

WWW: The Whys and Whats of Web Archiving

Maureen Pennock
Digital Curation Centre
UKOLN, University of Bath



This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 2.5 UK: Scotland License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/2.5/scotland/>; or, (b) send a letter to Creative Commons, 543 Howard Street, 5th Floor, San Francisco, California, 94105, USA.

Funded by:



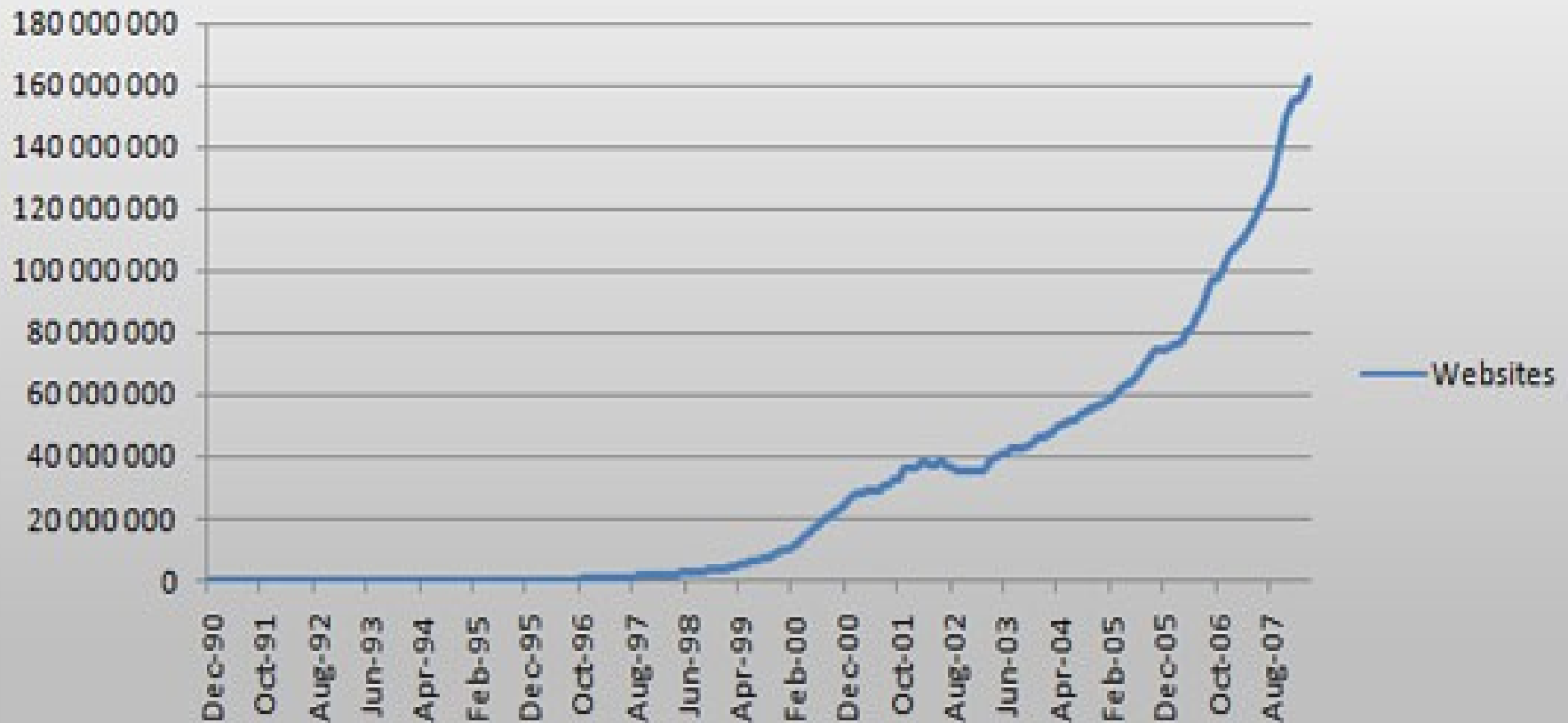
Overview

- What exactly is the World Wide Web?
- Why bother archiving websites?
- What you should think about about
- What you should know about
- What you could do next...



