Mesa 8: Preservação de Conteúdo Web

Michael Day
Digital Preservation Research Lead
The British Library
Michael.Day@bl.uk

V SINPRED (Seminário Internacional de Preservação Digital), Universidade Estadual de Campinas, Brazil, 13 May 2021, via Zoom

Outline

Contexts:

- The British Library
- Legal Deposit in the United Kingdom

The UK Web Archive:

- The selective archiving of content, from 2004
- Part of Non-Print Legal Deposit, since 2013

Practicalities:

- Collections
- Tools
- Collaboration and engagement
- Digital preservation

Lessons learned

The British Library

Some facts and figures:

- The national library of the United Kingdom
- Established by British Library Act 1972

Based on two sites:

- London (St Pancras)
- Boston Spa, Yorkshire

One of six UK Legal Deposit Libraries

Collections:

- >170 million collection items
- >746 miles of shelving
- >1 PB of digital content, and growing …

Legal Deposit in the UK

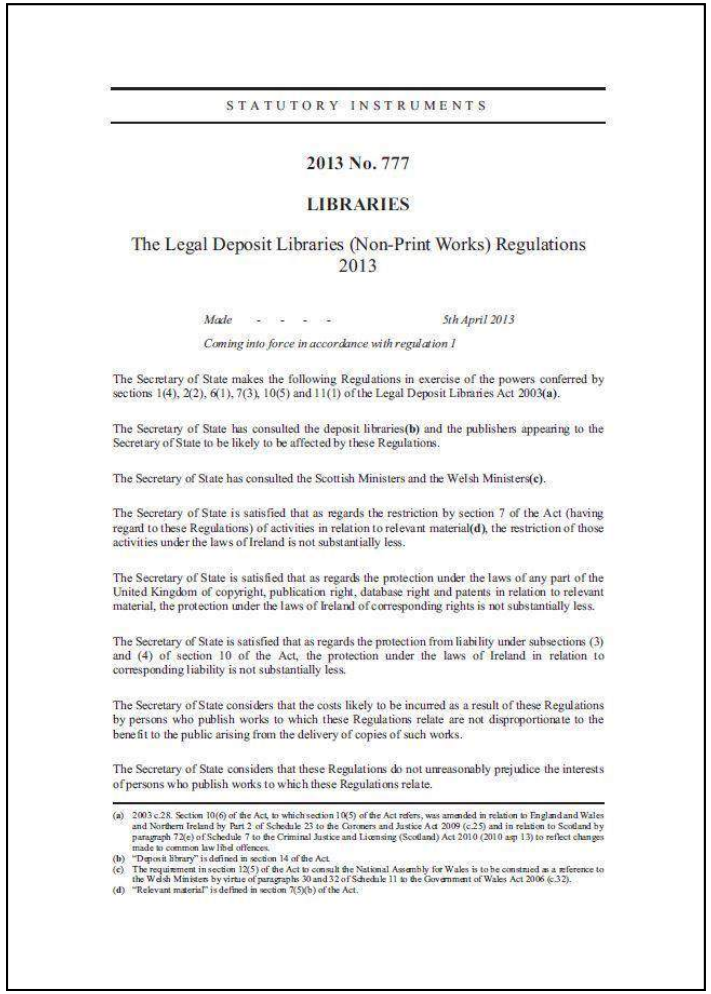
- Preserving a published record of UK culture, politics, society, technology, etc.
- Legal Deposit first established in the UK in 1662
- Rights were granted to the British Museum in 1757 (the British Library inherited these in 1973)
- Six Legal Deposit Libraries (LDLs) since 1911
- Legal Deposit is the statutory right for the six LDLs to receive copies of all printed works published in the UK and Republic of Ireland (with some exceptions)
- Legal Deposit was extended to cover Non-Print Legal Deposit (NPLD) publications in 2013

Non-Print Legal Deposit (NPLD)

The Legal Deposit Libraries Act 2003

Legal Deposit (Non-Print Works) Regulations 2013

- Came into operation on 6th April 2013
- LD provisions extended to cover material published digitally and online, “so that the Legal Deposit Libraries can provide a national archive of the UK's non-print published material, such as websites, blogs, e-journals and CD-ROMs.”
- Access is restricted to on-site readers
- Collection focus to date on: eJournals; eBooks; geospatial data, music scores … and the web http://www.legislation.gov.uk/ukdsi/2013/9780111533703



STATUTORY INSTRUMENTS

2013 No. 777

LIBRARIES

The Legal Deposit Libraries (Non-Print Works) Regulations 2013

Made - - - - 5th April 2013

Coming into force in accordance with regulation 1

The Secretary of State makes the following Regulations in exercise of the powers conferred by sections 1(4), 2(2), 6(1), 7(3), 10(5) and 11(1) of the Legal Deposit Libraries Act 2003(a).

