

Archiving the UK Web

Helen Hockx-Yu
Head of Web Archiving
British Library

20 January 2014

6th April 2013...

Legal Deposit Libraries (Non-Print Works) Regulations 2013

<http://www.youtube.com/watch?v=PhHpRMsoq64>



Regulations cover

- *Collecting the material*

1. Harvesting the open web (Legal Deposit UK web archive)

2. Publisher deposit of online works (e.g. e-journals/e-books...) by mutual agreement

3. Harvesting behind pay walls and password barriers

4. Publisher deposit of offline works (on CD-ROM...) by obligation

- Micro-businesses excluded from 3 & 4 until April 2014

- Print or non-print version, not both

- *Using the material*

- Access only on Libraries' premises

- One concurrent user per work

- No digital copying

- Restricted printing

- "Perpetual copyright"?

Other issues

- Post-implementation review

- Dispute resolution procedure

UK Legal Deposit libraries



The British Library

Bodleian Libraries of the University of Oxford

Cambridge University Library

The National Library of Scotland

The Library of Trinity College, Dublin

The National Library of Wales

UK Web Archive

- **Open UK Web Archive**
- Since 2004
- Based on websites owners' permissions
- Over 14,000 selected UK websites & 60,000 instances, 18.8TB
- **Legal Deposit UK Web Archive**
- Since April 2013
- Legal deposit content, all UK websites
- Only accessible on Legal Deposit Libraries' premises

Translate to Welsh

UK WEB ARCHIVE
preserving uk websites

You are here: Home

Provided by:
BRITISH LIBRARY

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

Welcome to the UK Web Archive
Thousands of UK websites have been collected since 2004 and the Archive is growing fast.
Here you can see how sites have changed over time, locate information no longer available on the live Web and observe the unfolding history of a spectrum of UK activities represented online. Sites that no longer exist elsewhere are found here and those yet to be archived can be saved for the future by nominating them.
The Archive contains sites that reflect the rich diversity of lives and interests throughout the UK. Search is by Title of Website, Full Text or URL, or browse by Subject, Special Collection or Alphabetical List.

Quick website links
What is the UK Web Archive?
Who is the UK Web Archive for?
How do I search the archive?
How can I nominate a website?

Quick search
Please enter text
Title (for a specific archived website)
Full text (across all the archived websites)
search
Advanced search

Explore the Special Collections
Special Collections are groups of websites brought together on a particular theme by librarians, curators and other specialists, often working in collaboration with key organisations in the field. They can be events-based (e.g The Olympic & Paralympic Games 2012), topical (e.g. The Credit Crunch Collection) or subject-oriented (e.g. The British Countryside Collections).

Education & Research
Arts & Humanities
Science & Technology
Government, Law & Politics
Society & Culture
Business, Economy & Industry
Medicine & Health
Browse by Subject

New! N-gram Search
N-gram Search

Initiated by:
UK WEB ARCHIVING CONSORTIUM

Notice and takedown | Terms and conditions | Privacy statement

<http://www.webarchive.org.uk>

Same job, just bigger, more complex

- Over 10 million .uk registered domains
 - 4th TLD after .com, .de and .net
 - UK organisations also use non .uk domain names (eg .com or .org) – scale unknown
- Legal change: implementing non-print Legal Deposit, no need to request / administer permissions
- Organisational change: collaborative effort of six Legal Deposit Libraries (LDLs), new ways of working together
- New workflows, new tools
- Access and use – how to manage user expectations related to restricted access
- Ongoing capability of archiving the evolving web

Collecting strategy

Domain Crawl

Event

- Domain harvesting:
- Broad sweep of .uk domain
 - Once or twice a year

Key sites

- Events & key sites:
- Events of national interest
 - Sites need to be captured frequently eg bbc.co.uk

Event

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

Progress

Description	Duration of crawl	#of seeds	# of URLs collected	Size
Focused crawls of websites related to NHS Reform	7 weeks	1,100	180 million	1.8TB
UK Domain Crawl 2013	10 weeks	3.8 million	1.4 billion	31.5TB

Plus:

Margaret Thatcher and Nelson Mandela collection

Ongoing crawls of 250+ selected websites & 500+ news sites

Access strategy

- Legal Deposit content cannot be accessed outside libraries' reading rooms.
- Online access to metadata and selected content to showcase the Legal Deposit web archive of the UK
 - Bibliographic metadata
 - Analysis and visualisation of *aggregated* content
 - Statistical and contextual data
 - Copy of deposited content with direct permission
- For sites from outside the UK, permissions both for harvesting and for public access will be required

Archived website as historical document

UK WEB ARCHIVE preserving uk websites

Translate to Welsh

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

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

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		

Quick search

Please enter text

Title (for a specific archived website)

Full text (across all the archived websites)

Advanced search

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

UK WEB ARCHIVE

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

Magnificent Maps

Opens Fri 30 April
Preview it online
Read Curators' blog

What's on

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

Site highlights

News

- 26 Apr 2010 Magnificent Maps: latest
- 12 Apr 2010 Event: Stem Cells - Panacea?
- 8 Apr 2010 Guardian: Mervyn Peake archive

Your library

- Business Support Centre
- Online Gallery
- Learning
- Support Us

British Library websites

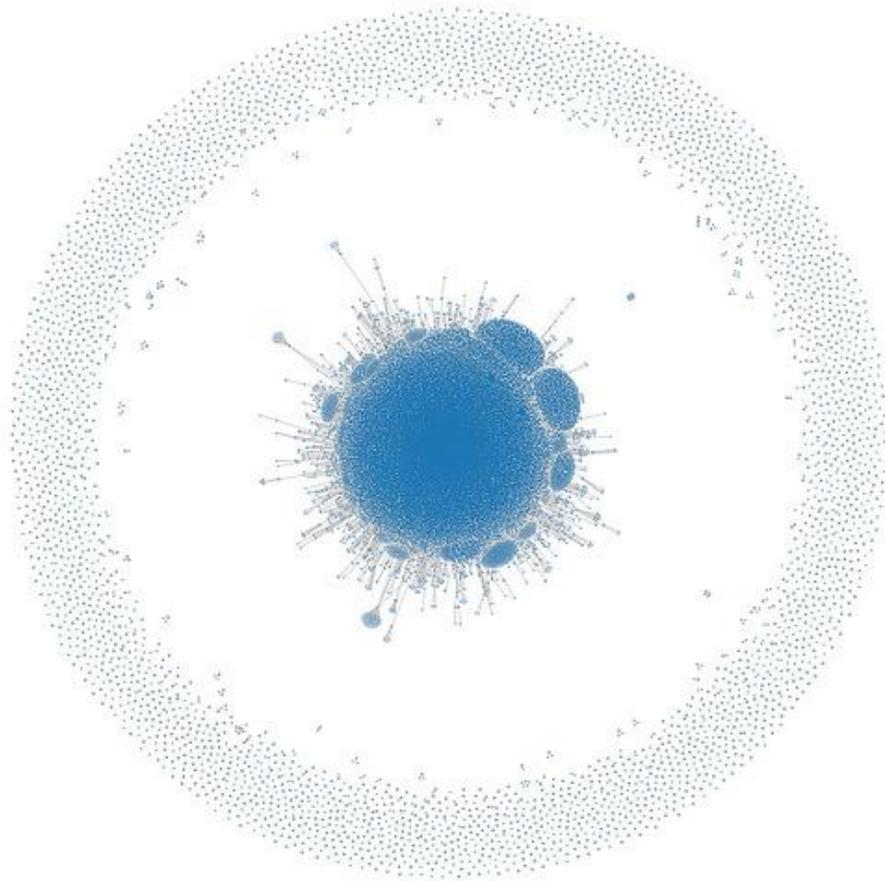
Accessibility | Terms of use | Freedom of information | Copyright © The British Library Board

Analytics and visualisation

- Shift of focus from the level of single webpages or websites to the entire web archive collection.
- Use web archives as datasets, access to metadata and knowledge about websites
- Support survey, annotation, contextualisation and visualisation
- Allows discovery of patterns, trends and relationships in inter-linked web pages

Visit [UK Web Archive](#) to see our work

How was the UK web linked in 1996?