What is the World Wide Web?

Number of websites (1990 - 2008)



pingdom

Source: Royal Pingdom

World Wide Web

The WorldWideWeb (W3) is a wide-area [hypermedia](#) information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an [executive summary](#) of the project, [Mailing lists](#) , [Policy](#) , [No W3 news](#) , [Frequently Asked Questions](#) .

[What's out there?](#)

Pointers to the world's online information, [subjects](#) , [W3 servers](#), etc.

[Help](#)

on the browser you are using

[Software Products](#)

A list of W3 project components and their current state. (e.g. [Line Mode](#) ,X11 [Viola](#) , [NeXTStep](#) , [Servers](#) , [Tools](#) , [Mail robot](#) , [Library](#))

[Technical](#)

Details of protocols, formats, program internals etc

[Bibliography](#)

Paper documentation on W3 and references.

[People](#)

A list of some people involved in the project.

[History](#)

A summary of the history of the project.

[How can I help ?](#)

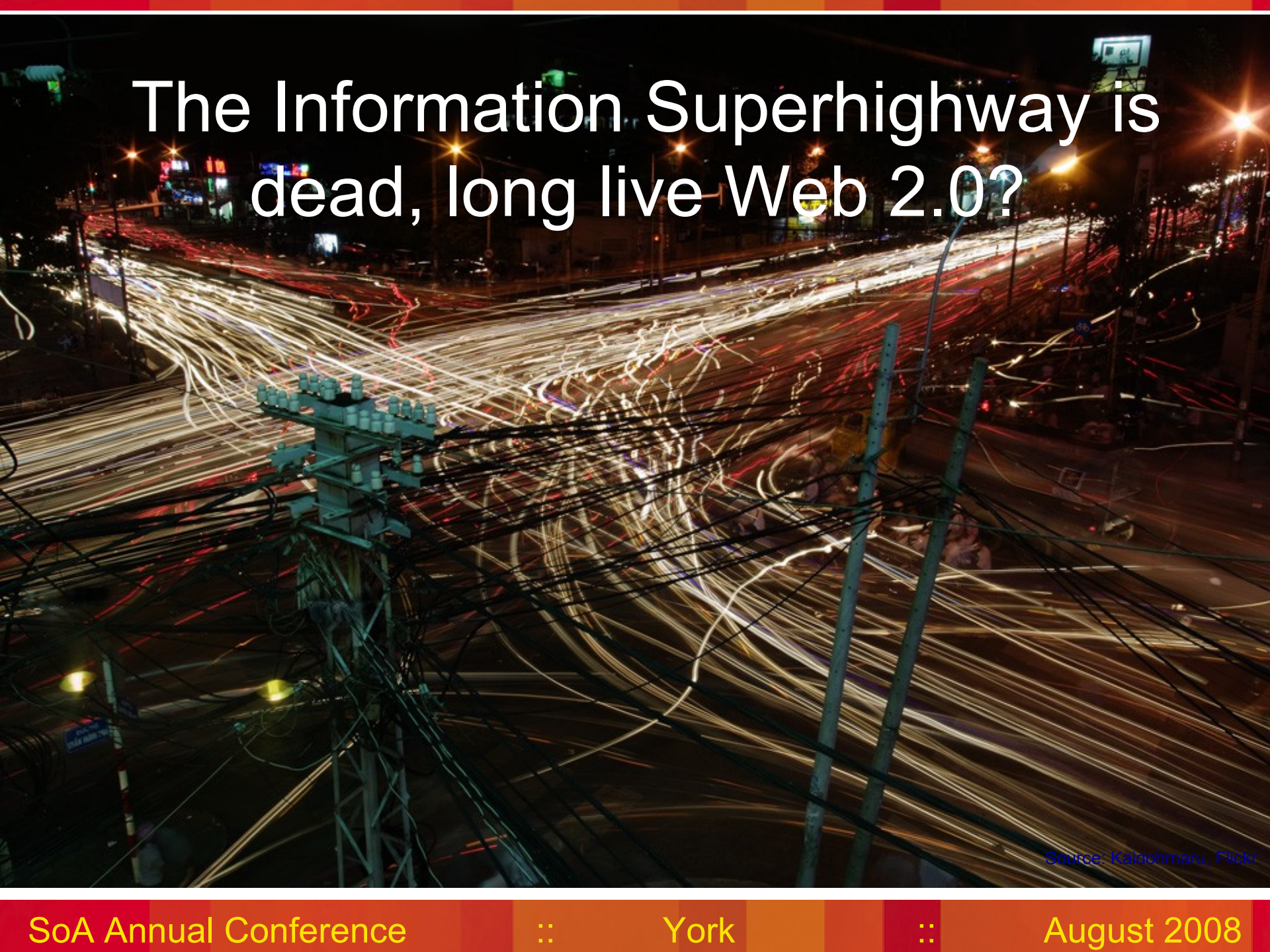
If you would like to support the web..

