

Archiving The UK Domain And UK Web Sites

Brian Kelly

UK Web Focus

UKOLN

University of Bath

Email

B.Kelly@ukoln.ac.uk

URL

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

Topics

- The Extent Of UK Web Space
- Issues In Preserving Web Sites

UKOLN is supported by:

re:source The Council for
Museums
Archives
and Libraries



The Extent Of UK Web Space (1)

How big is UK Web space?

First thoughts:

- Why should this matter?

On reflection there is a need to:

- Estimate the nos. of Web servers
- Estimate the total size of Web sites
- Profile Web sites (e.g. proportion of dynamic or personalised Web sites)
- ...

in order to inform discussions on a Web preservation strategy

The Extent Of UK Web Space (2)

Second thoughts:

- Can this be measured?

Some would say:

- The Web is so complex that it is not currently sensible to talk about measuring the Web

Others would argue:

- Measuring the size of TV audiences (or the size of the Universe) is also difficult, but we do do this

Both camps would probably agree that measuring the Web is difficult, and that we are at the early stages of developing statistically valid methodologies for interpreting the figures

Web Estimates – The Difficulties

What Do We Mean By UK Web Space?

- Web sites under the '.uk' top-level domain (which will miss .org, .com, etc.)
- Web sites hosted on servers physically in the UK (which will miss UK Web sites hosted elsewhere)
- Web sites owned by UK citizens and/or organizations
- Web sites that host content published by UK citizens and/or organisations

<http://groups.yahoo.com/group/english-parliament-news/>

Mailing list archives for campaigners for an English parliament. Web site based in US, with a .com domain.

Web Estimates – The Difficulties

Estimating the size and extent of UK Web space poses some challenges.

How Do We Measure UK Web Space?

- Examine DNS for *.uk
- Examine search engines for their coverage of .uk
- Statistical sampling
- Get figures from Web auditing companies or the research community
- ...

What Challenges Will We Find?

- What about the “*Invisible Web*” – Web resources which cannot easily be indexed by search engines (dynamic sites, proprietary formats, etc.)
- Legal and ethical issues of auditing tools
- ...



Web Estimates – Numbers

Netcraft (www.netcraft.com):

- Polls Web servers (over 36 million) and reports on Web server software usage, trends, etc.
- Based in Bath
- They do store information on the .UK Web sites, but information is not reusable (batches of 2,000)

Many thanks to Netcraft for supplying this information

co.uk	2,750,706
org.uk	170,172
sch.uk	16,852
ac.uk	14,124
ltd.uk	8,527
gov.uk	2,157
net.uk	580
plc.uk	570
nhs.uk	215
police.uk	66
mod.uk	26
bl.uk	25
...	
Total	2,964,056

Web Estimates – Numbers

OCLC (www.oclc.org):

- Have a Web Characterization Project (WCP)
- In 2001 UK Web sites consists of 3% of total of 8,443,000 (i.e. 253,290 unique Web sites)
- See `<http://wcp.oclc.org/>` and `<http://wcp.oclc.org/stats/explanation.html>` (information on sampling methodology)

