Workshop B: Archiving the Web

Michael Day and Maureen Pennock
Digital Curation Centre
UKOLN, University of Bath
http://www.ukoln.ac.uk/

Driving the Long-Term Preservation of Electronic Records, London, 26-28 September 2006

DCC a centre of expertise in data curation and preservation

Workshop outline

- Session 1: The context of Web archiving - Michael Day (30 minutes)
- Session 2: Archival perspectives on Web archiving - Maureen Pennock (45 minutes)
- Coffee break (15 minutes)
- Session 3: Web archiving in practice - Michael Day (45 minutes)
- Session 4: Looking to the future - Michael Day (30 minutes)

Workshop: Archiving the Web, 28 September 2006

Session 1: The context of Web archiving

Michael Day

DCC a centre of expertise in data curation and preservation

The World Wide Web (1)

- Origins in scientific community
  - CERN (early 1990s)
  - Now part of the common 'cyberinfrastructure' of science and scholarship
  - Scientists 'increasingly reliant' on Web for supporting research activities (James Hendler, 2003)
  - Helps to promotes 'open access' principles (peer-reviewed publications, data resulting from publicly-funded research)
  - Other educational roles - e.g., e-learning

Workshop: Archiving the Web, 28 September 2006

The World Wide Web (2)

- Scholarly concern with the longevity of Internet references
  - Link rot problem
  - A study of three leading peer-reviewed journals showed that 13 percent of links were inactive after 3 years (Dellavalle, *et al*., 2003)
  - Same trends demonstrated in biomedicine, computer science, information science, …
  - Wallace Koehler's longitudinal studies show that after seven years, just 33.8 percent of a sample of Web pages persisted at their original URL

DCC a centre of expertise in data curation and preservation

The World Wide Web (3)

- The Web now widely used across many different communities:
  - Commerce, marketing, publishing
  - Government information (e-government)
  - Personal communication
    - e.g., 44 percent of US Internet users in a 2003 survey had contributed some kind of content to the Internet
  - "The information source of first resort for millions of readers" - Peter Lyman (2002)

Workshop: Archiving the Web, 28 September 2006

Why preserve the Web? (1)

- Cultural importance
  - National Library of Australia noted its responsibility to develop collections of library materials, *regardless of format*
  - Many national libraries have now developed operational or pilot Web archives, e.g.
    - Australia, Austria, China, Czech Republic, Denmark, Finland, France, Iceland, Japan, New Zealand, Norway, Slovenia, UK, USA, etc.
  - Some have made changes to legal deposit laws to accommodate Web content

DCC a centre of expertise in data curation and preservation

Why preserve the Web (2)

- Cultural importance
  - Internet Archive
    - not-for-profit organisation, based in San Fransciso
    - Acquired Web content from Alexa Internet and its own Web crawls, provides access through the Wayback Machine (http://www.archive.org/)
    - Co-operates with memory institutions on developing special collections, e.g. Library of Congress, The National Archives (UK)
    - Part of International Internet Preservation Coalition
    - Mirror of Wayback Machine at Bibliotheca Alexandrina (Egypt)

Workshop: Archiving the Web, 28 September 2006

Why preserve the Web? (3)

- Web content are records of evidence
  - National archives guidance for Web managers
  - Some collection of Web sites has started
    - The National Archives UK Government Web Archive, joint project with Internet Archive
    - US National Archives and Records Administration collected snapshot of federal agency Web sites at end of the Clinton Administration
- Scholarly interest
  - Politics (Archipol), social history (Occasio), Chinese studies (DACHS)

DCC a centre of expertise in data curation and preservation

Why preserve the Web? (4)

- Joint approaches
  - The UK Web Archiving Consortium
    - Led by the British Library
    - Partners include The National Archives, the national libraries of Wales and Scotland, the Joint Information Systems Committee, and the Wellcome Trust
    - Sharing costs, risks and experiences
    - Each partner focuses on sites relevant to their own interests

Approaches (1)

- Automatic harvesting
  - Web crawler programs
  - National libraries tend to focus on national Web domains, e.g. Kulturarw^3 (Sweden)
  - Harvester fed set of links, pages fetched, analysed, etc., etc.
  - Internet Archive uses same approach for whole Web, since 1996 has generated ~2 petabytes
    - Problems with functionality and country representation (but still a very valuable resource)
  - Development of Heritrix crawler program

DCC a centre of expertise in data curation and preservation

Approaches (2)

- Selective capture or deposit
  - Pioneered by National Library of Australia (PANDORA)
  - Development of selection guidelines, selection of sites, negotiation with site owners, capture using gathering or mirroring tools
  - Used by UK Web Archiving Consortium
  - Sites can also be captured and deposited by Web site owners
    - e.g., NARA 2001

Workshop: Archiving the Web, 28 September 2006

Approaches (3)

- Combined approaches
  - Some selective capture, periodic whole domain harvesting
  - Reflects relative strengths of the two approaches
    - Harvesting approach much cheaper per terabyte, enables large collections to be built up
    - More detailed attention can be paid to complex sites, e.g. database driven (deep Web) sites
  - Approach pioneered by Bibliothèque nationale de France (BnF)
  - Recent Australian whole domain harvest

DCC a centre of expertise in data curation and preservation

Approaches (4)

- International Internet Preservation Consortium (IIPC)
  - Group of national libraries and the Internet Archive, led by BnF
  - Co-operation on coverage and access - a global distributed collection
  - Development of tools
    - Harvesting - Heritrix, DeepArc
    - Storage - ARC, BAT
    - Search and navigation - NutchWAX, WERA, Zinq
    - Web Archiving Metadata Set

Issues (1)

- What is the Web?
  - A conceptual problem
  - Components of the Web easier to understand than the whole
  - What is is that we want to preserve?
    - Content? - easy for HTML pages, more difficult for databases (or database-driven sites)
    - Interfaces?
  - Personalisation features
  - Web 2.0

DCC a centre of expertise in data curation and preservation

Issues (2)

- Legal problems
  - Legal environment in many countries does not take Web archives into account (Charlesworth, 2003)
  - Problems with:
    - Copyright
    - Archives could be deemed to be the "publishers" of defamatory or otherwise illegal content, or held responsible for breaches of data protection legislation
  - Remedies = select content or restrict access

Workshop: Archiving the Web, 28 September 2006

Issues (3)

- Scale
  - Web is large (and growing)
  - Regular snapshots grow even bigger
  - Internet Archive: almost 2 petabytes, growing at >20 terabytes a month
  - Differences in Web archive size depending on domain:
    - Finland (2002) 500 gigabytes
    - Portugal (2003) 78 gigabytes
    - Australia (2005) 6.69 terabytes

DCC a centre of expertise in data curation and preservation

Issues (4)

- Dynamic nature of the Web
  - Pages, sites, domains, constantly changing
    - e.g. new top level domains
    - Web content disappearing (link rot)
  - Some *ad hoc* focus on the ephemeral
    - Political elections, sports events, 9/11, Hurricanes Katrina and Rita
  - Changes in Web technologies
    - Personalised delivery of content
    - Increased interactivity, Web 2.0, etc.

