Collecting and preserving web content

Adrian Brown

Head of Digital Preservation

The National Archives

the national archives

Background

- UKWAC evaluation report
  - Evaluation of new collection methods
  - Digital preservation working group established to address UKWAC preservation requirements

UK WEB ARCHIVING CONSORTIUM
www.webarchive.org.uk

Defining the website

- Database-driven content
- Personalisation
- Syndicated content
- Scripting
- Multimedia

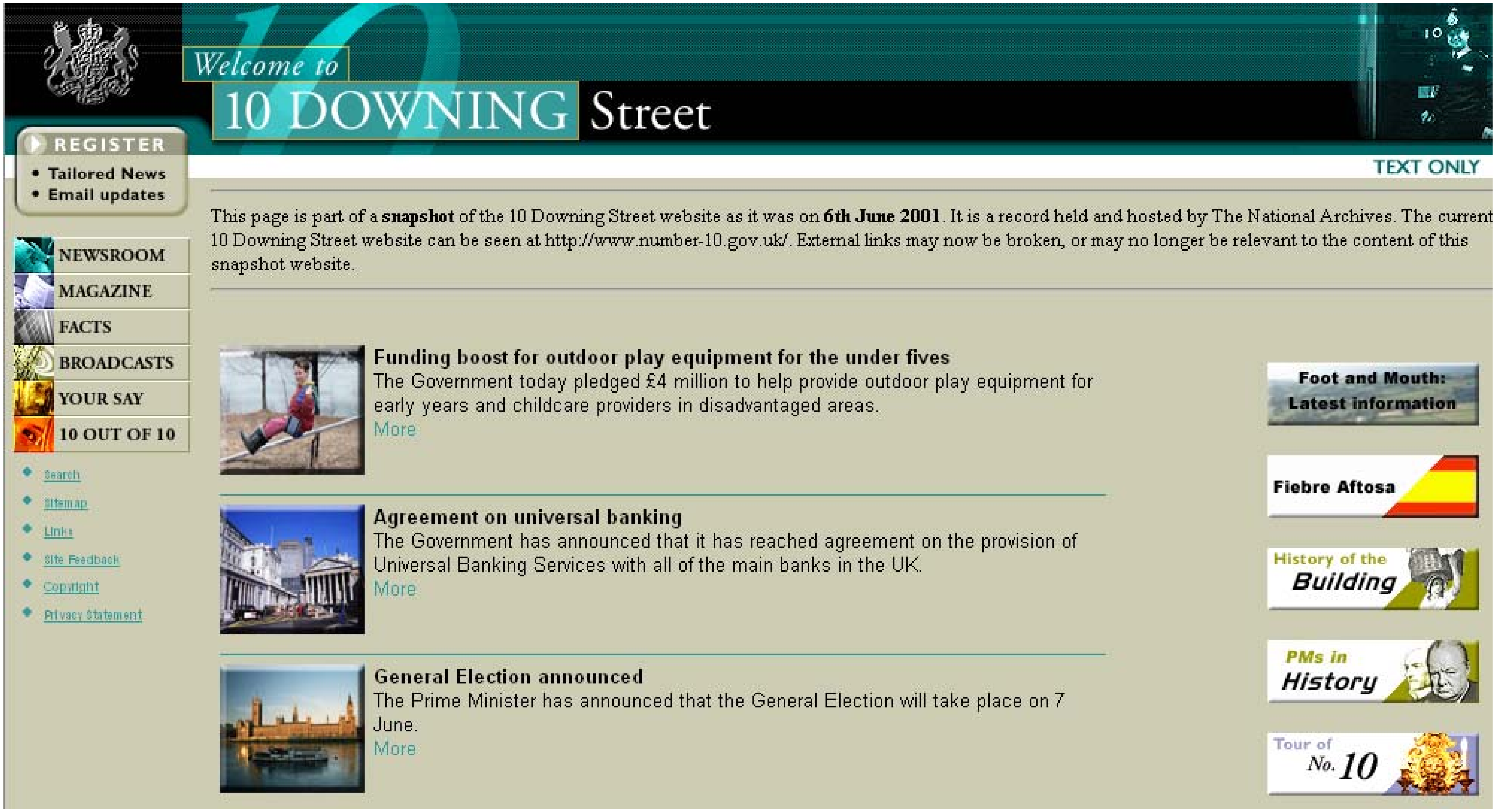
Defining the website

- Experience arises from interaction (transactions) between web server and client
- Content-driven view
    - Sum of all available content - set of all possible transactions
- Event-driven view
    - Actual transactions – subset of content delivered

