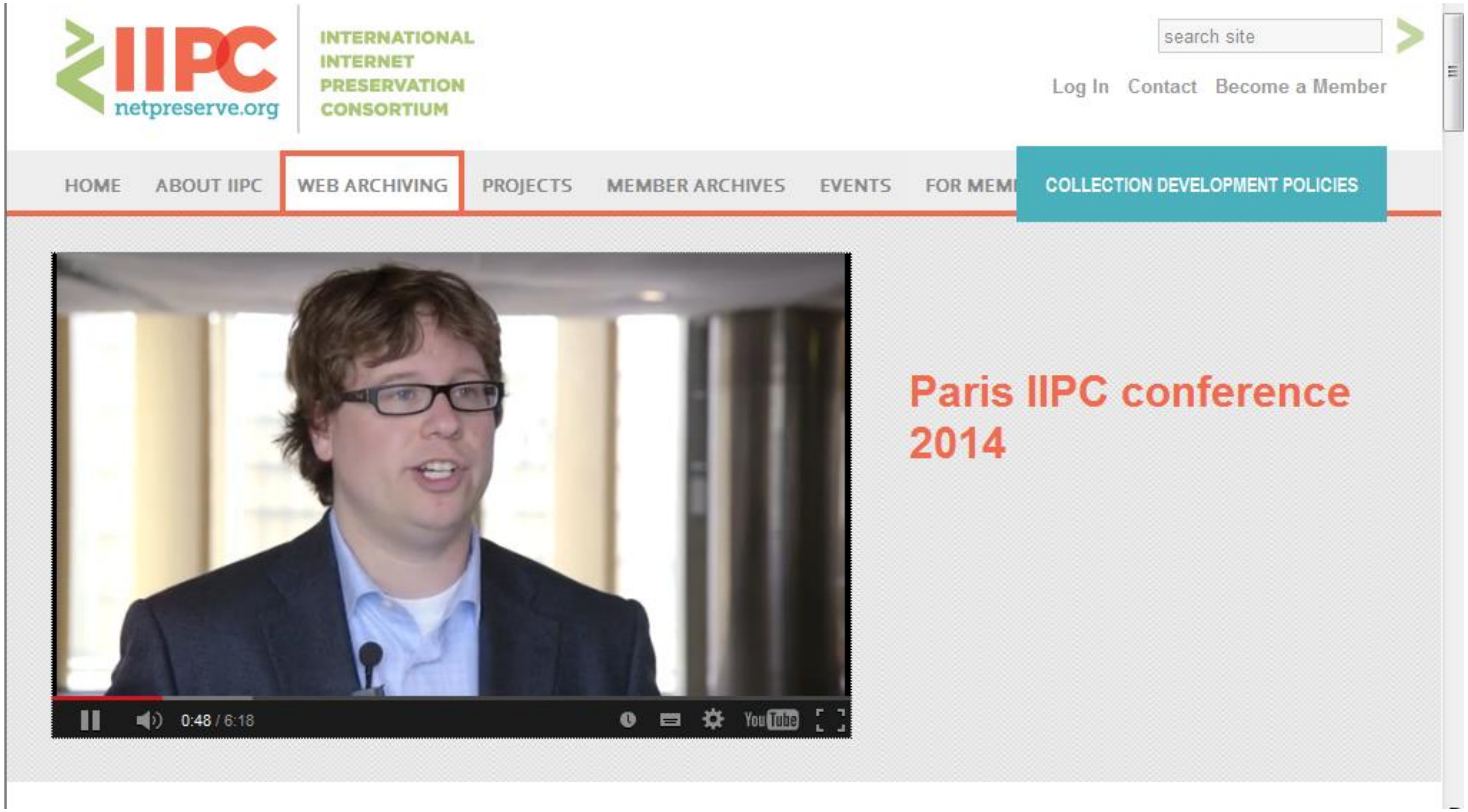
Accessibility Terms of use Freedom of information Copyright © The British Library Board