Workshop: Archiving the Web, 28 September 2006

Issues (5)

- Access
  - Problem of linking content stored in multiple, distributed archives
  - Need for co-operation
  - A role for International Internet Preservation Consortium?
- Digital preservation and curation
  - What this might mean for the Web has not been explored in detail
  - Web archives need to fit into the wider landscape of digital preservation and curation initiatives

DCC a centre of expertise in data curation and preservation

Initial conclusions

- The Web is culturally important
- To date, Web archiving initiatives have collected a significant amount of content
- Different capture techniques compliment each other
- There has been a major improvement in the tools being used to harvest and manage content, e.g. the IIPC toolkit
- Co-operation - the IIPC provides one venue for this. Are others needed?
- Many significant issues remain to be solved

Session 2: Archival perspectives on Web archiving

Maureen Pennock

[separate presentation]

DCC a centre of expertise in data curation and preservation

Session 3: Web archiving in practice

Michael Day

Contexts

- National and research libraries
  - National domains
  - Special collections
- National archives
  - Snapshots of government Web sites

DCC a centre of expertise in data curation and preservation

Selection (1)

- Develop selection policy
  - Exact criteria will depend on the purpose of the Web archive
  - National libraries will tend to focus on their role as the custodian of the nation's documentary heritage, e.g.
    - National Library of Australia
      - Selected content needs to be relevant to Australia (or written by an Australian)
      - But there is a higher degree of selectivity than in the traditional environment
      - Boundaries of document type are not so clear cut
    - Other partners in PANDORA focus on specific content
      - States, film and music, war

Workshop: Archiving the Web, 28 September 2006

Selection (2)

- UK Web Archiving Consortium
  - Different member organisations focus on different content types, e.g.:
    - Medical sites (Wellcome Library), project web sites (JISC), Wales (National Library of Wales), …
- Archives will focus on the role of Web sites as records, e.g.
  - Recording interactions between state and citizen (e-Government)

- Frequency
  - Decisions also need to be made on the frequency of capture
  - The National Archives (UK) collects some sites weekly, others biannually

DCC a centre of expertise in data curation and preservation

Collection and ingest (1)

- Collection methods

	Content-driven	Event-driven
Client-side	Remote harvesting	
Server-side	Direct transfer Database archiving	Transactional archiving

- *Source*: Adrian Brown (TNA): http://www.dcc.ac.uk/events/fpw-2006/

Workshop: Archiving the Web, 28 September 2006

Collection and ingest (2)

- Direct transfer
  - Examples:
    - NARA snapshots at the end of the Clinton Administration (2001)
    - 10 Downing Street site (2001 General Election)
  - Can be problematic, effectively a migration to a different technical environment
- Database archiving
  - IIPC tool developed for capture of the deep Web (DeepARC)
  - Non trivial task, mapping relational DBs into XML schema, migrating content into an XML document

DCC a centre of expertise in data curation and preservation

Collection and ingest (3)

- Remote harvesting
  - The most commonly used capture method
  - Uses crawler programs similar to those used by search engines
    - To date, various crawler programs have been developed (or adapted)
    - The Internet Archive has led the development of a crawler program focused on the capture of Web content (Heritrix)
  - Collection can be focused at different levels
    - Domain capture (national domain defined in various ways), used by some national libraries
    - Focused collections, capture of selected sites

UKOLN
UNIVERSITY OF BATH
Workshop: Archiving the Web, 28 September 2006

Collection and ingest (4)

- Software available to manage the capture and ingest process
  - PANDAS (Pandora Digital Archiving System)
  - For setting up crawler programs, identifying base URLs, managing harvesting parameters (for selective approach)
  - Creation of metadata
- Limitations of the harvesting approach:
  - Does not deal effectively with database-driven sites (deep Web)
  - Little quality-control of content harvested

DCC a centre of expertise in data curation and preservation

Collection and ingest (5)

- Harvesting can also be contracted out:
  - Contracts with the Internet Archive/European Archive
    - The National Archives
      - » UK Government Web Archive
      - » Regular capture of selected government Web pages
    - Library of Congress, *et al*.
      - » September 11 Web Archive
      - » Hurricanes Katrina and Rita Web Archive

Workshop: Archiving the Web, 28 September 2006

Preservation and access (1)

- Preservation
  - Is about maintaining accessibility over time
  - About maintaining the authenticity of content (knowing that it is what it claims to be)
  - The 'significant properties' of objects are important
- Web archiving initiatives have, until now, mostly been about collecting content rather than preserving it
  - Reflects the rapidly changing nature of the Web
  - An essential first step
  - Preservation is a much harder issue to solve

DCC a centre of expertise in data curation and preservation

Preservation and access (2)

- Preservation involves
  - The development of a secure repository system
    - e.g., based on the Reference Model for an Open Archival Information System (ISO 14721:2003)
  - Good system administration
    - Access control, management of storage (media refreshment, backup and replication), disaster recovery
  - Activities specific to digital preservation:
    - Identifying the significant properties of objects
    - Identifying and implementing appropriate preservation strategies
    - Preservation planning (dealing with future uncertainty)

Workshop: Archiving the Web, 28 September 2006

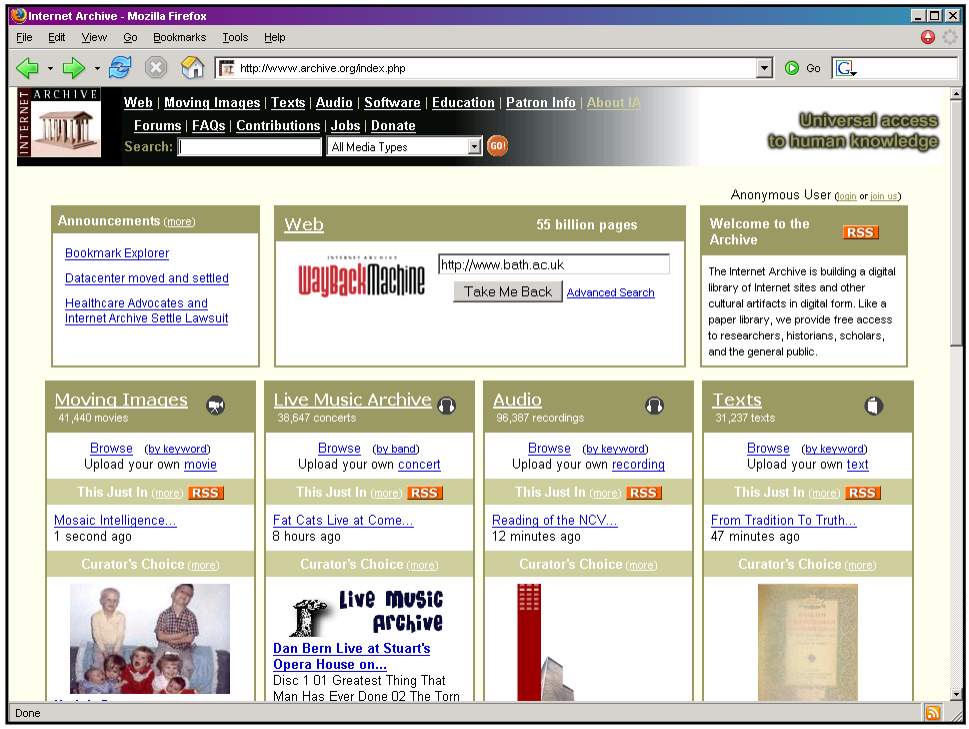
Preservation and access (3)