- By Rainer Simon using UK Host-Level Link Graph (1996-2010) dataset.
- Based on the 1996 portion: 58,842 hosts (nodes); 184,433 host-to-host links (edges)
- UK web as part of the global web
- Scalability issues with large dataset over time

My Job

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

FINANCIAL TIMES

ft.com/management

[Home](#) [UK](#) [World](#) [Companies](#) [Markets](#) [Global Economy](#) [Lex](#) [Comment](#)[Business Education](#) [Entrepreneurship](#) [Business Books](#) [Business Travel](#) [Recruitment](#) [The](#)

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.



More

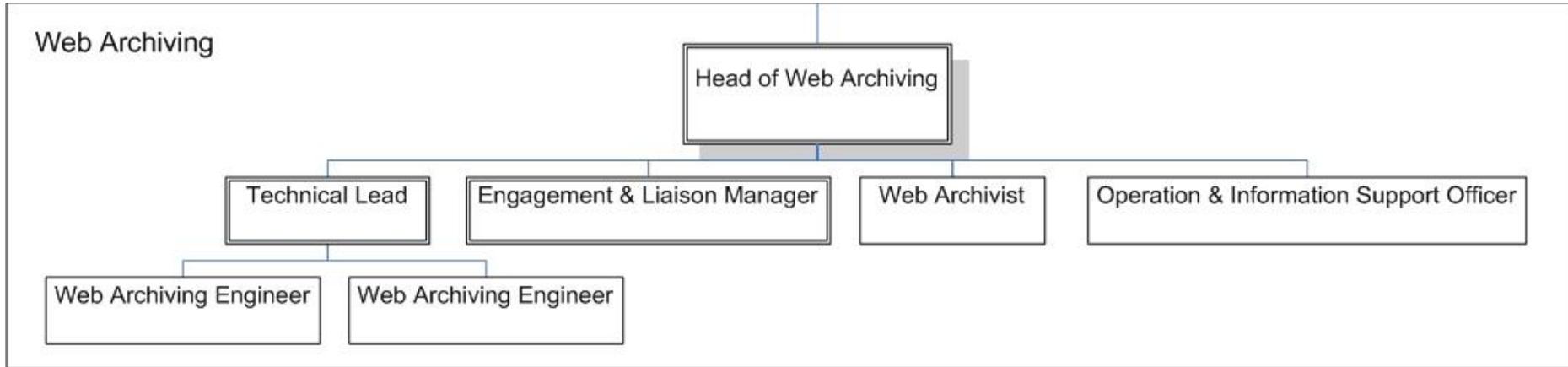
ON THIS STORY

[The Job Posthumous repatriation agent](#)[The job Work and political psychologist](#)[The Job Falconer](#)[The Job Aalster](#)

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.

My Team



Skills profile: cross-disciplinary expertise

- IT
- Collection management, digital curation
- Management
- Communications
- Web Archiving

My week (20-24 January)

Meetings (9 hours)

- Team meeting, one to one meetings with team members, with line manager and team leaders of wider business unit; with line manager and Advisory Group chair – all regular meetings
- Library-wide Digital Content Development Group
- Project board / project meetings
- New project: document harvesting
- IT restructuring / funding cuts

Emails!

Tasks

- DPC student conference
- International Internet Preservation Consortium (IIPC): planning for General Assembly in May
- Contribute to UK Web Archive / IIPC Twitter accounts
- Compile reports and read papers for meetings, complete actions required
- Project management: Open Wayback
- HR tasks: recruitment
- Planning: roadmap, resource planning, hardware
- Policy: open access strategy for web archives, notice and takedown

Some reflections

- Our jobs do not just cover a single field. Require general understanding of multiple areas but still helpful to specialise in one area
- Implementing NDLP: policy, technology, people
- Continuous learning / training / professional development
 - Linux Admin, Django, Git, IT service management
 - Digital research, methodology, practice and requirements
 - Publications / knowledge of the field
- Think and plan ahead
 - mobile web
 - development of key software

Some “advice”

- Employability: "the capability of getting and keeping satisfactory work".
- Employability skills: “a set of achievements, understandings and personal attributes that make individuals more likely to gain employment and to be successful in their chosen occupations”. - Peter Knight & Mantz Yorke
- A degree is not enough, it unlocks doors
- Communication skills are the most important – eg self-presentation (on application and interviews)
- Need to be open-minded and flexible to deal with changes

Join us

- [Web Archiving Engagement and Liaison Officer](#)
 - Closing date 4 February
 - Interviews on 11 February 2014.
- Twitter: @ukwebarchive