Web Estimates – Size

You can use search engines to count the numbers of pages indexed by domain

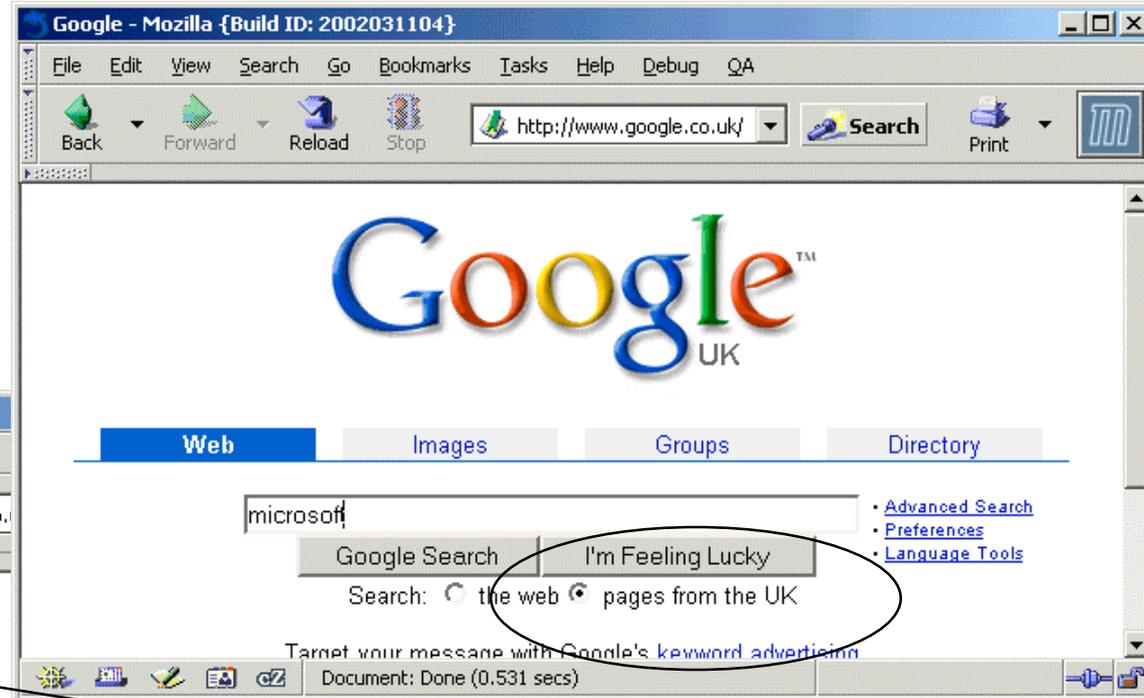
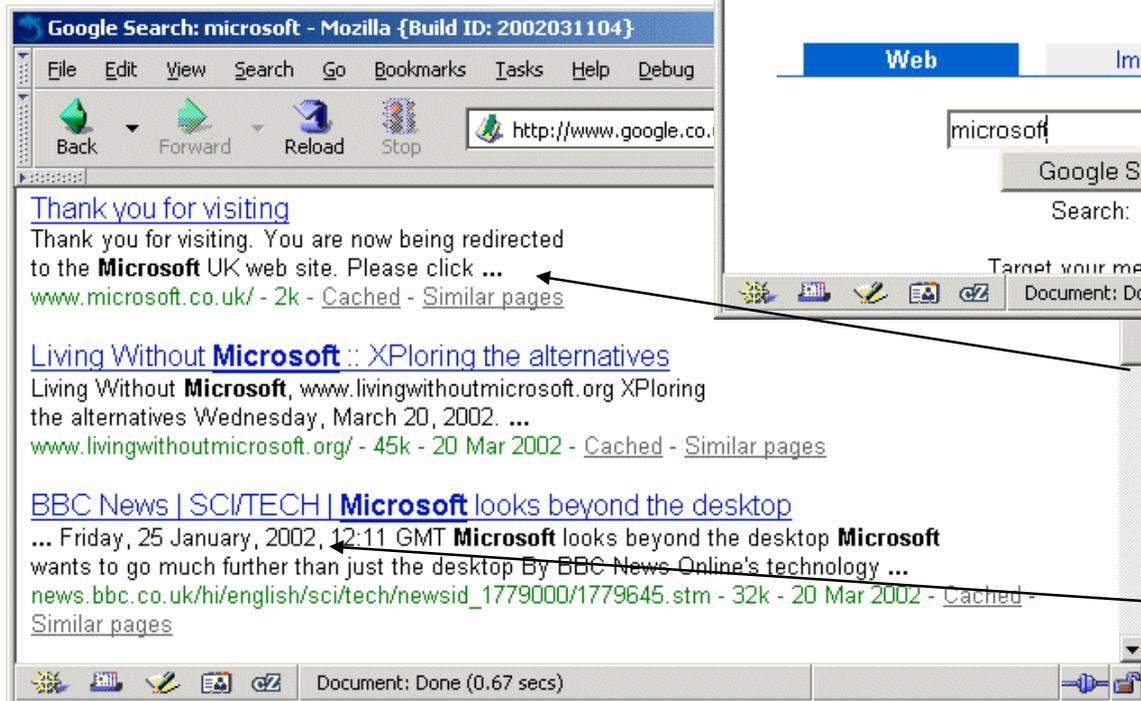
Search Term (AltaVista)	No. of Pages	Search Term (Google)	No. Of Pages
url:*.ac.uk	5,598,905	site:.ac.uk uk	2,080,000
url:*.co.uk	15,040,793	site:.co.uk uk	3,570,000
url:*.org.uk	1,644,322	site:.org.uk uk	898,000
url:*.gov.uk	975,506	site:.gov.uk uk	343,000
url:*.uk	24,862,369	site:.uk uk	4,760,000

Remember tools index the public and indexable Web and findings are subject to interpretation (dynamic pages, duplicate pages, inconsistencies, fluctuations, etc.)