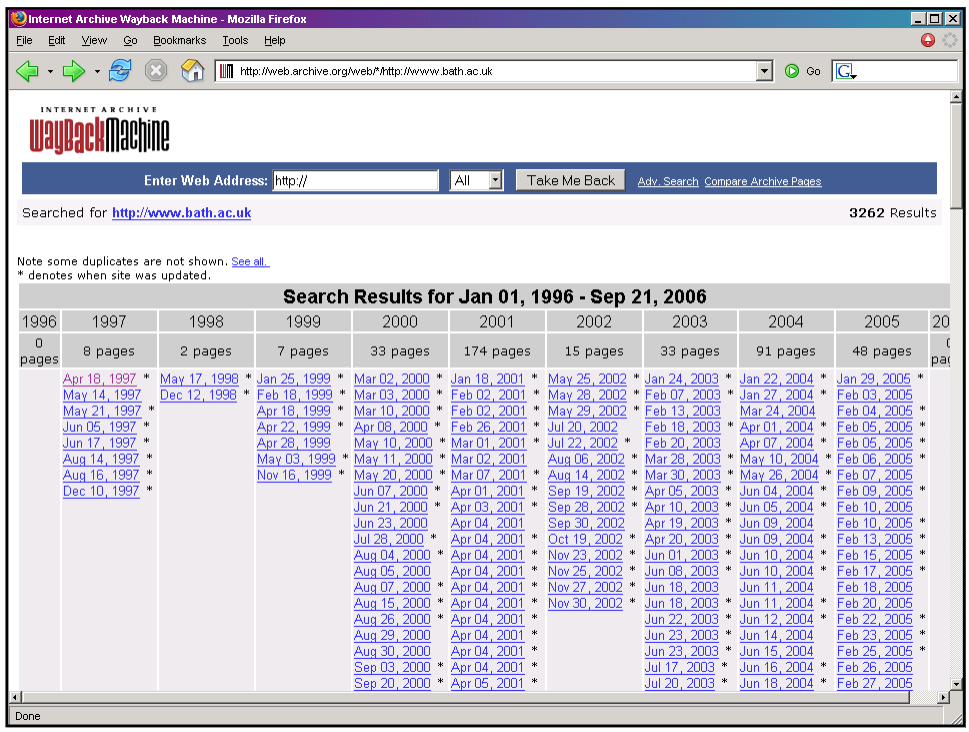
- Access
  - Many challenges (see IIPC Use Cases)
  - Legal reasons mean that many Web archiving initiatives do not provide significant end-user access
  - Especially true for domain harvesting initiatives (national libraries)
  - However, some selective initiatives already allow access to captured content:
    - UK Web Archiving Consortium
    - The Pandora Archive
  - As does:
    - The Internet Archive …

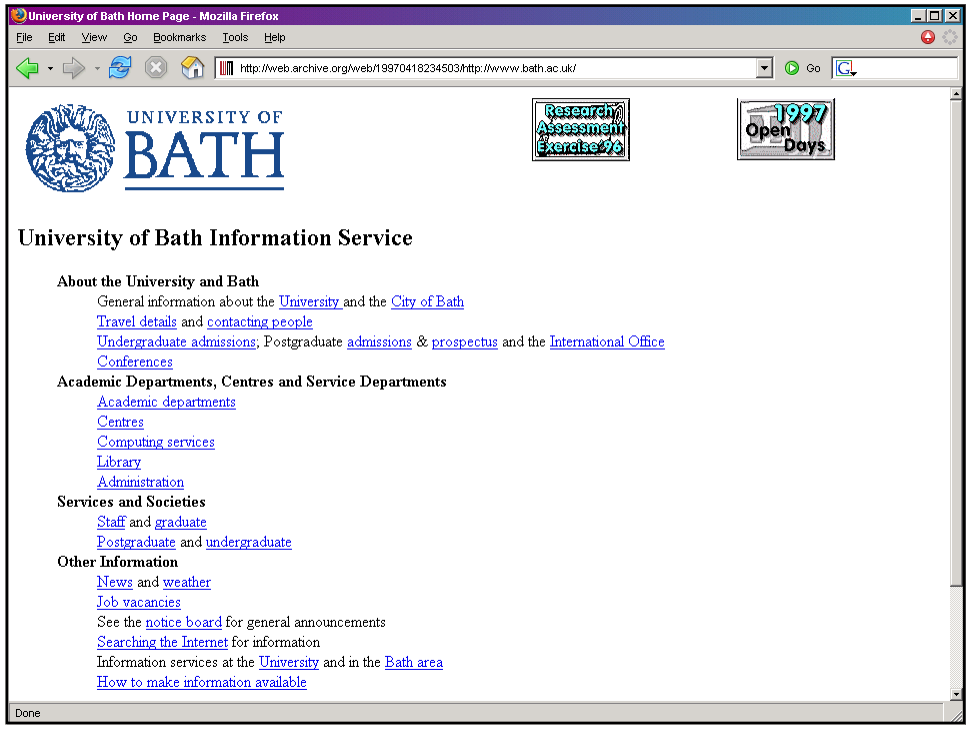
DCC a centre of expertise in data curation and preservation

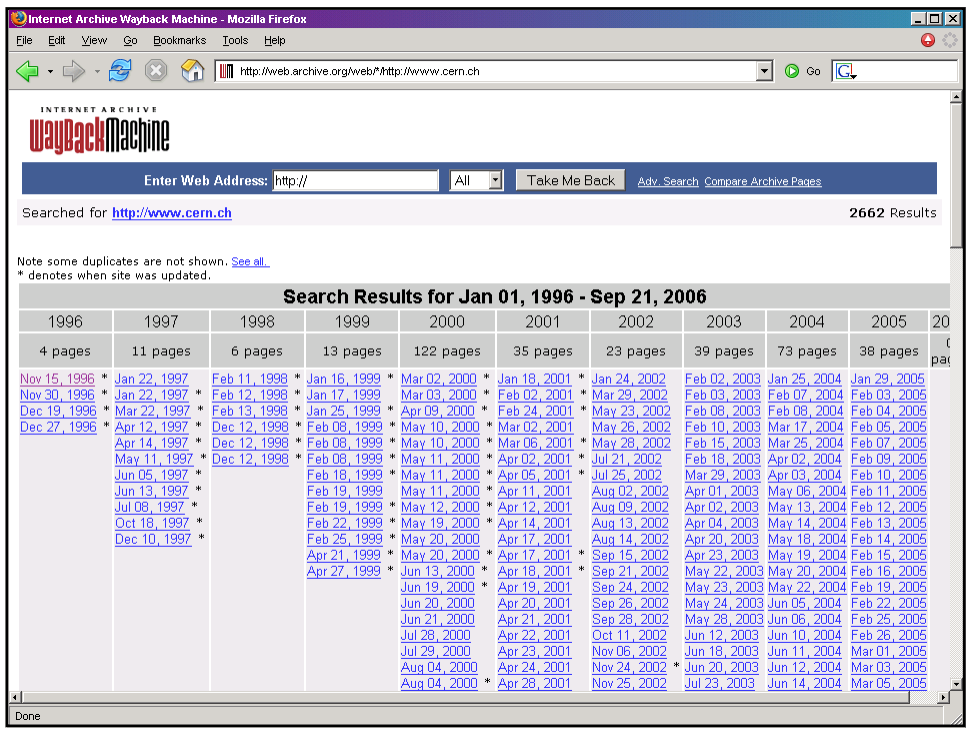
The Wayback Machine

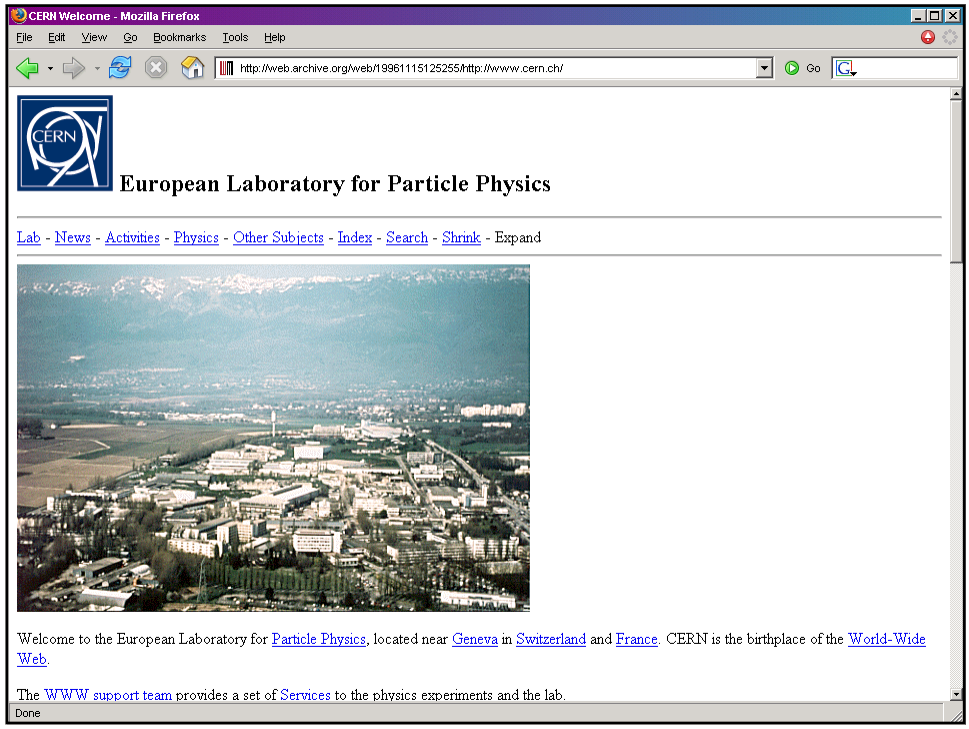
Workshop: Archiving the Web, 28 September 2006