The Secretary of State has consulted the deposit libraries(b) and the publishers appearing to the Secretary of State to be likely to be affected by these Regulations.

The Secretary of State has consulted the Scottish Ministers and the Welsh Ministers(c).

The Secretary of State is satisfied that as regards the restriction by section 7 of the Act (having regard to these Regulations) of activities in relation to relevant material(d), the restriction of those activities under the laws of Ireland is not substantially less.

The Secretary of State is satisfied that as regards the protection under the laws of any part of the United Kingdom of copyright, publication right, database right and patents in relation to relevant material, the protection under the laws of Ireland of corresponding rights is not substantially less.

The Secretary of State is satisfied that as regards the protection from liability under subsections (3) and (4) of section 10 of the Act, the protection under the laws of Ireland in relation to corresponding liability is not substantially less.

The Secretary of State considers that the costs likely to be incurred as a result of these Regulations by persons who publish works to which these Regulations relate are not disproportionate to the benefit to the public arising from the delivery of copies of such works.

The Secretary of State considers that these Regulations do not unreasonably prejudice the interests of persons who publish works to which these Regulations relate.

(a) 2003 c.28. Section 10(6) of the Act, to which section 10(5) of the Act refers, was amended in relation to England and Wales and Northern Ireland by Part 2 of Schedule 23 to the Coroners and Justice Act 2009 (c.25) and in relation to Scotland by paragraph 72(e) of Schedule 7 to the Criminal Justice and Licensing (Scotland) Act 2010 (2010 asp 13) to reflect changes made to common law libel offences.
(b) "Deposit library" is defined in section 14 of the Act.
(c) The requirement in section 12(5) of the Act to consult the National Assembly for Wales is to be construed as a reference to the Welsh Ministers by virtue of paragraphs 30 and 32 of Schedule 11 to the Government of Wales Act 2006 (c.32).
(d) "Relevant material" is defined in section 7(5)(b) of the Act.

The UK Web Archive - origins

- First established in 2004:
- Based on the recommendations of a feasibility study commissioned by JISC and the Wellcome Library (2003)
- A collaborative endeavour from the start (UK Web Archiving Consortium)
- Collected websites after getting permission from website owners
- Resource intensive from a curation perspective
- Content made openly available through: www.webarchive.org.uk
- Collection = >15K websites

Non-Print Legal Deposit (NPLD)

The Legal Deposit Libraries (Non-Print Works) Regulations

- From April 2013
- Enabled the collection of UK websites at scale for the first time

Regulations define parameters as works:

- a) made available to the public from a website with a domain name which relates to the UK; or
- b) made available to the public by a person and any of that person's activities relating to the creation or the publication of the work take place within the United Kingdom.

In practical terms covers:

- UK top-level domains, including .uk, .scot, .cymru, .London, etc.
- Geo-IP database look up for UK servers

Non-Print Legal Deposit (NPLD)

The regulations do not cover the collection of:

- Film and recorded sound where audio-visual content predominates
- Private intranets and emails
- Personal data in social networking sites or that are only available to restricted groups.

Restricted access:

- Unlike the ‘Open’ web archive, end-user access to NPLD content is restricted to the reading rooms of the UK Legal Deposit Libraries

UK Web Archive - Collections

Annual capture of the UK domain + selected websites on a more frequent basis

- Domain crawl collects ca. 6-10 websites (>2 billion ‘items’), 70-100 TB of compressed data

Special collections:

- ‘Topics and Themes’ section of the UK Web Archive website contains more than 100 curated collections of websites related to research, life and events in the UK

Users of the UK Web Archive are also invited to submit information about UK websites that they think it should be archiving

UK Web Archive – Tools

Heritrix 3: web crawler / harvester software (maintained by International Internet Preservation Consortium)

Annotation and Curation Tool (ACT), facilitates a number of tasks:

- Adding metadata
- Specifying how often a website is captured
- Quality control
- Changing access parameters (e.g., for open-access content)

Webrecorder (a tool developed by Rhizome)

- Experiments with collecting more-interactive content, e.g. social media for 2019 UK General Election: https://blogs.bl.uk/webarchive/2020/05/index.html

Collaboration and engagement

The UK Web Archive is still a very highly collaborative venture:

- A partnership of all six UK Legal Deposit Libraries
- Special collections developed in collaboration with many other organisations, individual researchers, etc.
- IIPC Collaborative Collections, e.g. on Olympic Games, Covid-19

Engagement with researchers, journalists, etc.