[Getting code](#)

Getting the code by [anonymous FTP](#) , etc.

Source: W3C.org

The Information Superhighway is dead, long live Web 2.0?

A long-exposure photograph of a busy city street at night. The image shows light trails from cars, with white and yellow streaks indicating forward motion and red streaks indicating brake lights. In the foreground, a dense network of black power lines and cables is visible, supported by metal poles. The background shows city lights and traffic signals.

Source: Kaidohmaru, Flickr

- What's on it now?
- More like what's
not on it!

•Source: [Sebastien Prooth](#), [Flickr](#) [Foobar poster from Fooboy]



axman (D.-Calif.), the new chairman of the House Oversight and
ast fall's election results. Waxman and others are miffed because CPCs
nection between abortion and breast cancer. He is also the Democrats'
abstain from sexual relations. Yet going after CPCs is unlikely to
ing voters. More insidious is the issue of
wa tion of abortion was losing them votes,
igh moderate on the issue and even recruited
ru erable success. Now that they have taken
ay educing abortion while not doing anything
al y-death base, and at the same time
mi on grays." These are American voters who
ish susceptible to appeals from either side
gra emist both the NARAL, Barbara Boxer
ort tion activists who believe every single
con of Roe v. Wade and political realities,
asi the time being, so pro-abortion forces
ore ach could not only divide and demonize
ong ic distress, and inflate the number of
co educing abortion and watch pro-lifers,

It looks like you're posting
bullshit to the Internet.

What would you like to do?

- Actually check your facts for once
- Take a basic course in logical reasoning
- Enter the reality-based community

Autosaved draft at 7:16:25 PM



Allow Comments?

Journal Default ?

Comment Screening:

Journal Default ?

Preview

Spell check

Source: [Marmanel, Flickr](#)

Why archive websites? (i)

- Web sites can contain uniquely available informative records and data
 - Users may act or take decisions based on this information, with important consequences
- Records of business transactions
- Accountability & transparency
 - To funding bodies
 - To stakeholders
 - For legal reasons
- Some examples...

Communicate

[Ask the PM](#)

[...from the PM](#)

[e-Petitions](#)

[Petition Responses](#)

Communicate

Ask the PM

Gordon Brown wants to hear about the issues that matter to you through a new YouTube initiative. Submit your video question and view the PM's answers in this regular feature.

[Go to the Number 10 YouTube channel](#)

...from the PM

Hear directly from the PM on specific issues of importance to the nation.

In addition to our regular newsletter



Newsletter

Sign up to our newsletter to keep updated with the latest information from Number 10.

[Click here to subscribe](#)

Around the Web


[flickr](#)

Latest Photos



You are here: Wiltshire County Council > [Home](#)

Search the site

Type your search here 

A to Z of Services

Council

Advice and Benefits

Business

Community and Living

Council and Democracy

Education and Learning

Environment and Planning

Health and Social Care

Jobs and Careers

Leisure and Culture

Transport and Streets

Customer Contact Centre

Email: [Customer Care](#)

Tel: 01225 713000

View [Out of hours numbers](#)

Opening Hours:

Wiltshire County Council



Change colours and background on this site



Increase text size [A](#) [A](#) [A](#)

Latest news articles

▶ [School admissions for 2009](#)

Wiltshire parents are being encouraged to apply for the school they would like their children to attend next September. The online application form ...