Internet Archive - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.archive.org/index.php
Go
INTERNET ARCHIVE
Web | Moving Images | Texts | Audio | Software | Education | Patron Info | About IA
Forums | FAQs | Contributions | Jobs | Donate
Search:
All Media Types
GO!
Universal access to human knowledge
Anonymous User (login or join us)
Announcements (more)
Bookmark Explorer
Datacenter moved and settled
Healthcare Advocates and Internet Archive Settle Lawsuit
Web
55 billion pages
WayBackMachine
http://www.bath.ac.uk
Take Me Back
Advanced Search
Welcome to the Archive
RSS
The Internet Archive is building a digital library of Internet sites and other cultural artifacts in digital form. Like a paper library, we provide free access to researchers, historians, scholars, and the general public.
Moving Images
41,440 movies
Browse (by keyword)
Upload your own movie
This Just In (more) RSS
Mosaic Intelligence...
1 second ago
Curator's Choice (more)
Live Music Archive
38,647 concerts
Browse (by band)
Upload your own concert
This Just In (more) RSS
Fat Cats Live at Come...
8 hours ago
Curator's Choice (more)
Live Music Archive
Dan Bern Live at Stuart's Opera House on...
Disc 1 01 Greatest Thing That Man Has Ever Done 02 The Torn
Audio
96,387 recordings
Browse (by keyword)
Upload your own recording
This Just In (more) RSS
Reading of the NCV...
12 minutes ago
Curator's Choice (more)
Texts
31,237 texts
Browse (by keyword)
Upload your own text
This Just In (more) RSS
From Tradition To Truth...
47 minutes ago
Curator's Choice (more)
Done


Internet Archive Wayback Machine - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://web.archive.org/web/*/http://www.bath.ac.uk
Go
INTERNET ARCHIVE
WayBackMachine
Enter Web Address: http://
All
Take Me Back
Adv. Search Compare Archive Pages
Searched for http://www.bath.ac.uk
3262 Results
Note some duplicates are not shown. See all.
* denotes when site was updated.
Search Results for Jan 01, 1996 - Sep 21, 2006
1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005
0 pages | 8 pages | 2 pages | 7 pages | 33 pages | 174 pages | 15 pages | 33 pages | 91 pages | 48 pages
Apr 18, 1997 * May 14, 1997 May 21, 1997 * Jun 05, 1997 * Jun 17, 1997 * Aug 14, 1997 * Aug 16, 1997 * Dec 10, 1997 *
May 17, 1998 * Dec 12, 1998 *
Jan 25, 1999 * Feb 18, 1999 * Apr 18, 1999 * Apr 22, 1999 * Apr 28, 1999 May 03, 1999 * Nov 16, 1999 *
Mar 02, 2000 * Mar 03, 2000 * Mar 10, 2000 * Apr 08, 2000 * May 10, 2000 * May 11, 2000 * May 20, 2000 * Jun 07, 2000 * Jun 21, 2000 * Jun 23, 2000 Jul 26, 2000 * Aug 04, 2000 * Aug 05, 2000 Aug 07, 2000 * Aug 15, 2000 * Aug 26, 2000 * Aug 29, 2000 Aug 30, 2000 Sep 03, 2000 * Sep 20, 2000 *
Jan 18, 2001 * Feb 02, 2001 * Feb 02, 2001 * Feb 26, 2001 * Mar 01, 2001 * Mar 02, 2001 Mar 07, 2001 * Apr 01, 2001 * Apr 03, 2001 * Apr 04, 2001 Apr 04, 2001 * Apr 04, 2001 * Apr 04, 2001 * Apr 04, 2001 * Apr 04, 2001 * Apr 04, 2001 * Apr 04, 2001 * Apr 04, 2001 * Apr 04, 2001 * Apr 05, 2001 *
May 25, 2002 * May 28, 2002 * May 29, 2002 * Jul 20, 2002 Jul 22, 2002 * Aug 06, 2002 * Aug 14, 2002 * Sep 19, 2002 * Sep 28, 2002 * Sep 30, 2002 Oct 19, 2002 * Nov 23, 2002 * Nov 25, 2002 * Nov 27, 2002 * Nov 30, 2002 *
Jan 24, 2003 * Feb 07, 2003 * Feb 13, 2003 Feb 18, 2003 * Feb 20, 2003 Mar 28, 2003 * Mar 30, 2003 * Apr 05, 2003 * Apr 10, 2003 * Apr 19, 2003 * Apr 20, 2003 * Jun 01, 2003 * Jun 08, 2003 * Jun 18, 2003 Jun 18, 2003 * Jun 22, 2003 * Jun 23, 2003 Jun 23, 2003 * Jul 17, 2003 * Jul 20, 2003 *
Jan 22, 2004 * Jan 27, 2004 * Mar 24, 2004 Apr 01, 2004 * Apr 07, 2004 * May 10, 2004 * May 26, 2004 * Jun 04, 2004 * Jun 05, 2004 * Jun 09, 2004 Jun 09, 2004 * Jun 10, 2004 * Jun 10, 2004 * Jun 11, 2004 Jun 11, 2004 * Jun 12, 2004 * Jun 14, 2004 Jun 15, 2004 Jun 16, 2004 * Jun 18, 2004 *
Jan 29, 2005 * Feb 03, 2005 Feb 04, 2005 * Feb 05, 2005 * Feb 05, 2005 * Feb 06, 2005 * Feb 07, 2005 Feb 09, 2005 * Feb 10, 2005 Feb 10, 2005 * Feb 13, 2005 * Feb 15, 2005 * Feb 17, 2005 * Feb 18, 2005 Feb 20, 2005 Feb 22, 2005 * Feb 23, 2005 * Feb 25, 2005 * Feb 26, 2005 Feb 27, 2005
Done


University of Bath Home Page - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://web.archive.org/web/19970418234503/http://www.bath.ac.uk/
Go
UNIVERSITY OF BATH
Research Assessment Exercise 96
1997 Open Days
University of Bath Information Service
About the University and Bath
General information about the University and the City of Bath
Travel details and contacting people
Undergraduate admissions; Postgraduate admissions & prospectus and the International Office
Conferences
Academic Departments, Centres and Service Departments
Academic departments
Centres
Computing services
Library
Administration
Services and Societies
Staff and graduate
Postgraduate and undergraduate
Other Information
News and weather
Job vacancies
See the notice board for general announcements
Searching the Internet for information
Information services at the University and in the Bath area
How to make information available
Done