Collection methods

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

Direct transfer



This page is part of a **snapshot** of the 10 Downing Street website as it was on **6th June 2001**. It is a record held and hosted by The National Archives. The current 10 Downing Street website can be seen at http://www.number-10.gov.uk/. External links may now be broken, or may no longer be relevant to the content of this snapshot website.

Funding boost for outdoor play equipment for the under fives
The Government today pledged £4 million to help provide outdoor play equipment for early years and childcare providers in disadvantaged areas.
More

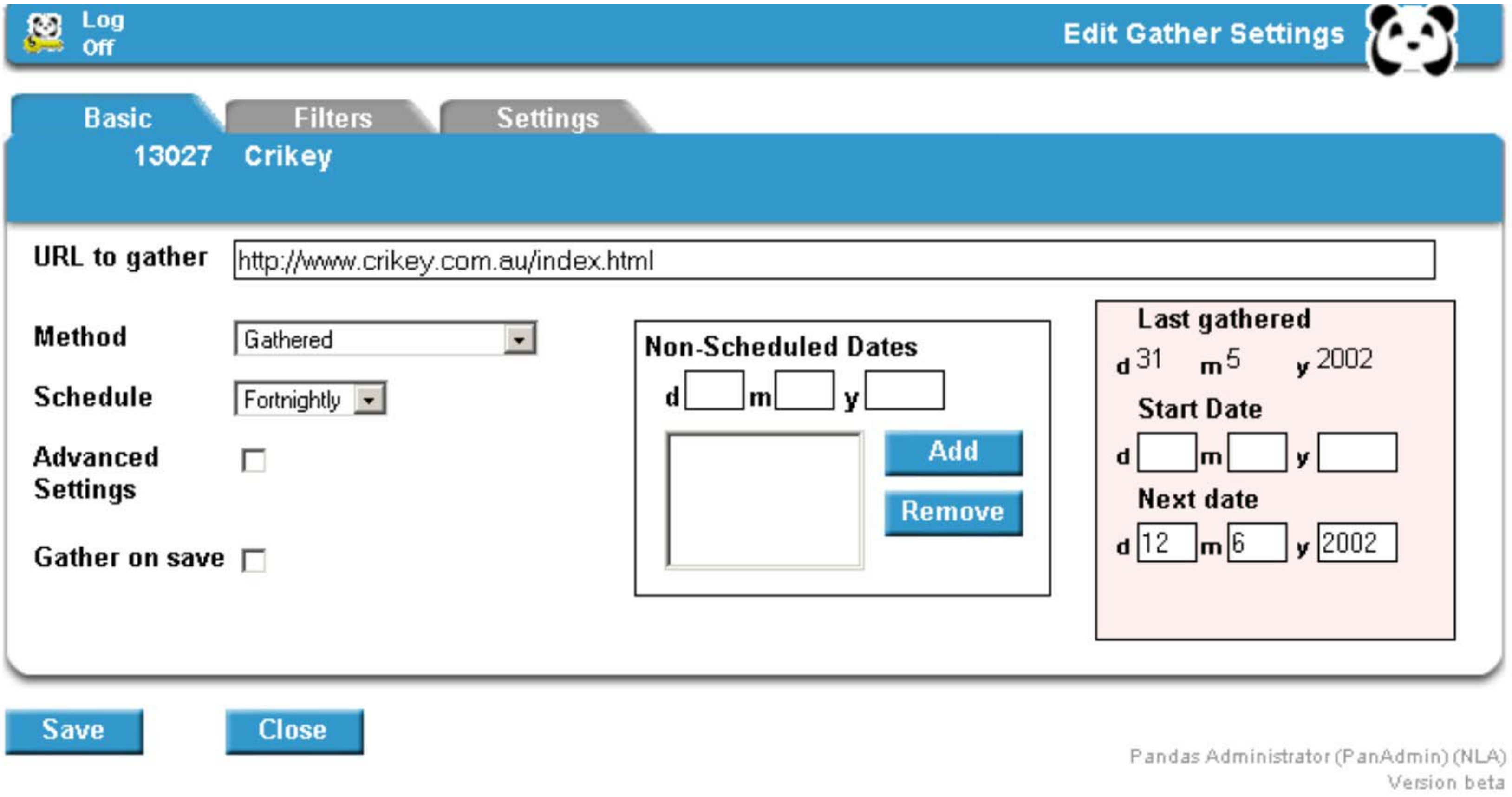
Agreement on universal banking
The Government has announced that it has reached agreement on the provision of Universal Banking Services with all of the main banks in the UK.
More

