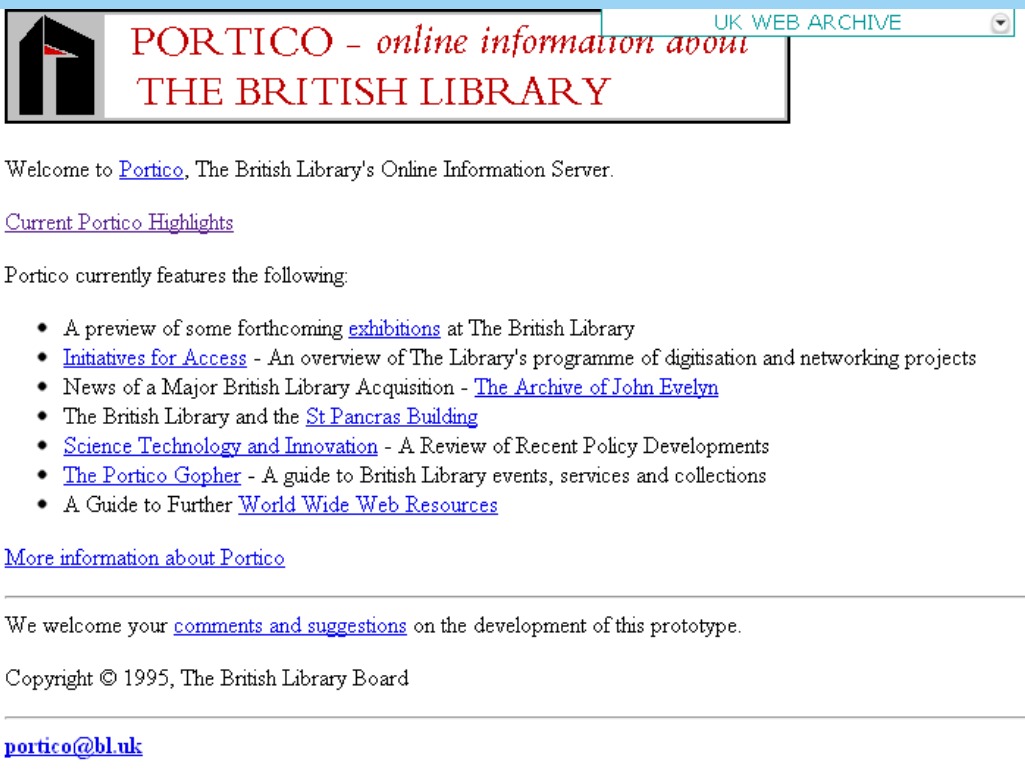


Enriching the UK Web Archive by ingesting non-harvested content



The screenshot shows a browser window with the title "UK WEB ARCHIVE" and a dropdown arrow. The main heading reads "PORTICO - online information about THE BRITISH LIBRARY". Below this, a welcome message states "Welcome to [Portico](#), The British Library's Online Information Server." A link for "Current Portico Highlights" is provided. A section titled "Portico currently features the following:" lists several items with bullet points, each containing a link to a specific page or resource. At the bottom, there is a link for "More information about Portico", a welcome message for comments and suggestions, a copyright notice for 1995, and an email address "portico@bl.uk".