Internet Archive Wayback Machine - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://web.archive.org/web/*/http://www.cern.ch
Go
INTERNET ARCHIVE
WayBackMachine
Enter Web Address: http://
All
Take Me Back
Adv. Search Compare Archive Pages
Searched for http://www.cern.ch
2662 Results
Note some duplicates are not shown. See all.
* denotes when site was updated.
Search Results for Jan 01, 1996 - Sep 21, 2006
1996 1997 1998 1999 2000 2001 2002 2003 2004 2005
4 pages 11 pages 6 pages 13 pages 122 pages 35 pages 23 pages 39 pages 73 pages 38 pages
Done



DCC

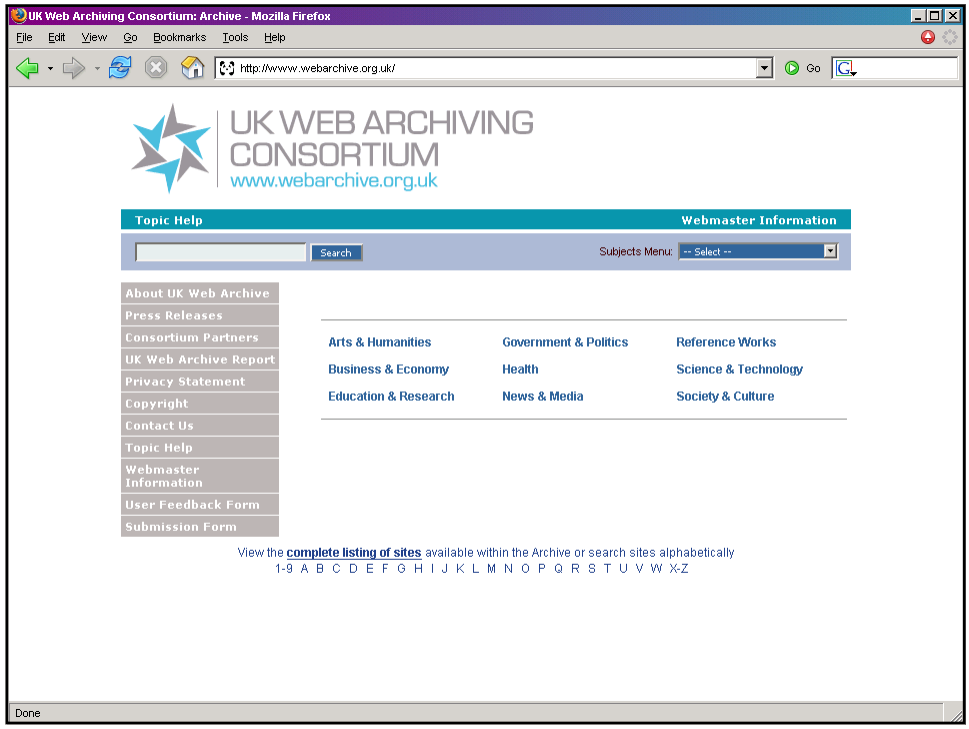
a centre of expertise in data curation and preservation