- Funded research projects, placements, etc.
- SHINE service: a portal able to search .uk websites collected by the Internet Archive between 1996 and 2013: https://www.webarchive.org.uk/shine

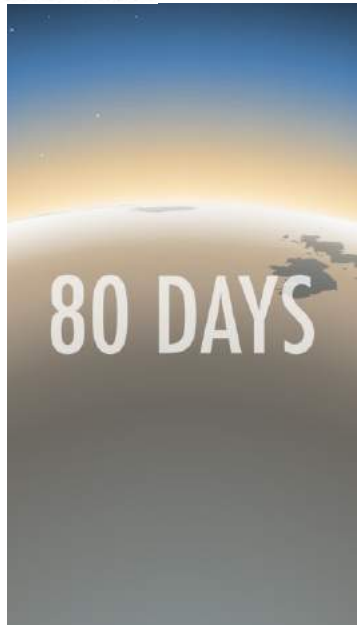
Emerging Formats

UK Legal Deposit Libraries exploring new kinds of published content and how they might be considered in terms of Legal Deposit

- eBooks published as Apps
- Web-based interactive narratives
- Structured data

Experiments with Annotation and Curation Tool (ACT) and Webrecorder to test the collection of interactive narratives: https://blogs.bl.uk/digital-scholarship/2019/04/collecting-emerging-formats.html

Collaboration with other organisations dealing with complex digital objects, e.g. digital art, computer games



Digital preservation

UK Web Archive

- The large-scale (and rapid growth) of web collections brings its own challenges
- Most content is packaged in the WARC format (aggregated archival files with associated information), but these ultimately represent many different formats and content types (e.g. HTML, images, audio-visual, software, etc.)

Preservation planning

- Web content may require different preservation planning approaches to other kinds of digital collections
- What should we be preserving? The user experience (browser apps)? The web as a graph or dataset? Individual documents?

Digital Preservation Awards 2020

‘15 Years of the UK Web Archive’ won The National Archives (UK) Award for Safeguarding the Digital Legacy, November 2020: https://blogs.bl.uk/webarchive/2020/11/web-archive-team-wins-2020-digital-preservation-award.html