In Google, searching for pages containing term “uk” (which may includes *uk* in domain name)

Google

How does Google define “pages from the UK”?



- What does Google do with redirects?
- Why is www.livingwithoutmicrosoft.org in the UK?

“Google uses a mix of heuristics, e.g. domain names, analysing redirects, links from UK directories, etc.”

Size Of The .UK Web

Number of Web servers

253,290 unique Web sites (OCLC WCP figures for 2001)

2,964,056 Netcraft (recent) figures

Number of pages (AV)

24,862,369

Further research can be carried out and the accuracy of these figures discussed, but let's move on

UKOLN Work

UKOLN has:

- Experiences with the BLRIC-funded WebWatch project (development and use of a robot for profiling various Web communities)
- Involvement in other harvesting work (EU-funded DESIRE project, RDN work, etc.)
- Published findings of semi-automated surveys across mainly UK HE Web sites, published in WebWatch column in Ariadne (after WebWatch funding finished and software developer left)
- Carried out pilot study of mirror eLib project Web sites

Archiving eLib Projects

Background

- Surveys of eLib project Web sites and EU Telematics For Libraries (TFL) projects showed that project Web sites were disappearing shortly after the funding had finished!
- A pilot study into the issues of archiving eLib project Web sites was carried out at the request of eLib Central Office.
- See `<http://www.exploit-lib.org/issue7/webwatch/>` for profiles (of 103 TFL projects 11 domains & 12 entry points had gone)
- See `<http://www.ariadne.ac.uk/issuexx/webwatch/>` for profiles of eLib Web sites

Archiving Pilot

What we did:

- Used a Web mirroring tool to mirror a number of the eLib project Web sites
- Observed problem areas
- Reflected on issues which emerged

Our main findings concerned:

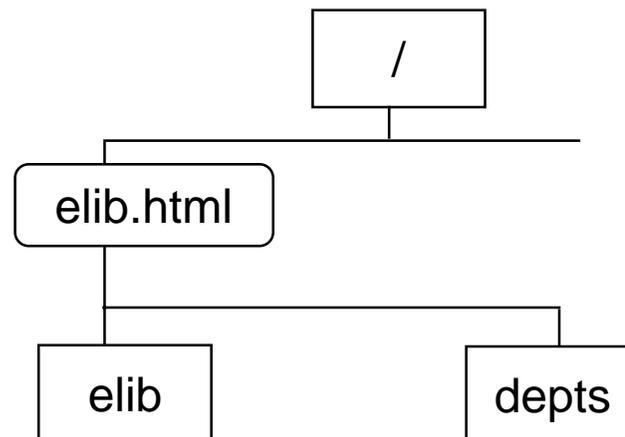
- Setup of the Web services
- Tools used to carry out mirroring
- Purpose of the mirroring exercise
- Legal, ethical, etc. issues

Issues

Issues which the pilot (and related work) revealed included:

- Should Web site be archived if use of robots is banned?
- At times difficult to identify a “site” – project Web site may be confused with entire organisational Web site

Mirroring
foo.ac.uk/
elib.html will
result in entire **foo**
Web site being
mirrored



Issues

Other Issues:

- What are we attempting to preserve:
 - Static documents on Web site
 - Functionality of Web site

Static documents are relatively easy to mirror

If the aim of a Web site is, say, to provide a subject gateway, there may be an expectation that the gateway service will be mirrored. This is not possible with a remote harvesting approach

Issues

Other issues include:

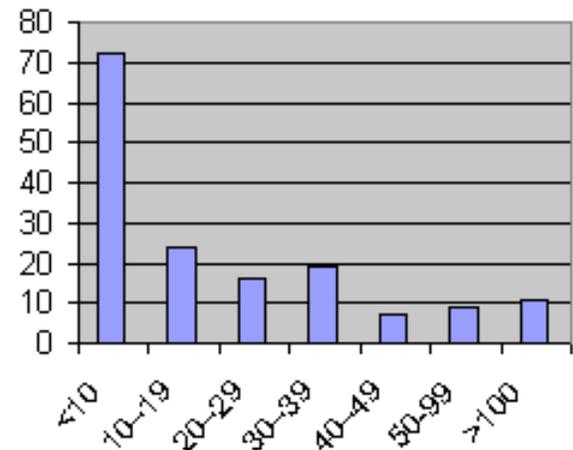
- **Copyright, Data Protection, etc.**
The mirror included copies of various logos, images, etc
- **Dynamic Web Sites**
How should dynamic Web pages be preserved?
- **Embedded Objects**
The mirror included text and images, but not necessarily CSS and JavaScript files
- **Frequency Of Archiving**
A one-off archive, regular archiving, archiving on demand (after major changes)
- **Absolute URLs, Server Redirects, etc.**
It is not clear what should be done if redirects are encountered and if absolute URLs are used

Profiling Nos. Of Servers

A survey of the numbers of Web servers in UK Universities was carried out in June 2000 and repeated recently.

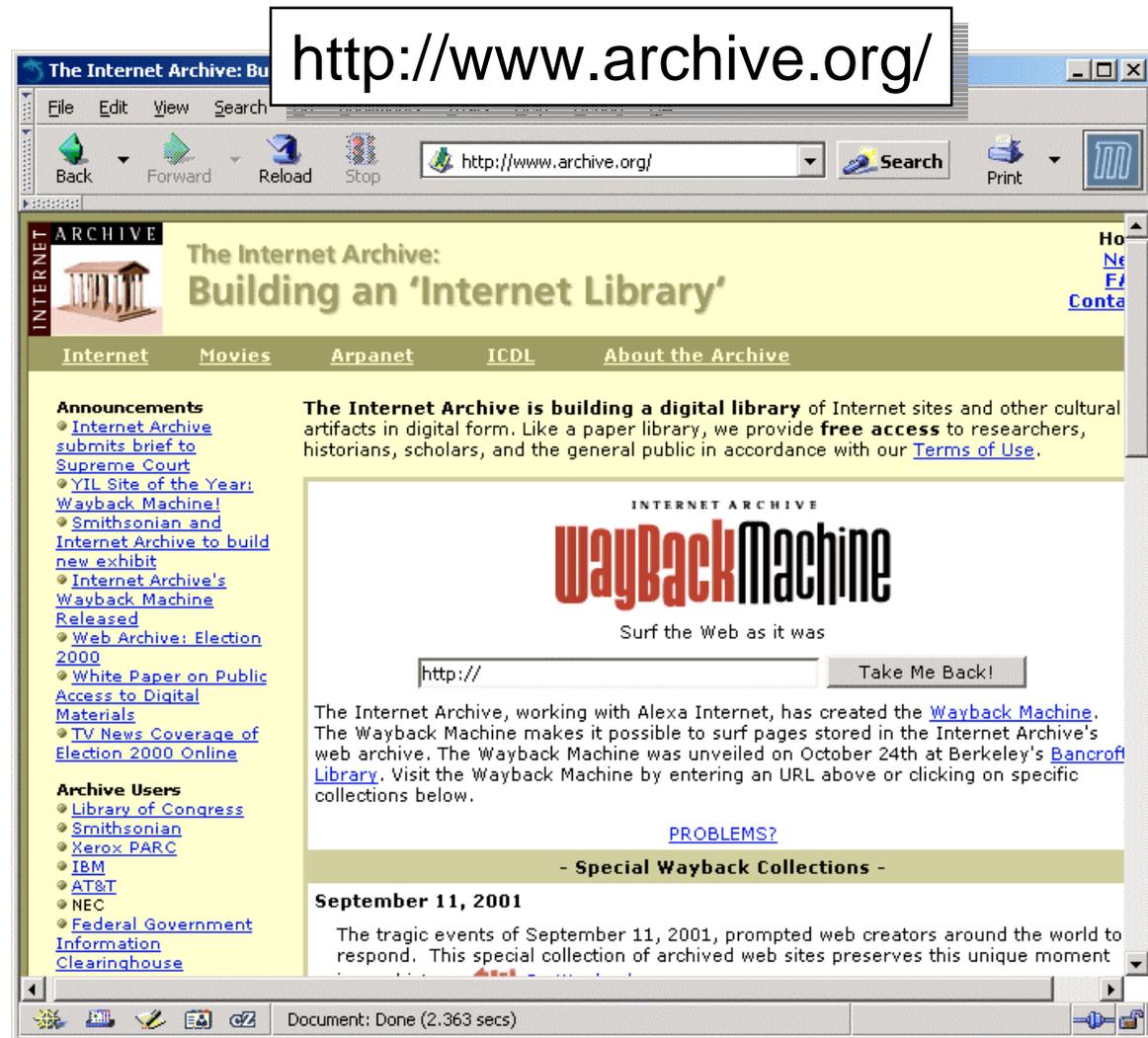
Cambridge has the most number of Web servers (now 369)

The average is 24.2 servers per institution



Internet Archive

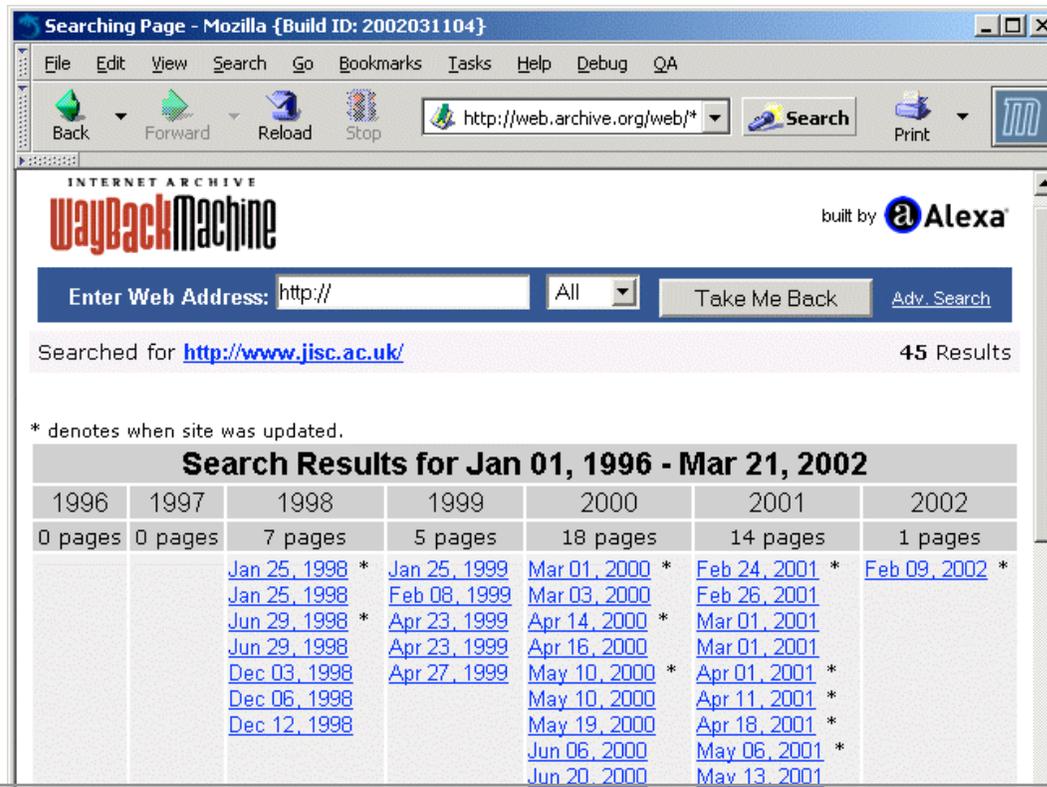
The **Internet Archive** is a “public nonprofit that was founded to build an ‘Internet library,’ with the purpose of offering permanent access for researchers, historians, and scholars to historical collections that exist in digital format.”



Has the Internet Archive solved the problems we experienced?

The Wayback Machine

The Wayback Machine is a public interface to the Internet Archive.