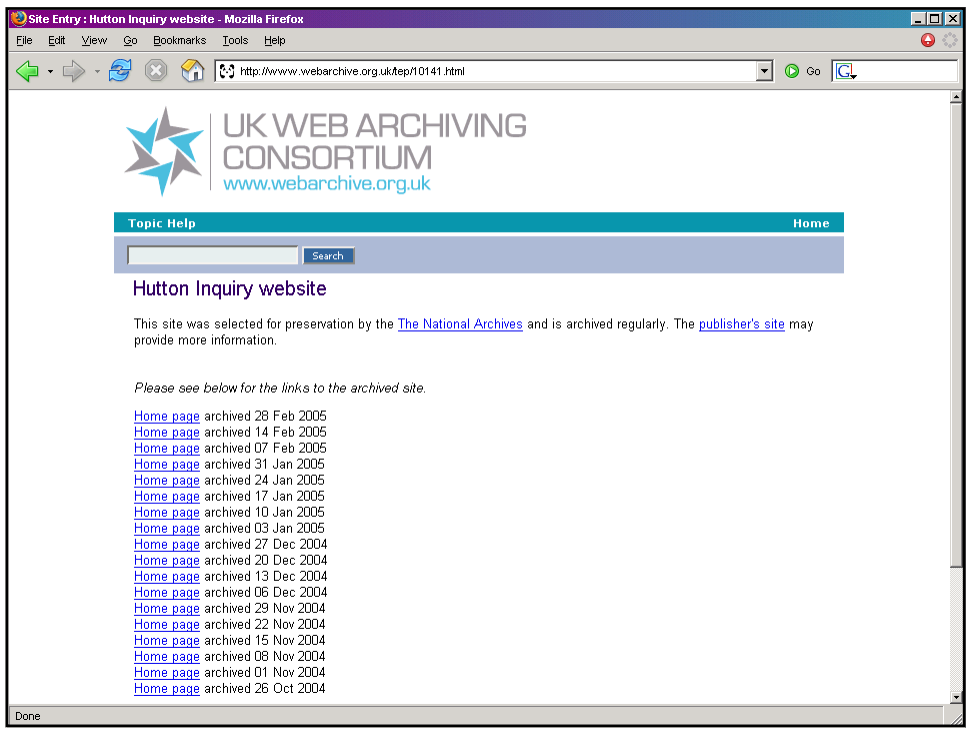
UK Web Archiving Consortium

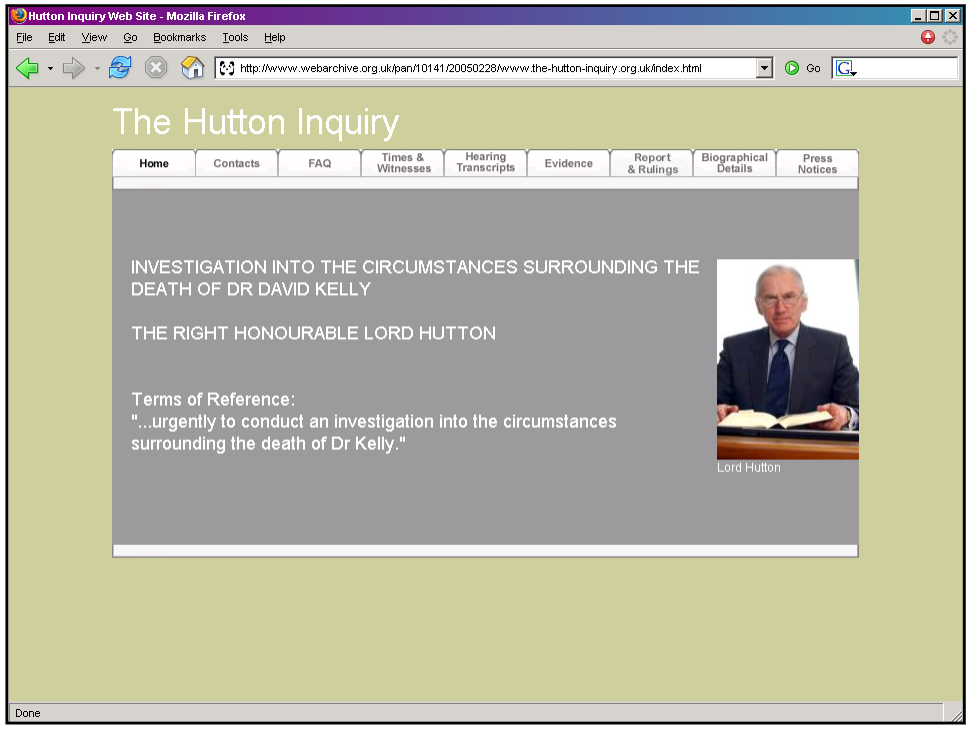
UKOLN

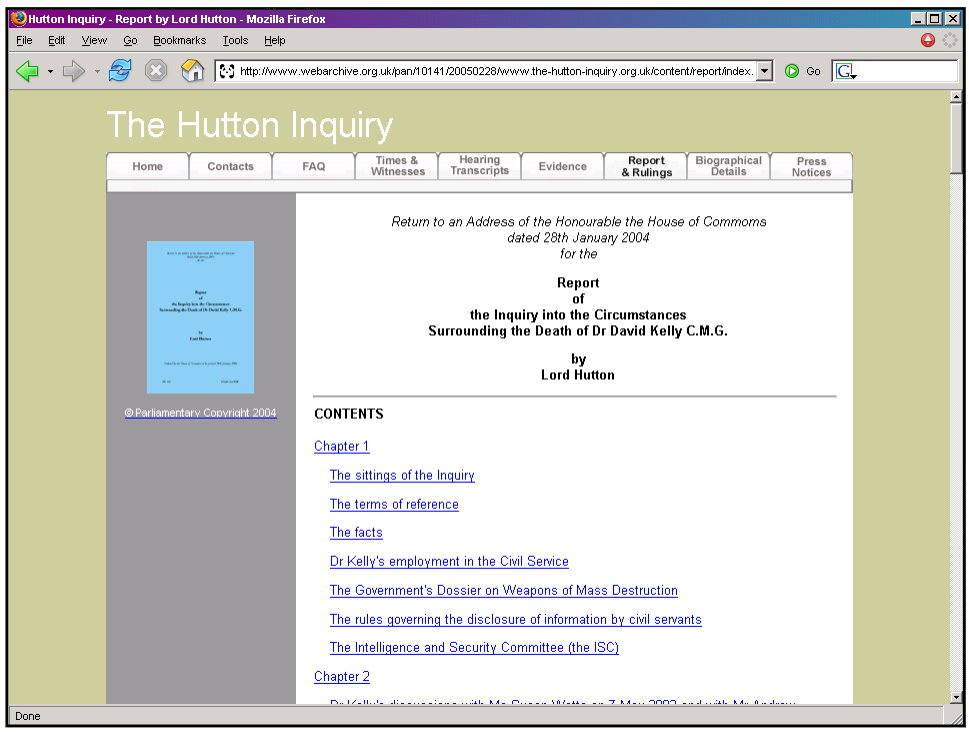
UNIVERSITY OF BATH

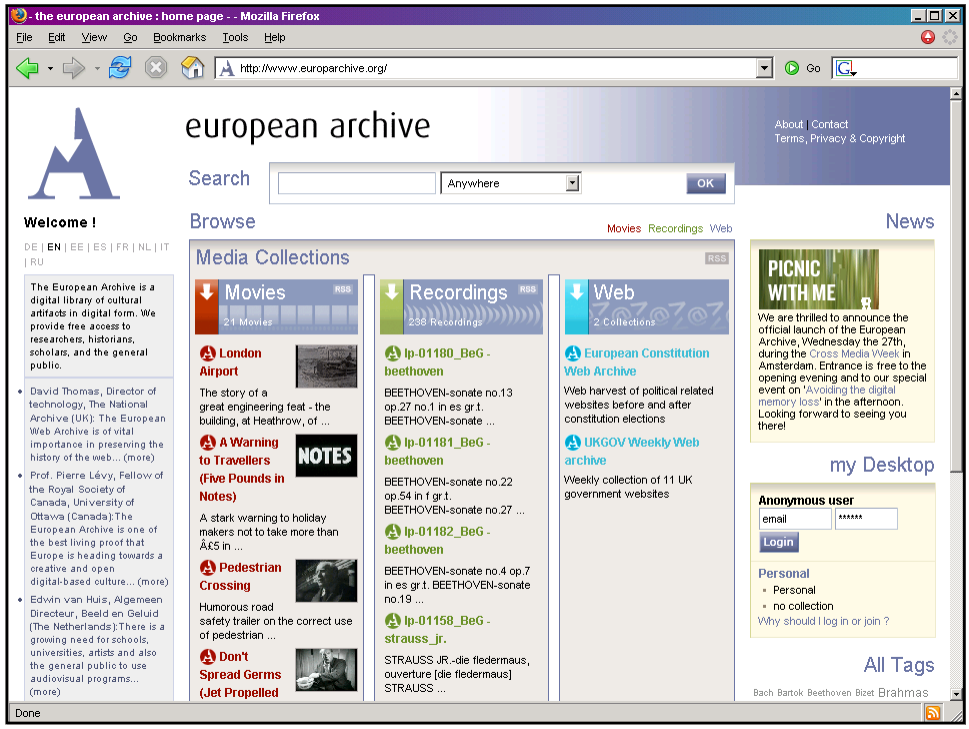
Workshop: Archiving the Web, 28 September 2006

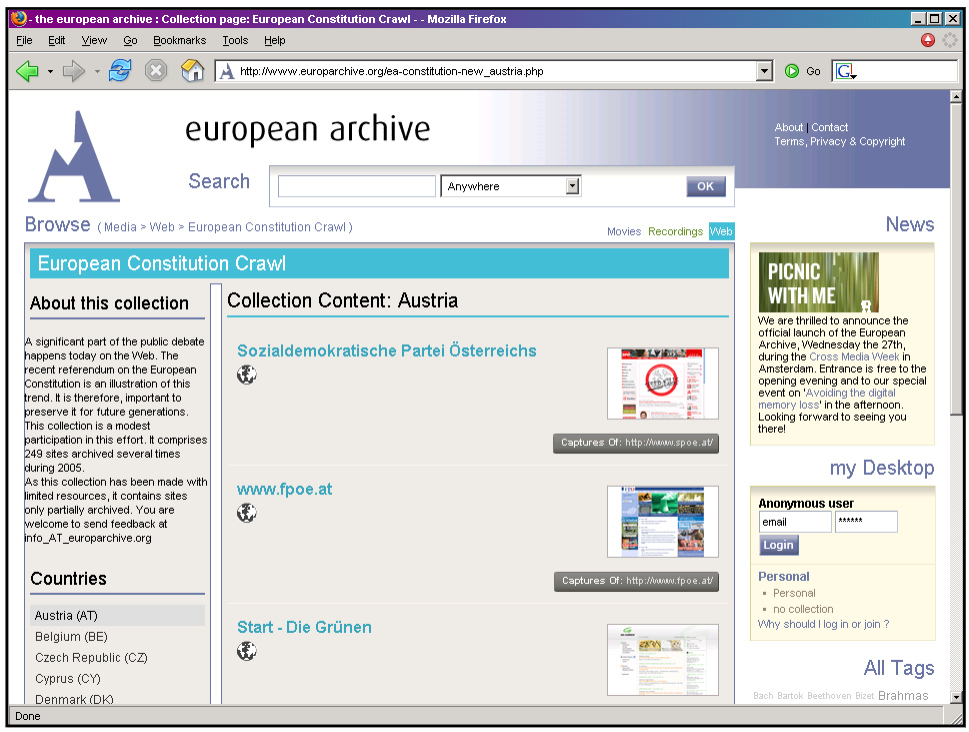

UK Web Archiving Consortium: Archive - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.webarchive.org.uk/
Go
UK WEB ARCHIVING CONSORTIUM
www.webarchive.org.uk
Topic Help
Webmaster Information
Search
Subjects Menu:
-- Select --
About UK Web Archive
Press Releases
Consortium Partners
UK Web Archive Report
Privacy Statement
Copyright
Contact Us
Topic Help
Webmaster Information
User Feedback Form
Submission Form
Arts & Humanities
Government & Politics
Reference Works
Business & Economy
Health
Science & Technology
Education & Research
News & Media
Society & Culture
View the complete listing of sites available within the Archive or search sites alphabetically
1-9 A B C D E F G H I J K L M N O P Q R S T U V W X-Z
Done