▶ [Wiltshire's new council takes to the road](#)

Wiltshire's new council is taking to the road to get local people involved in its future. Wiltshire Council will come into effect in April 2009, when ...

▶ [More News...](#)

Featured Items

- ▶ [Can't find what you're looking for...](#)
- ▶ [Towards One Council](#)
- ▶ [Equality Action Plan Consultation](#)
- ▶ [Learn about the Beyond the Immediate Project](#)

What's On

- ▶ [The Wiltshire Open at Salisbury Library](#)
- ▶ [Warminster Library Readers Group](#)
- ▶ [Noisy Storytime at Salisbury Library](#)
- ▶ [Harry Potter Team Read Special at Trowbridge Library](#)

Popular areas

- ▶ [Do It Online](#)
- ▶ [Jobs online](#)
- ▶ [Wiltshire Strategic](#)
- ▶ [Community Histor](#)
- ▶ [Find your Council](#)
- ▶ [Council Meetings](#)
- ▶ [Compliments and Concerns](#)
- ▶ [Public Documents](#)
- ▶ [A to Z of Highway](#)
- ▶ [Roadworks](#)
- ▶ [Planning](#)





Welcome to the University of Bath



[text view](#) | [campus home](#)

[a-z](#) | [contacts](#) | [find people](#) | [experts](#) | [webmail](#)

Wed 27 Aug 2008, 20:52

Prospective students

[Our programmes](#)

[Applying to Bath](#)

[Undergraduate prospectus](#)

[Postgraduate prospectus](#)

[International](#)

Research & innovation

[Research](#)

[Business services](#)

Development & Alumni

[Alumni](#)



Students: Welcome to Bath! ➤

News and events

- ❖ [Action imminent for Team GB's University of Bath based modern pentathlon contingent](#)
- ❖ [Researchers study how cancer cells 'come unstuck' & produce secondaries](#)
- ❖ [Student designs prize-winning Olympic starting blocks](#)
- ❖ [University of Bath launches MSc in Innovation & Technology Management](#)

[more news »](#)
[podcasts and videos](#) | [video intro »](#)

University of Bath, Bath, BA2 7AY, UK · tel 01225 388388

© 2007 · [disclaimer](#) · [privacy statement](#) · [FoI](#) · updated: 27 August, 2008 by the web team

about

[the University](#)
[our departments](#)
[conferencing](#)
[job vacancies](#)

visit

[open days](#)
[virtual tour](#)
[getting here](#)

experience


[sport](#) : [arts](#)
[student life](#)
[city of Bath](#)
[community links](#)

Are you:

My Dundee

ATTENTION: All Users!



my dundee >> summer 2008 

Virus Warning**

Users are reminded not to open e-mails from unknown senders which contain attachments, particularly if the subject line refers to "Statement of Fees 2008/09" or similar. For more information, see the guidance from ICS at

<http://www.dundee.ac.uk/ics/services/virus/>

System Upgrade

My Dundee has now been upgraded to version 8.0.260.7. Thank you to the ICS and Learning Centre teams for achieving this.

- ◆ Please report any problems or glitches to vle@dundee.ac.uk in the usual way.

- ◆ Please log any problems with the [ICS Service Desk](#) and

Terms of Use

Using "My Dundee" means you are subject to, and agree to abide by the following:

[Code of Conduct](#) and
[Regulations for the Use of Computing Facilities](#)

Login Here

Enter login information here and click the Login button below.