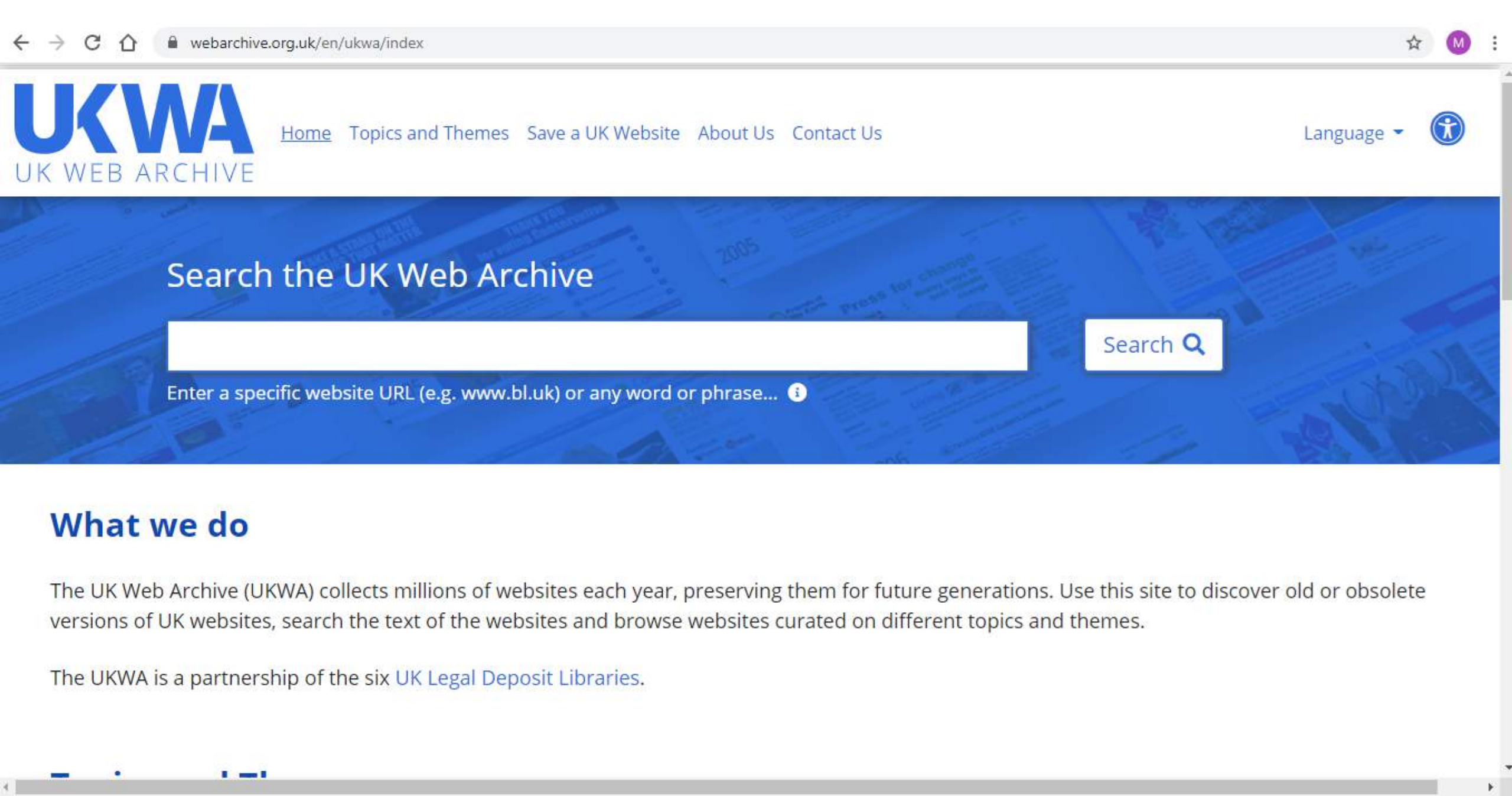
Lessons learned

People and skills are important – a combination of technical and curatorial expertise is required

Collaborate wherever possible – e.g. on development of tools and approaches, on collections, and on the use of archives

Engage with potential users – current restrictions on access to Non-Print Legal Deposit collections means that web archives have to be pro-active in facilitating use

Keep thinking about preservation requirements – we need to understand more about the challenges of preserving web archives, and how to integrate them with other digital collections


webarchive.org.uk/en/ukwa/index
UKWA
UK WEB ARCHIVE
Home
Topics and Themes
Save a UK Website
About Us
Contact Us
Language
Search the UK Web Archive
Search
Enter a specific website URL (e.g. www.bl.uk) or any word or phrase...
What we do
The UK Web Archive (UKWA) collects millions of websites each year, preserving them for future generations. Use this site to discover old or obsolete versions of UK websites, search the text of the websites and browse websites curated on different topics and themes.
The UKWA is a partnership of the six UK Legal Deposit Libraries.

webarchive.org.uk/en/ukwa/search?text=tower+of+london&search_location=full_text&reset_filters=false&content_type=Web+Page

UKWA
UK WEB ARCHIVE

Home Topics and Themes Save a UK Website About Us Contact Us Language

Clear all filters

Access: Viewable online

Document type (Include):

"Web Page" ×

tower of london

Search

Enter a specific website URL (e.g. www.bl.uk) or any word or phrase...

Tips/Notes for using the UK Web Archive

Accessing Content

- [x] Viewable Online (375,743,677)
- [ ] At libraries (3,087,321,707)

Domain

Search results: 375,743,677 results for "tower of london"

Sort by Most relevant

Items per page 50

1 2 3 4 5 Next

Domain

- gov.uk (29,111,862)
- bbc.co.uk (23,182,755)
- twitter.com (13,458,686)
- Show more

Document Type

- Web Page (375,743,677)
- Show more

Suffix

- gov.uk (115,423,674)

Search results

http://images.hrp.org.uk/en/set/show_content_page.html?category=24&set=9&qw=

Palace highlights - **Tower of London** Use search (top right) to find more 200 Assets X Preview Download
Date collected: 2015-12-05