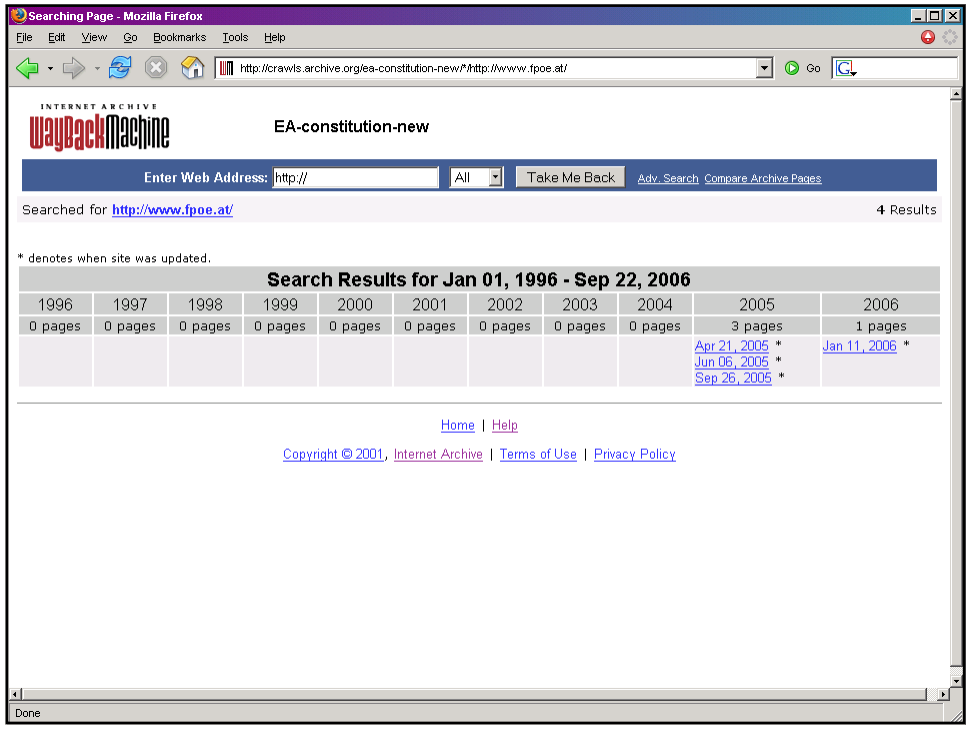

Site Entry : Hutton Inquiry website - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.webarchive.org.uk/tep/10141.html
Go
UK WEB ARCHIVING CONSORTIUM
www.webarchive.org.uk
Topic Help
Home
Search
Hutton Inquiry website
This site was selected for preservation by the The National Archives and is archived regularly. The publisher's site may provide more information.
Please see below for the links to the archived site.
Home page archived 28 Feb 2005
Home page archived 14 Feb 2005
Home page archived 07 Feb 2005
Home page archived 31 Jan 2005
Home page archived 24 Jan 2005
Home page archived 17 Jan 2005
Home page archived 10 Jan 2005
Home page archived 03 Jan 2005
Home page archived 27 Dec 2004
Home page archived 20 Dec 2004
Home page archived 13 Dec 2004
Home page archived 06 Dec 2004
Home page archived 29 Nov 2004
Home page archived 22 Nov 2004
Home page archived 15 Nov 2004
Home page archived 08 Nov 2004
Home page archived 01 Nov 2004
Home page archived 26 Oct 2004
Done


Hutton Inquiry Web Site - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.webarchive.org.uk/pan/10141/20050228/www.the-hutton-inquiry.org.uk/index.html
Go
The Hutton Inquiry
Home
Contacts
FAQ
Times & Witnesses
Hearing Transcripts
Evidence
Report & Rulings
Biographical Details
Press Notices
INVESTIGATION INTO THE CIRCUMSTANCES SURROUNDING THE DEATH OF DR DAVID KELLY
THE RIGHT HONOURABLE LORD HUTTON
Terms of Reference:
"...urgently to conduct an investigation into the circumstances surrounding the death of Dr Kelly."
Lord Hutton
Done


Hutton Inquiry - Report by Lord Hutton - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.webarchive.org.uk/pan/10141/20050228/www.the-hutton-inquiry.org.uk/content/report/index.
Go
The Hutton Inquiry
Home
Contacts
FAQ
Times & Witnesses
Hearing Transcripts
Evidence
Report & Rulings
Biographical Details
Press Notices
© Parliamentary Copyright 2004
Return to an Address of the Honourable the House of Commoms
dated 28th January 2004
for the
Report
of
the Inquiry into the Circumstances
Surrounding the Death of Dr David Kelly C.M.G.
by
Lord Hutton
CONTENTS
Chapter 1
The sittings of the Inquiry
The terms of reference
The facts
Dr Kelly's employment in the Civil Service
The Government's Dossier on Weapons of Mass Destruction
The rules governing the disclosure of information by civil servants
The Intelligence and Security Committee (the ISC)
Chapter 2
Done


- the european archive : home page - - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.europarchive.org/
Go
european archive
About | Contact
Terms, Privacy & Copyright
Search
Anywhere
OK
Welcome !
DE | EN | EE | ES | FR | NL | IT | RU
The European Archive is a digital library of cultural artifacts in digital form. We provide free access to researchers, historians, scholars, and the general public.
David Thomas, Director of technology, The National Archive (UK): The European Web Archive is of vital importance in preserving the history of the web... (more)
Prof. Pierre Lévy, Fellow of the Royal Society of Canada, University of Ottawa (Canada):The European Archive is one of the best living proof that Europe is heading towards a creative and open digital-based culture... (more)
Edwin van Huis, Algemeen Directeur, Beeld en Geluid (The Netherlands):There is a growing need for schools, universities, artists and also the general public to use audiovisual programs... (more)
Browse
Movies Recordings Web
Media Collections
Movies
21 Movies
London Airport
The story of a great engineering feat - the building, at Heathrow, of ...
A Warning to Travellers (Five Pounds in Notes)
A stark warning to holiday makers not to take more than Â£5 in ...
Pedestrian Crossing
Humorous road safety trailer on the correct use of pedestrian ...
Don't Spread Germs (Jet Propelled
Recordings
298 Recordings
lp-01180_BeG - beethoven
BEETHOVEN-sonate no.13 op.27 no.1 in es gr.t. BEETHOVEN-sonate ...
lp-01181_BeG - beethoven
BEETHOVEN-sonate no.22 op.54 in f gr.t. BEETHOVEN-sonate no.27 ...
lp-01182_BeG - beethoven
BEETHOVEN-sonate no.4 op.7 in es gr.t. BEETHOVEN-sonate no.19 ...
lp-01158_BeG - strauss_jr.
STRAUSS JR.-die fledermaus, ouverture [die fledermaus] STRAUSS ...
Web
2 Collections
European Constitution Web Archive
Web harvest of political related websites before and after constitution elections
UKGOV Weekly Web archive
Weekly collection of 11 UK government websites
News
PICNIC WITH ME
We are thrilled to announce the official launch of the European Archive, Wednesday the 27th, during the Cross Media Week in Amsterdam. Entrance is free to the opening evening and to our special event on 'Avoiding the digital memory loss' in the afternoon. Looking forward to seeing you there!
my Desktop
Anonymous user
email
Login
Personal
Personal
no collection
Why should I log in or join ?
All Tags
Bach Bartok Beethoven Bizet Brahms
Done