The screenshot shows a Mozilla browser window titled "Searching Page - Mozilla {Build ID: 2002031104}". The address bar contains "http://web.archive.org/web/*". The page displays the Wayback Machine logo, built by Alexa, and a search interface. The search results for "http://www.jisc.ac.uk/" are shown, indicating 45 results. A table of search results for the period from Jan 01, 1996 to Mar 21, 2002 is displayed, showing the number of pages and the dates of updates for each year.

INTERNET ARCHIVE
WayBackMachine built by 

Enter Web Address: All

Searched for <http://www.jisc.ac.uk/> 45 Results

* denotes when site was updated.

Search Results for Jan 01, 1996 - Mar 21, 2002						
1996	1997	1998	1999	2000	2001	2002
0 pages	0 pages	7 pages	5 pages	18 pages	14 pages	1 pages
		Jan 25, 1998 *	Jan 25, 1999	Mar 01, 2000 *	Feb 24, 2001 *	Feb 09, 2002 *
		Jan 25, 1998	Feb 08, 1999	Mar 03, 2000	Feb 26, 2001	
		Jun 29, 1998 *	Apr 23, 1999	Apr 14, 2000 *	Mar 01, 2001	
		Jun 29, 1998	Apr 23, 1999	Apr 16, 2000	Mar 01, 2001	
		Dec 03, 1998	Apr 27, 1999	May 10, 2000 *	Apr 01, 2001 *	
		Dec 06, 1998		May 10, 2000	Apr 11, 2001 *	
		Dec 12, 1998		May 19, 2000	Apr 18, 2001 *	
				Jun 06, 2000	May 06, 2001 *	
				Jun 20, 2000	May 13, 2001	

See Greg Notess's article "*The Wayback Machine: The Web's Archive*" in *Online*, Mar/Apr 2002, Vol. 26, No. 2. ISSN 0146-5422

Also at <http://www.infotoday.com/online/mar02/OnTheNet.htm>

Using The Wayback Machine

When using the Wayback Machine:

- You get a fairly faithful view (images included, unlike Google's cache)
- You stay in the machine when you follow links (to closest date)

The image shows two side-by-side screenshots of a Mozilla browser window. The left window displays a list of events under the heading 'Past events'. The right window displays a detailed page for 'Videoconferencing - Getting started'.

Left Window: JISC Events - Mozilla {Build ID: 2002031104}

assist@jisc.ac.uk

JANET User Support Workshop, 8 - 10 September 1998, University of Cambridge

Networking Strategy Workshop, 16 - 17 December 1998, Robinson College, Cambridge.

Past events

Networking Strategy Workshop 16-17 December 1997, Trinity & All Saints University College, Leeds.

JISC INFORMATION STRATEGIES CONFERENCE 2nd December 1997, School of Oriental and African Studies, London

Videoconferencing - Getting started A joint UKERNA/JTAP event on 20th November 1997, at the Royal Geographical Society, London (with the British Geographers) For further information, Email: workshop@ukerina.ac.uk see URL: <http://ctiweb.cf.ac.uk/jtap/getstar.htm>

Facing the legal challenges of providing internet access in higher education establishments, (one-day seminar), Thursday 23 October 1997 at ECU, University of Loughborough. The aim of the conference was to provide an overview of the main legal issues which might arise from the provision of internet access in Further and Higher Education establishments.

Right Window: Videoconferencing - Getting Started - Mozilla {Build ID: 2002031104}

Videoconferencing
Getting started

A JOINT UKERNA/JTAP EVENT

THURSDAY 20th NOVEMBER, 1997

Royal Geographical Society, London
(with The Institute of British Geographers)

The use of videoconferencing in higher education is growing, mainly due to the fact that it is an application that has enormous potential in education and training and is becoming established as an educational technology for the advancement of teaching and learning. Videoconferencing is also set to play an increasingly important role in the expansion not only of intra-UK activities but also of European and international educational delivery roles.

Many academic staff now need to know how videoconferencing may work effectively for them and seek

United Kingdom Education and Research Networking Association

JTAP

Conclusions

To conclude:

- Measuring the size of the UK Web is difficult
- There is a need to define our terminology
- If measuring is difficult, preserving Web sites that we can't count will be even more difficult!
- Experiences of robot developers, Web indexers, etc. will provide useful information