General Election announced
The Prime Minister has announced that the General Election will take place on 7 June.
More

Direct transfer

- Strengths
  - Potentially most authentic rendition
- Limitations
  - Manual and resource intensive
  - Potentially requires support for multiple technologies

Remote harvesting



13027 Crikey

URL to gather: http://www.crikey.com.au/index.html

Method: Gathered

Schedule: Fortnightly

Advanced Settings □

Gather on save □

Non-Scheduled Dates

d □ m □ y □

Last gathered

d 31 m 5 y 2002

Start Date

d □ m □ y □

Next date

d 12 m 6 y 2002

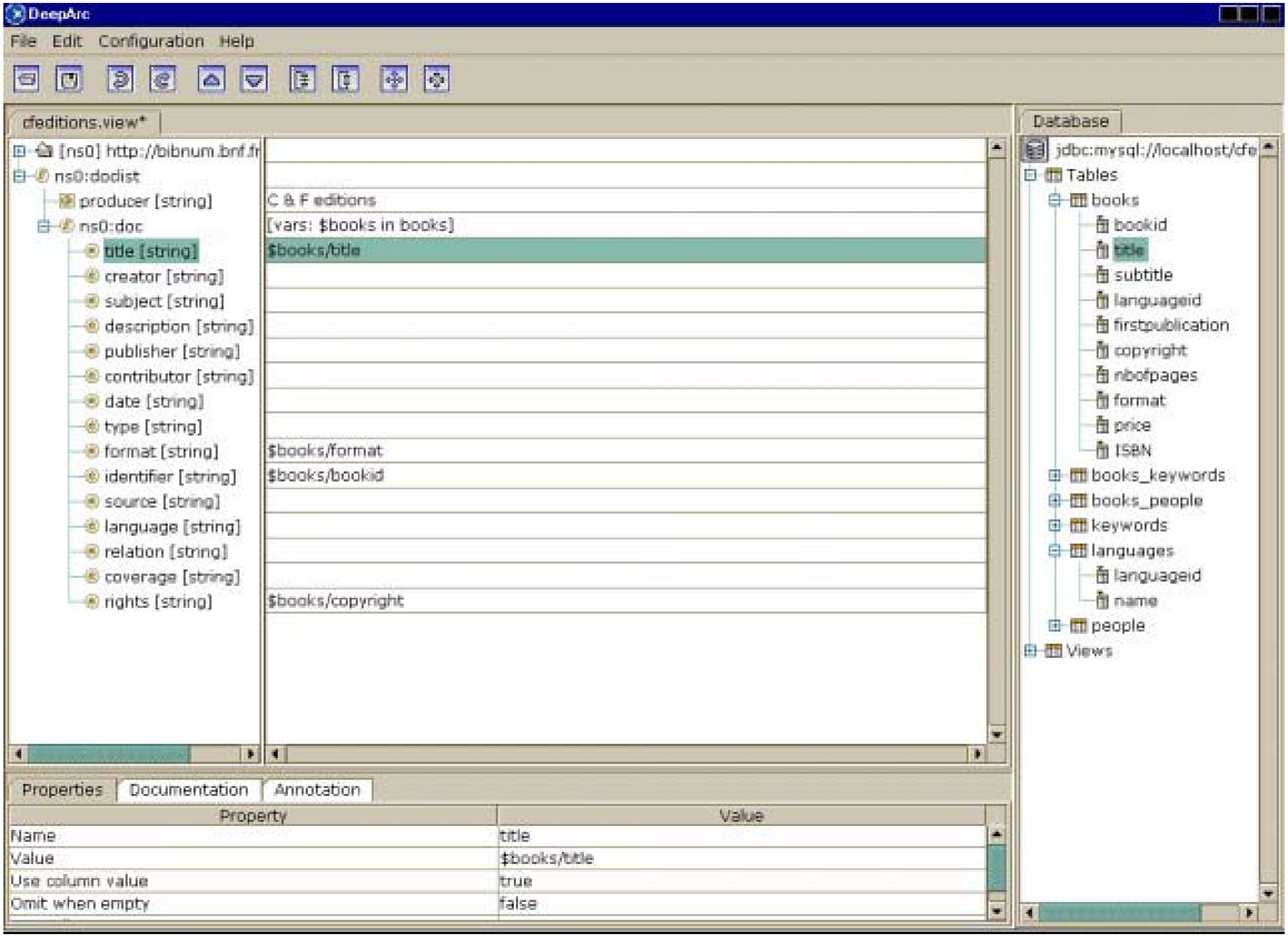
Pandas Administrator (PanAdmin) (NLA)
Version beta