- the european archive : Collection page: European Constitution Crawl - - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.europarchive.org/ea-constitution-new_austria.php
Go
european archive
About | Contact
Terms, Privacy & Copyright
Search
Anywhere
OK
Browse (Media » Web » European Constitution Crawl)
Movies Recordings Web
European Constitution Crawl
About this collection
A significant part of the public debate happens today on the Web. The recent referendum on the European Constitution is an illustration of this trend. It is therefore, important to preserve it for future generations. This collection is a modest participation in this effort. It comprises 249 sites archived several times during 2005.
As this collection has been made with limited resources, it contains sites only partially archived. You are welcome to send feedback at info_AT_europarchive.org
Countries
Austria (AT)
Belgium (BE)
Czech Republic (CZ)
Cyprus (CY)
Denmark (DK)
Collection Content: Austria
Sozialdemokratische Partei Österreichs
Captures Of: http://www.spoe.at/
www.fpoe.at
Captures Of: http://www.fpoe.at/
Start - Die Grünen
News
PICNIC WITH ME
We are thrilled to announce the official launch of the European Archive, Wednesday the 27th, during the Cross Media Week in Amsterdam. Entrance is free to the opening evening and to our special event on 'Avoiding the digital memory loss' in the afternoon. Looking forward to seeing you there!
my Desktop
Anonymous user
email
Login
Personal
Personal
no collection
Why should I log in or join ?
All Tags
Bach Bartok Beethoven Bizet Brahms
Done


Searching Page - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://crawls.archive.org/ea-constitution-new/*/http://www.fpoe.at/
Go
INTERNET ARCHIVE
WayBackMachine
EA-constitution-new
Enter Web Address: http://
All
Take Me Back
Adv. Search
Compare Archive Pages
Searched for http://www.fpoe.at/
4 Results
* denotes when site was updated.
Search Results for Jan 01, 1996 - Sep 22, 2006
1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006
0 pages 0 pages 0 pages 0 pages 0 pages 0 pages 0 pages 0 pages 0 pages 3 pages 1 pages
Apr 21, 2005 *
Jun 06, 2005 *
Sep 26, 2005 *
Jan 11, 2006 *
Home | Help
Copyright © 2001, Internet Archive | Terms of Use | Privacy Policy
Done


www.fpoe.at - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://crawls.archive.org/ea-constitution-new/20050421135430/http://www.fpoe.at/
Go
FPÖ
KONTAKT | LOGIN
DIE FPÖ HC | STRACHE DAFÜR STEHEN WIR PRESSESERVICE INTERAKTIV
Die FPÖ Landesgruppen
Bundesland:
Heimniederlage für Haider: Kärntner bevorzugen die Blauen
RH-Bericht: Eurofighter nur ein teurer Schönwetterflieger
Asylgesetz weich: Prokop will FPÖ-Forderungen streichen
LANDESPARTEITAG FPÖ-WIEN
FOTOALBUM
20. April 2005
Strache für Eurofighter-Untersuchungsausschuss
Vernichtender RH-Bericht zu Eurofighter-Kauf
Heimniederlage für Haider
19. April 2005
LPO Werner Neubauer: "Steinkellner wird sich Recht beugen müssen!"
18. April 2005
Rosenkranz: Keine Verpflichtungen mehr Regierungspolitik mitzutragen
Nittmann: BZÖ vierter ÖVP-Bund
Kabas schließt OÖ-Obmann Steinkellner aus
Schnell: Haiders Zick-Zack-Kurs ist jetzt Problem von Schüssel
BZÖ?
Riesen Kluft zwischen Anspruch und Wirklichkeit.
VOLLTEXTSUCHE
http://crawls.archive.org/ea-constitution-new/20050421135430/http://www.fpoe.at/index.php?id=477&backPID=390&tt_news=3530

Session 4: Looking to the future

Michael Day

DCC a centre of expertise in data curation and preservation

Legal issues (1)

- General observations
  - I am not a lawyer!
  - There is much legal uncertainty in the digital domain, not least about jurisdiction
- Intellectual property
  - Copyright regimes getting more stringent (e.g., DCMA)
  - Rights holders more determined to protect IPR
  - This is the reason why the UK Web Archiving Consortium negotiates deposit of content with rights holders
    - But it can still be difficult to identify who holds the rights in multi-partner project Web sites

Workshop: Archiving the Web, 28 September 2006

Legal issues (2)

- Content liability:
  - In the UK, providing access to a preserved Web site counts as "publication," raising the issue of content liability for:
    - Defamation
      - Most UK case law relates to the role of ISPs, but Web archives would seem to be liable if defamatory content is "republished"
    - Data Protection
      - Where Web pages might contain personal information, Web archives need to comply with DP legislation

DCC a centre of expertise in data curation and preservation

Legal issues (3)

- Content liability (continued)
  - Illegal content
    - Some types of pornography, Holocaust denial
    - Wide variance internationally, but care still needs to be taken
- If you are thinking about doing Web archiving, you will at some point need to consider legal issues, even if only to dismiss them!

Workshop: Archiving the Web, 28 September 2006

Future proofing your web site (1)

- Some general principles
  - From John Kunze (California Digital Library)
  - 3 Rs
    - Reduce dependencies
    - Redirect URLs
    - Replicate
  - Prioritise
    - Focus on that content that is most important (or may contain essential business records)
  - Look for simple solutions
    - Focus on the things that may have the widest impact

DCC a centre of expertise in data curation and preservation

Future proofing your Web site (2)

- Basics:
  - Develop a *strategy* for managing Web sites over the short to medium term
    - Plan for the future, try to obtain sufficient funding
  - Maintain domain names
    - Expired names can be reused by Web site pirates
    - This can cause severe embarrassment
  - Where possible, use standards
    - Validate standards
      - Some tools exist to do this (e.g. for X/HTML)
    - Open standards are better than proprietary formats
    - Avoid browser-specific features

UNIVERSITY OF BATH
Workshop: Archiving the Web, 28 September 2006

Future proofing your Web site (3)

- If there is no possibility of maintaining the pages yourself:
  - Record the fact that the pages are no longer being updated
  - If necessary, hand over the site to be managed by someone else
    - A role for third party hosting services? National Libraries? The UK Web Archiving Consortium?
  - This is not just a problem for organisations, personal (or hobby) sites are probably even worse off …

DCC a centre of expertise in data curation and preservation

Conclusions

- The Web is culturally important [and also contains records]
- To date, Web archiving initiatives have collected a significant amount of content [and this is growing rapidly]
- Different capture techniques compliment each other [but significant progress has been on the development of models for selection and ingest]
- There has been a major improvement in the tools being used to harvest and manage content, e.g. the IIPC toolkit [this work continues]
- Co-operation - the IIPC provides one venue for this. Are others needed? [Web archiving is one aspect of a much wider digital preservation problem]
- Many significant issues remain to be solved

Workshop: Archiving the Web, 28 September 2006

Further reading

- **Adrian Brown, *Archiving Websites: a practical guide for information management professionals* (Facet, 2006)**
- **Julien Masanès, ed., *Web archiving* (Springer, 2006)**
- **Michael Day, *Collecting and preserving the World Wide Web* (JISC, 2003): http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf**
- **Andrew Charlesworth, Legal issues relating to the archiving of Internet resources … (JISC, 2003): http://www.jisc.ac.uk/uploaded_documents/archiving_legal.pdf**
- **UK Web Archiving Consortium: http://www.webarchive.org.uk/**
- **Internet Archive: http://www.archive.org/**
- **European Archive: http://www.europarchive.org/**
- **International Internet Preservation Consortium: http://netpreserve.org/**

DCC a centre of expertise in data curation and preservation

Acknowledgements

UKOLN is funded by the Museums, Libraries and Archives Council, the Joint Information Systems Committee (JISC) of the UK higher and further education funding councils, as well as by project funding from the JISC, the European Union, and other sources. UKOLN also receives support from the University of Bath, where it is based.

http://www.ukoln.ac.uk/

The *Digital Curation Centre* is funded by the JISC and the UK Research Councils' e-Science Core Programme.

http://www.dcc.ac.uk/

Workshop: Archiving the Web, 28 September 2006