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

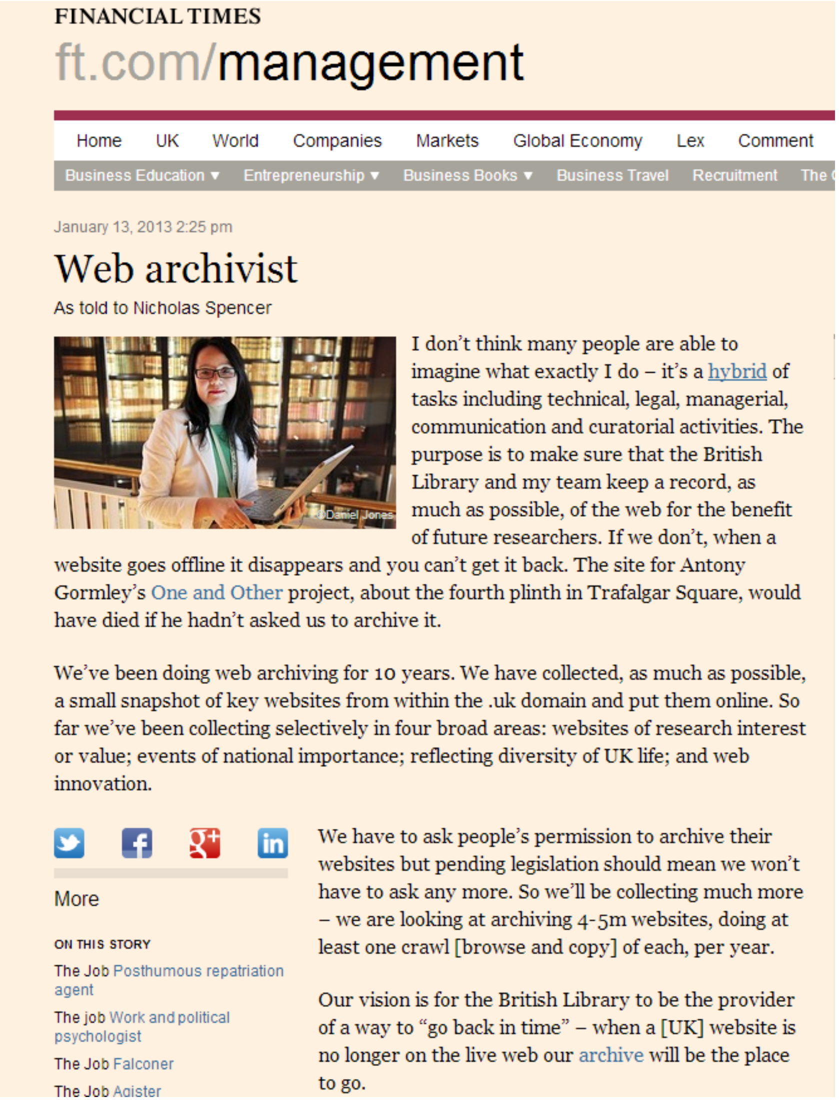


Paris IIPC conference 2014

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

My Job

"…a hybrid of tasks including technical, legal, managerial, communication and curatorial activities."



January 13, 2013 2:25 pm

Web archivist

As told to Nicholas Spencer

I don't think many people are able to imagine what exactly I do – it's a hybrid of tasks including technical, legal, managerial, communication and curatorial activities. The purpose is to make sure that the British Library and my team keep a record, as much as possible, of the web for the benefit of future researchers. If we don't, when a website goes offline it disappears and you can't get it back. The site for Antony Gormley's One and Other project, about the fourth plinth in Trafalgar Square, would have died if he hadn't asked us to archive it.

We've been doing web archiving for 10 years. We have collected, as much as possible, a small snapshot of key websites from within the .uk domain and put them online. So far we've been collecting selectively in four broad areas: websites of research interest or value; events of national importance; reflecting diversity of UK life; and web innovation.

We have to ask people's permission to archive their websites but pending legislation should mean we won't have to ask any more. So we'll be collecting much more – we are looking at archiving 4-5m websites, doing at least one crawl [browse and copy] of each, per year.

Our vision is for the British Library to be the provider of a way to "go back in time" – when a [UK] website is no longer on the live web our archive will be the place to go.

ON THIS STORY

The Job Posthumous repatriation agent

The job Work and political psychologist

The Job Falconer

The Job Agister

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