Remote harvesting

- Strengths
  - Cost effective and simple to manage
  - Fastest method for large-scale collecting
  - Improved,mature tools available (e.g. Heritrix)
- Limitations
  - Can’t capture much dynamic content
  - Requires careful configuration

Database archiving

- Tools developed by IIPC:
    - DeepArc: Extracts database to XML repository
    - Xinq: Provides standarised search/browse interface to XML repository



[ns0] http://bibnum.bnf.fr	
ns0:doclist	
producer [string]	C & F editions
ns0:doc	[vars: $books in books]
title [string]	$books/title
creator [string]	
subject [string]	
description [string]	
publisher [string]	
contributor [string]	
date [string]	
type [string]	
format [string]	$books/format
identifier [string]	$books/bookid
source [string]	
language [string]	
relation [string]	
coverage [string]	
rights [string]	$books/copyright

- jdbc:mysql://localhost/cfe
  - Tables
    - books
      - bookid
      - title
      - subtitle
      - languageid
      - firstpublication
      - copyright
      - nbofpages
      - format
      - price
      - ISBN
    - books_keywords
    - books_people
    - keywords
    - languages
      - languageid
      - name
    - people
  - Views

Property	Value
Name	title
Value	$books/title
Use column value	true
Omit when empty	false

Database archiving

- Strengths
  - Allows database-driven content to be archived
- Limitations
  - Does not preserve original look and feel
  - Immature with limited DB support
  - Can only capture snapshots
  - Requires webmaster participation

Transactional archiving

- Archives every materially-different response from a web server
- Allows transactions to be archived
- Tools available:
    - PageVault
    - Vignette WebCapture

Transactional archiving

- Strengths
  - Records what users actually experienced
  - Can collect static and dynamic content
- Limitations
  - Does not collect content which has not been requested
  - Possible impact on web server performance

The preservation challenge

To maintain the accessibility and authenticity of electronic records over time, across changing technical environments

- Accessibility depends upon a complex network of technical dependencies
- Authenticity derives from the significant properties of the record
- Preservation requires transformation

Preservation strategies

- Transform the source object to enable access within a new environment
    - Normalisation and migration
- Transform the means of access to enable continued access to the original object within a new environment
    - Emulation
    - Virtual computers

