

Beyond the Harvest: Long Term Preservation of the UK Web Archive

JISC, the DPC and the UK Web Archiving Consortium Workshop
Missing links: the enduring web

Maureen Pennock

Digital Preservation project manager, BL Web Archiving Programme

July 2009

The UK Web Archive: Some background

- JISC Feasibility study (2003) led to foundation of UKWAC
- UK Web Archive collection started in 2004
 - PANDAS software from the NLA
 - Web Curator Tool from BL & NLNZ
- Now contains almost 5,000 different titles
- Exclusively comprised of selectively harvested material
 - High quality control measures
- Over 4TB in size
 - Potentially unlimited range of file types

An Enduring Prospect?

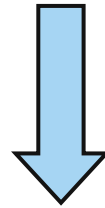
- Objectives:
 - Build a comprehensive web archive as part of the British Library's digital collection
 - Preserve the archive so that it remains accessible into the future
 - Put in place [the necessary] people, process and systems
- Challenge of digital preservation:
 - Managing changes in technology so that web archives remain reliably accessible over time
 - Size and structure of web archives make this a huge challenge
 - Most activity so far concentrated on harvesting process

A task of many parts...

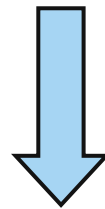


1. Documenting System Dependencies

PANDORA Web Archiving System



Web Curator Tool



...?

2. Containers & Metadata Standards

ARC

WARC

PREMIS

METS

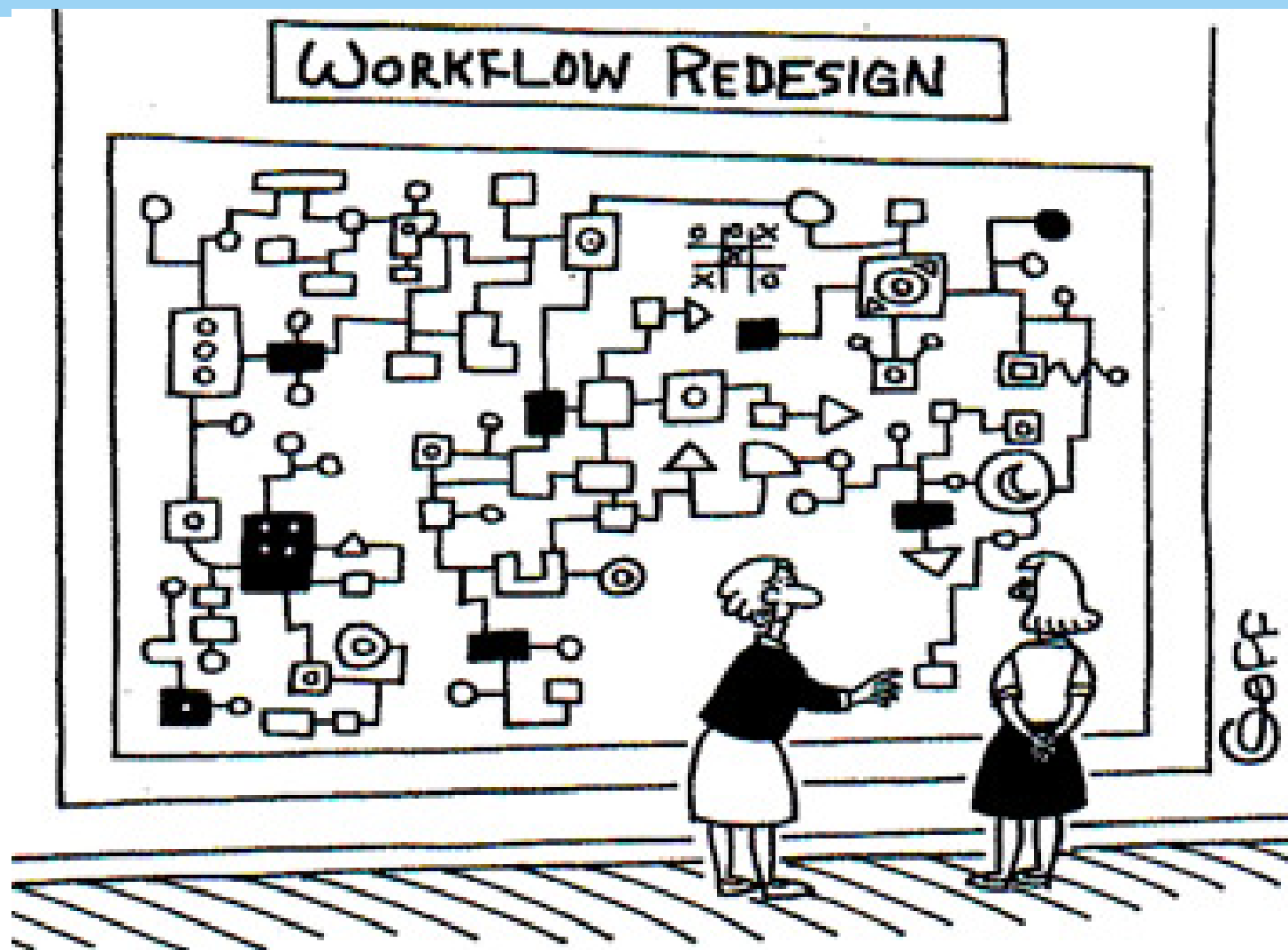


3. Metadata for Archived Websites

- What are we describing?
- What is the document model?
- What about preservation?
- Variations for different information packages
- How can we support retrieval?



4. Preservation Workflow



"And this is where our ED workflow redesign team went insane."

5. Technology Watch Blog

BRITISH LIBRARY

UK WEB ARCHIVE

Technology Watch

bl.uk • Latest • About blog • UK Web Archive website • Subscribe 



Maintained by British Library staff in the Digital Preservation Team and Web Archiving Programme

26 June 2009

Opera Unite: DIY hosting?

Opera Software has launched a new version of its Opera browser called Opera Unite. Currently available as an alpha build, Opera Unite offers users the ability to connect with each other in a more direct manner, rather than through third party servers. Opera Unite browsers have an integrated server component

6. Defining our Preservation Strategy

- A long term activity
 - Integrating and combining previous activities
 - Preservation planning
 - Significant properties research
 - Establishing a preservation workbench
 - Putting in place satisfactory virus protection
 - Revisiting established 'best practice' re – harvesting
 - Regular Risk Assessment
 - Preserving web archives in BL's digital library system (DLS)
- Continued collaboration
- Consistency with BL's wider digital preservation strategy

To infinity and beyond?

- But what of the future?
 - Growth of online, closed communities
 - Alternative means of access
 - Virtual online worlds
 - Personalisation and personalised browsing
 - Deep web (still)
 - Capturing transactions
 - Browser and plug-in development
 - Increases in size of web and websites
 - And then some...



Not the end, but the beginning of a long journey!

maureen.pennock@bl.uk

<http://www.webarchive.org.uk/>