Username:

Password:

Login

Change Your Password

Forgotten or need to reset your password? Visit the ICS [Passw Portal](#). Remember this will change your password for the network Groupwise and *My Dundee*. More password [information and adv](#) available from ICS. (Close the new browser window to return to *Dundee* - if you have reset your password you may need to log

Why archive websites? (ii)

- Cultural heritage objects: reflection of modern society

Why archive websites? (ii)

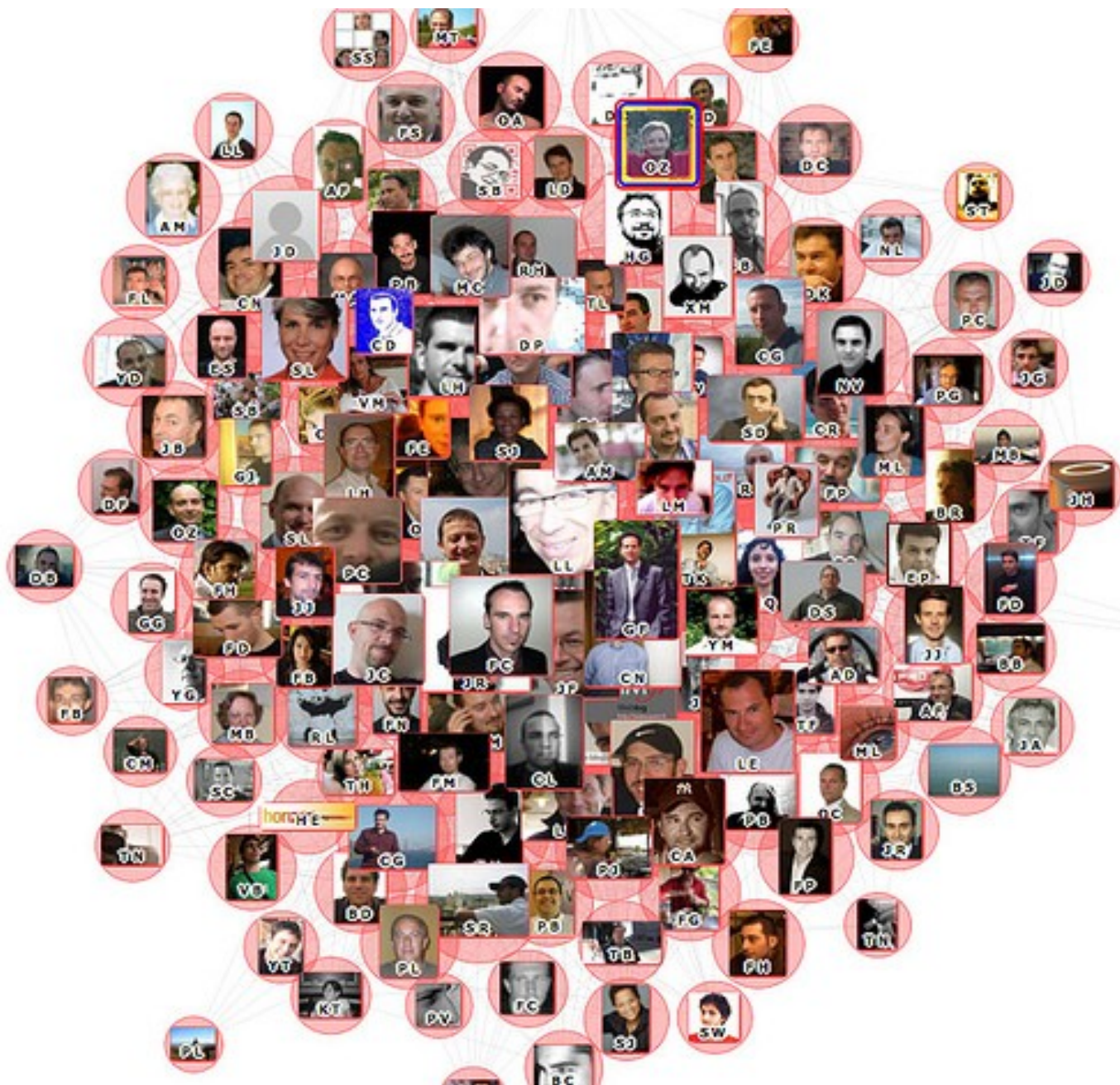
- Cultural heritage objects: reflection of modern society



Why archive websites? (ii)

- Cultural heritage objects: reflection of modern society





Social
networks:
linking
people on
a global
scale

Source: [Luc Legay](#), Flickr

Questions to ask yourself (i)

- What exactly do I want to archive?
 - My own website?
 - Just the back-end data?
 - Records hosted on the website?
 - Metadata about the website?
 - User generated website data?
 - Other people's websites?
- Why do I want to archive them?