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

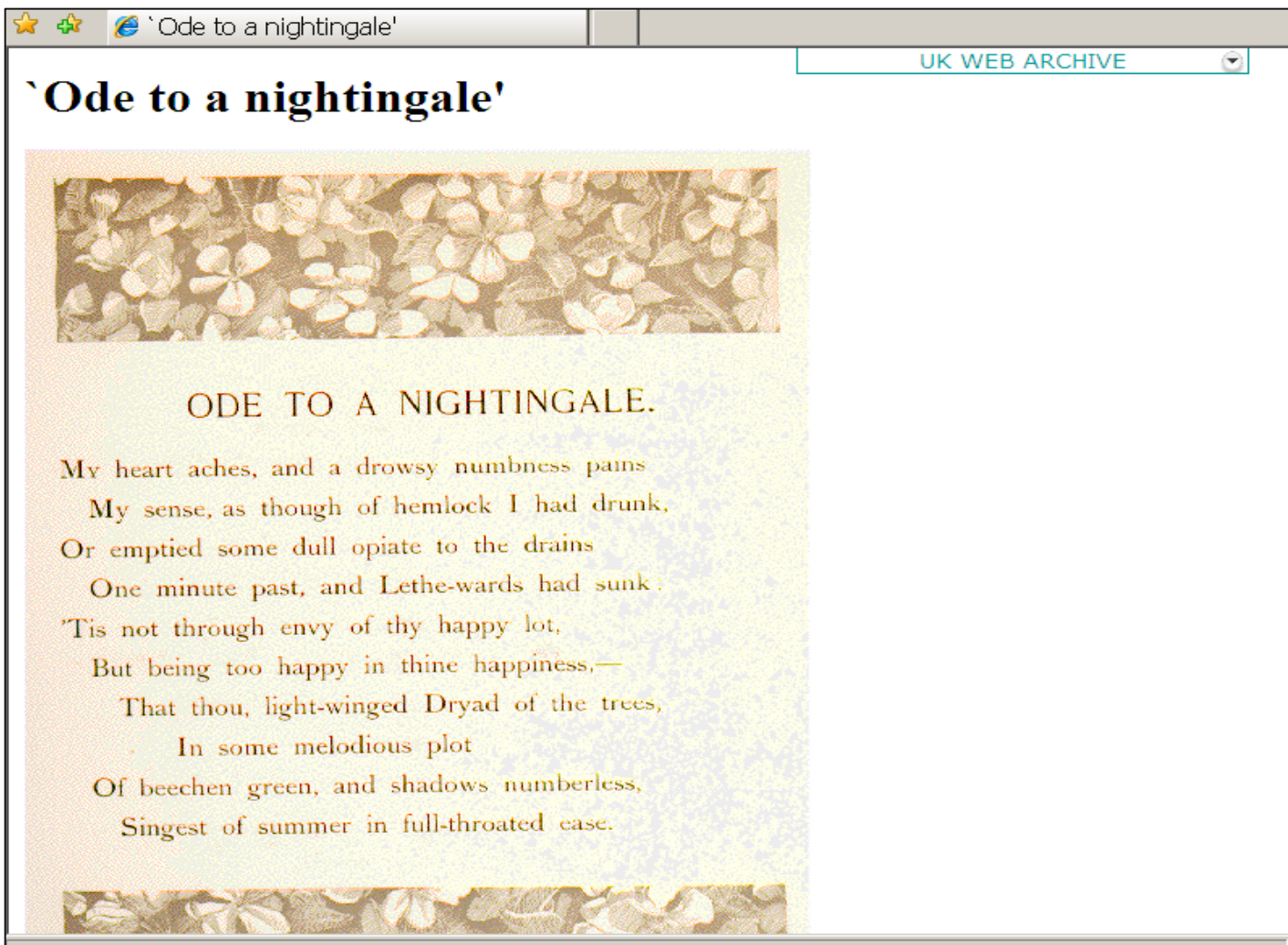
Helen Hockx-Yu

Head of Web Archiving
British Library

Roger Coram

Web Archiving Engineer
British Library

John Keats: *Ode to a Nightingale*



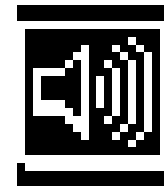
★ + 'Ode to a nightingale'

UK WEB ARCHIVE

'Ode to a nightingale'

ODE TO A NIGHTINGALE.

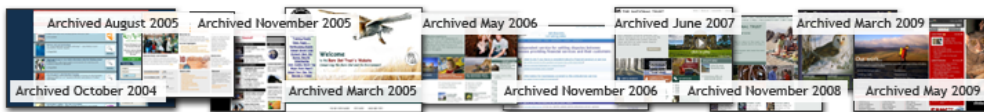
My heart aches, and a drowsy numbness pains
My sense, as though of hemlock I had drunk,
Or emptied some dull opiate to the drains
One minute past, and Lethe-wards had sunk:
'Tis not through envy of thy happy lot,
But being too happy in thine happiness,—
That thou, light-winged Dryad of the trees,
In some melodious plot
Of beechen green, and shadows numberless,
Singest of summer in full-throated ease.



(266K
bytes)

Earliest instance in a web archive?

[Translate to Welsh](#)



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

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

Quick search

Please enter text

- Title (for a specific archived website)
- Full text (across all the archived websites)

[Advanced search](#)

Ingesting non-harvested content

- Content in a web archive is commonly harvested (downloaded or crawled) using a crawler software
 - arguably the most cost-effective method of collecting websites automatically
- But harvesting is not always possible
 - Website no longer live
 - Website owners prefer to deposit files
 - Archiving organisations choose to obtain files directly
- Import from disk and ingest into archive

The 1995 BL website

- BL's first explorations into hypertext and embedded images from collection material with links to larger images, sound files and further information.
- A test copy of the BL website found on an internal server
- Not the complete dataset
- Linked content hosted then on a Gopher server not recovered
- Received as 12MB zip file in the original site's directory structure, dated 18 April 1995
 - .html, .gif, .au

Steps taken

- Unpack zip and convert to WARC
- Use stored procedure to create entry in database
- Copy files to disk, triggering indexing process
- Access via Wayback
 - Had to change Wayback setting and amend *org.archive.wayback.util.Timestamp* which defines YEAR_LOWER_LIMIT as 1996
- Amendment of SIP
 - No logs, crawler configuration
 - MET profile
- Note by web archivist in UI, explaining “unusual” archival process

Developing a Formal Process

- Currently doing this only as exceptions, upon specific requests and reviewed case by case
- Could require too much extra resource outside normal workflow if not managed
- Are there any curatorial / archival issues? Eg. authenticity?
- Need to formalise the process and clarify all details with website owners

How far do we go?