Search results

http://images.hrp.org.uk/?service=category&action=show_content_page&language=en&category=12

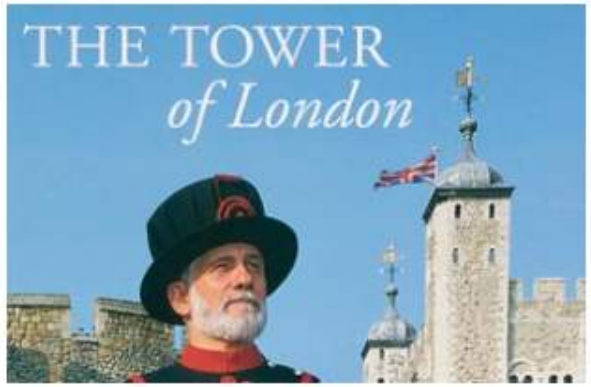
- The Imperial Crown **of** India - **Tower of London** X Preview Download Comp HRP06131 HRP06131 - The Imperial
Date collected: 2015-12-05

Search results

http://images.hrp.org.uk/?service=category&action=show_content_page&language=en&category=7

Warder uniform - **Tower of London** X Preview Download Comp HRP11092 HRP11092 - The Chief Yeoman Warder
Date collected: 2015-12-05

Further reading



The bibliography of the Tower of London is vast. The following list represents some published works believed to represent the history of the site most accurately, though some less-scholarly references are also included.

General history of the Tower

webarchive.org.uk/en/ukwa/collection

Topics and Themes

Explore over 100 collections of websites brought together by librarians, curators and other experts in response to a wide range of events and diverse topics and themes.

19th Century English Literature

Aging

This collection looks at aging

Arnhem 75

This is a collection of websites

Black and Asian Britain

Collection focussing on Black

Cardiff Branch, Glamorgan Family History Society (@Cardiff_GFHS) on Twitter

Viewable only on Library premises

The Twitter account for Cardiff Branch, Glamorgan Family ...

Archived date: 2020-02-19

https://twitter.com/cardiff_gfhs/

Cardiganshire Family History Society | Cymdeithas Hanes Teuluoedd Ceredigion

Family History

Archived date: 2012-03-29

http://www.cgnfhs.org.uk/

Cathays Cemetery Monuments Archive

As of 2018, Cathays Cemetery Monuments Archive contained ...

Archived date: 2018-09-20

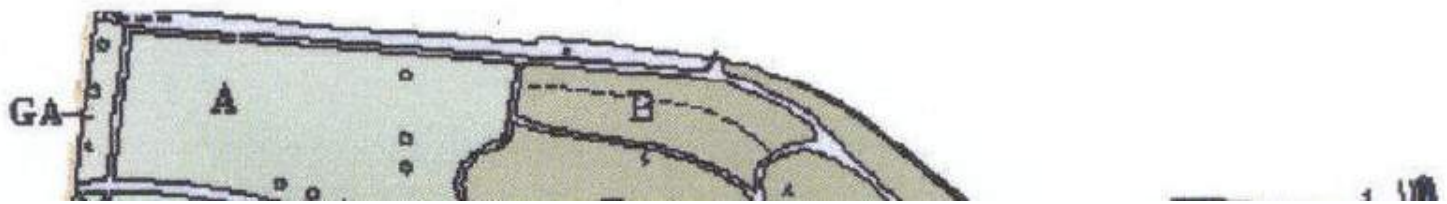
http://cardiffnorthwalkers.coffeecup.com/

Cathy's Home Page

Cathays Cemetery Monuments Archive

SECTION SELECTION

Click on Section required.



Links and further reading

- UK Web Archive: https://www.webarchive.org.uk/
- UK Web Archive blog: https://britishlibrary.typepad.co.uk/webarchive/
- British Library collection guide: https://www.bl.uk/collection-guides/uk-web-archive
- British Library Digital Preservation: https://www.bl.uk/digital-preservation
- Papers:
    - Nicola Bingham, Quality Assurance Paradigms in Web Archiving Pre and Post Legal Deposit (2014): https://doi.org/10.7227/ALX.0020
    - Nicola Bingham and Helena Byrne, Archival strategies for contemporary collecting in a world of big data: Challenges and opportunities with curating the UK web archive (2021): https://doi.org/10.1177/2053951721990409

BRITISH LIBRARY

Thank you

Obrigado