Questions to ask yourself (ii)

- Who owns the websites and their content?
- Do I have legal permission to collect and re-host the content?
- Which websites do I want to archive?
- Where will the money for this come from?
- How will I do all this?
 - Should I do it myself or rely on others?
- Who else is doing it and what tools do they have?



Good job there's an Internet Archive then!

Source: Mr Wright, Flickr

But...

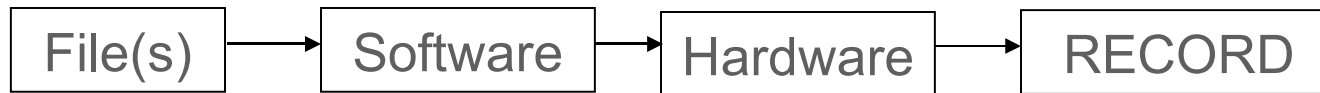
[Skip to: Content](#) |

[The crest for Number 10 Downing Street](#)

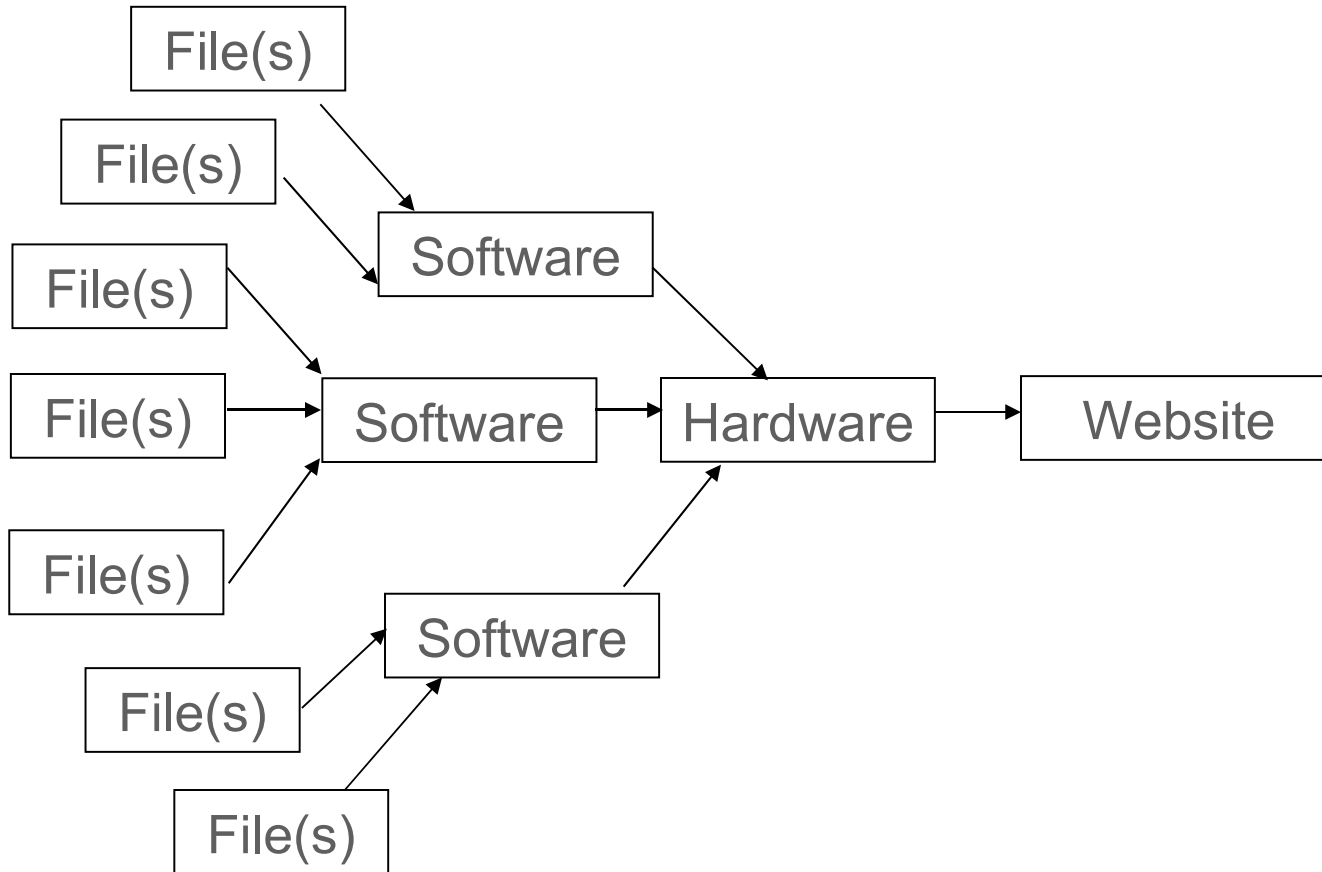
Main menu

- ♦ [prime minister](#)
 - ◊ [contact](#)
 - ◊ [biography](#)
 - ◊ [speeches](#)
 - ◊ [PM's office](#)
 - ◊ [big issues](#)
- ♦ [government](#)
 - ◊ [cabinet](#)
 - ◊ [guide to legislation](#)
 - ◊ [guide to government](#)
 - ◊ [in your area](#)
 - ◊ [links](#)
- ♦ [newsroom](#)
 - ◊ [latest news](#)
 - ◊ [media centre](#)
 - ◊ [email updates](#)
 - ◊ [photo galleries](#)
 - ◊ [webchats](#)
 - ◊ [other languages](#)
- ♦ [downing street](#)

So what's the technical problem?



So what's the technical problem?



UKWAC



- UK Web Archiving Consortium (6 members)
 - British Library, National Library of Scotland, National Library of Wales, The National Archives, Wellcome Library, JISC
- Collects Web content selectively according to individual interests
 - Uses modified collection/harvesting software developed by the National Library of Australia
 - Permission is sought from site owners in advance
 - Allocates Persistent Identifier URLs
 - Partners assumes responsibility for their 'own' sites
 - Central repository of metadata
 - The collections are publicly accessible
- Website: <http://www.webarchive.org.uk/>

IIPC