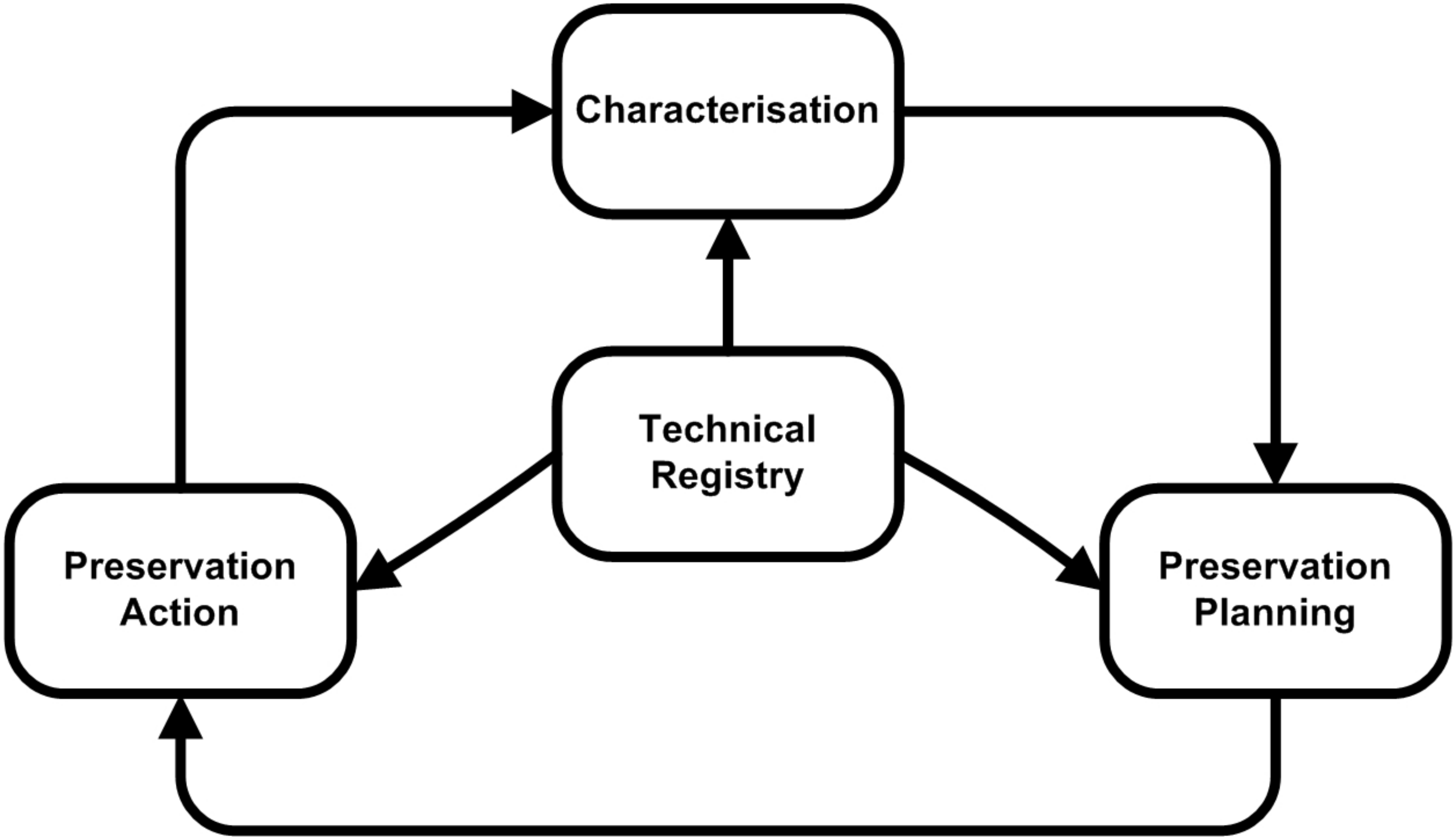
Preservation management

- Passive preservation
  - Preserving the bits
- Active preservation
  - Preserving the record
- Managing multiple manifestations

Passive preservation

- Security and access control
  - Physical and system security, user and system access control
- Integrity management
- Storage management
  - Media selection, management and refreshment, redundancy and backup
- Disaster recovery

Active preservation



Active preservation

- Characterisation
  - Identification
  - Validation
  - Property extraction
- Preservation planning
  - Risk assessment
  - Technology watch
  - Impact assessment
  - Preservation plan generation

Active preservation

- Preservation action
  - Enact preservation plan
  - Validate results – characterise transformed objects and compare significant properties with source objects

Preserving web content

- Preserving complex objects
  - Preserving relationships
  - Interconnected preservation actions
- Preserving behaviour
  - Input based
  - Output based

Legal issues

- Copyright
  - Rights to copy, adapt or reverse-engineer digital objects, or to circumvent DRM technologies for preservation may be constrained
- Regulatory compliance
  - Defining standards for legal admissibility

Preservation tools

- Web-specific
  - LOCKSS migration-on-demand
  - Virtual Remote Control
  - IIPC WARC format
- Generic
  - JHOVE
  - PRONOM and DROID

Next steps

- Develop UKWAC preservation requirements
- Input to infrastructure developments
- Review available tools
- Develop forward strategy