- International Internet Preservation Consortium
 - Co-ordinated by the National and University Library of Iceland
 - UK National Archives also a member
- Online toolkit that addresses:
 - Acquisition
 - Focussed selection and verification
 - Collection storage & maintenance
 - Access & finding aids
- Website: <http://www.netpreserve.org>

Web Continuity project

- Led by the UK National Archives
- Comprehensive of government websites by TNA
- Redirection tool to direct users to archived content if original is no longer available
- Guidance for government webmasters on best practice for design and maintenance
- Due to complete in November 2008
- See
<http://www.nationalarchives.gov.uk/webcontinuity/>

What you could do next (i)

- Identify your requirements: what do I want to archive and why?
- Seek help from the experts
- Develop a written policy and strategy to support activities and help secure resources
- Take a life-cycle approach to support curation and preservation planning
- Be aware of other people's web archiving activities: check other heritage collections before gathering!

What you could do next (ii)

- Plan for preservation activities to maintain access to authentic resources over time and avoid incurring extra costs
- Determine access and user requirements costs
- Re-assess your strategy on a very regular basis

Thank You

Maureen Pennock

m.pennock@ukoln.ac.uk

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



This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 2.5 UK: Scotland License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/2.5/scotland/>; or, (b) send a letter to Creative Commons, 543 Howard Street, 5th Floor, San Francisco, California, 94105, USA.

Funded by:

