The National Archives

nationalarchives.gov.uk

The National Archives

Content and Context

Delivering coordinated UK web archives to user communities

Cathy Smith

21 July 2009

Agenda

- The study: “Delivering coordinated UK web archives to user communities”
- Findings
- Recommendations
- Summary

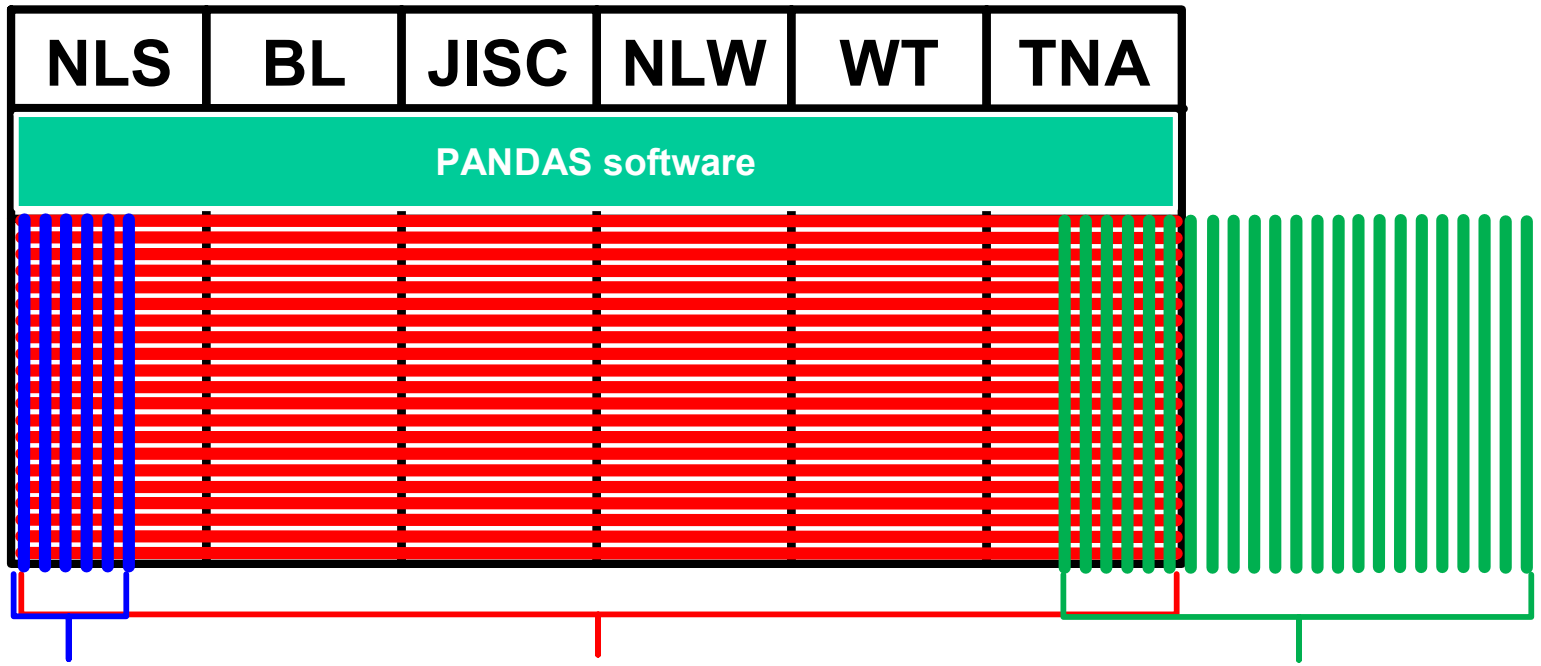
Workshop questions

From the workshop web page:

“What audiences should web archives anticipate and what does this mean for selection, ingest and preservation?”

“What will the web be like as an historical source, and what use will be made of archived web sites by future generations?”

Background: UKWAC



- Nearly 7,000 websites >20,000 snapshots
- About 10 TeraBytes

The study

- Funded as part of the JISC Preservation Programme
  - Formal tendering process
  - Contract awarded to Marc Fresko, of Inforesight Limited

- Conducted May – July 2009

Study approach

Documentation

Desk research

Interviews

Phone Interviews

Analysis

Consultation

Draft report

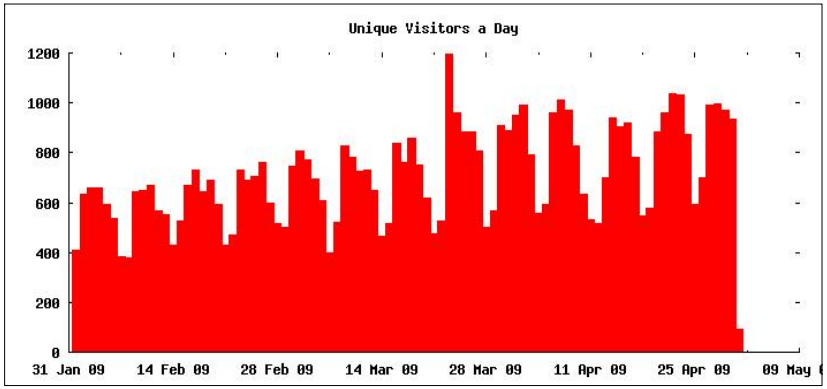
UK WEB ARCHIVING CONSORTIUM
www.webarchive.org.uk

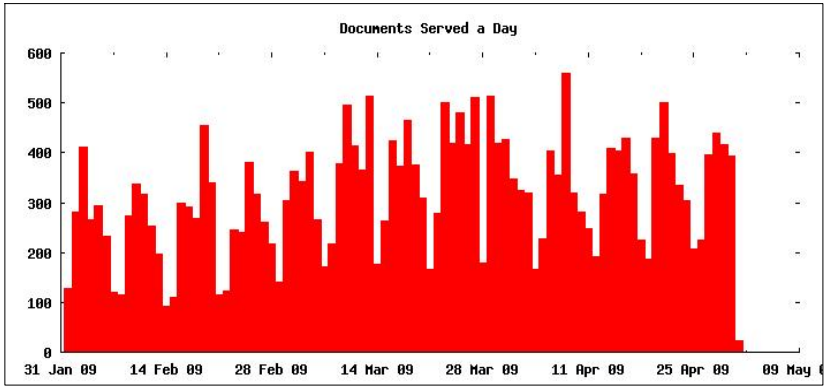
Findings: User base

- Web archives in general are being used by:
  - journalists and investigative reporters
  - litigants and detectives
  - civil servants
  - web designers
  - academic researchers
- UKWAC user base is small – so far!
  - Archives are still ‘young’
  - UK Web Archive is in ‘beta’
- Difficult to identify individual users of UKWAC archives

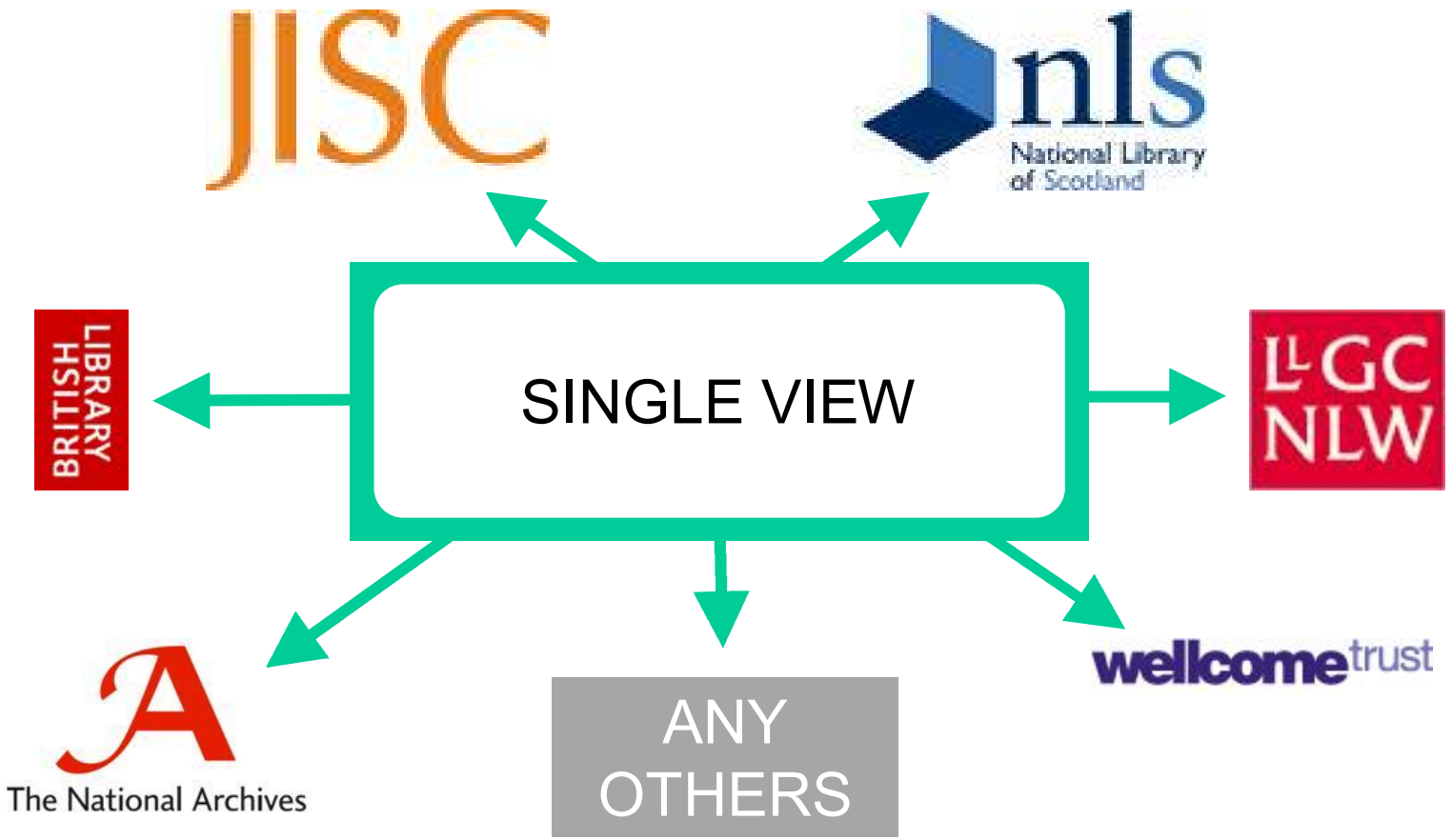
The National Archives

Findings: Usage is growing steadily





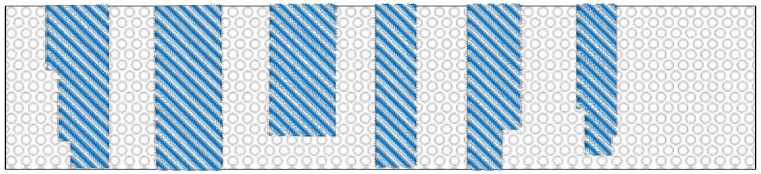
Findings: Searching across collections



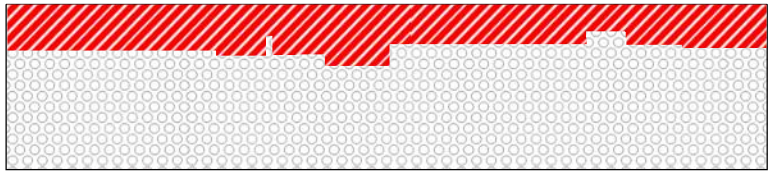
The National Archives

Findings: How to archive

- An estimated 5 million websites fall in scope
- UKWAC Partners currently harvest selectively about 7,000 websites
- Selective harvesting alone will not scale up satisfactorily
- Whole domain harvesting is needed
  - Through regulations
  - If ‘shallow’ then complemented by ‘deep’ selective harvesting



SELECTIVE



WHOLE DOMAIN

Recommendations:
National Collection of Websites

The National Archives

Recommendations: Individual collections

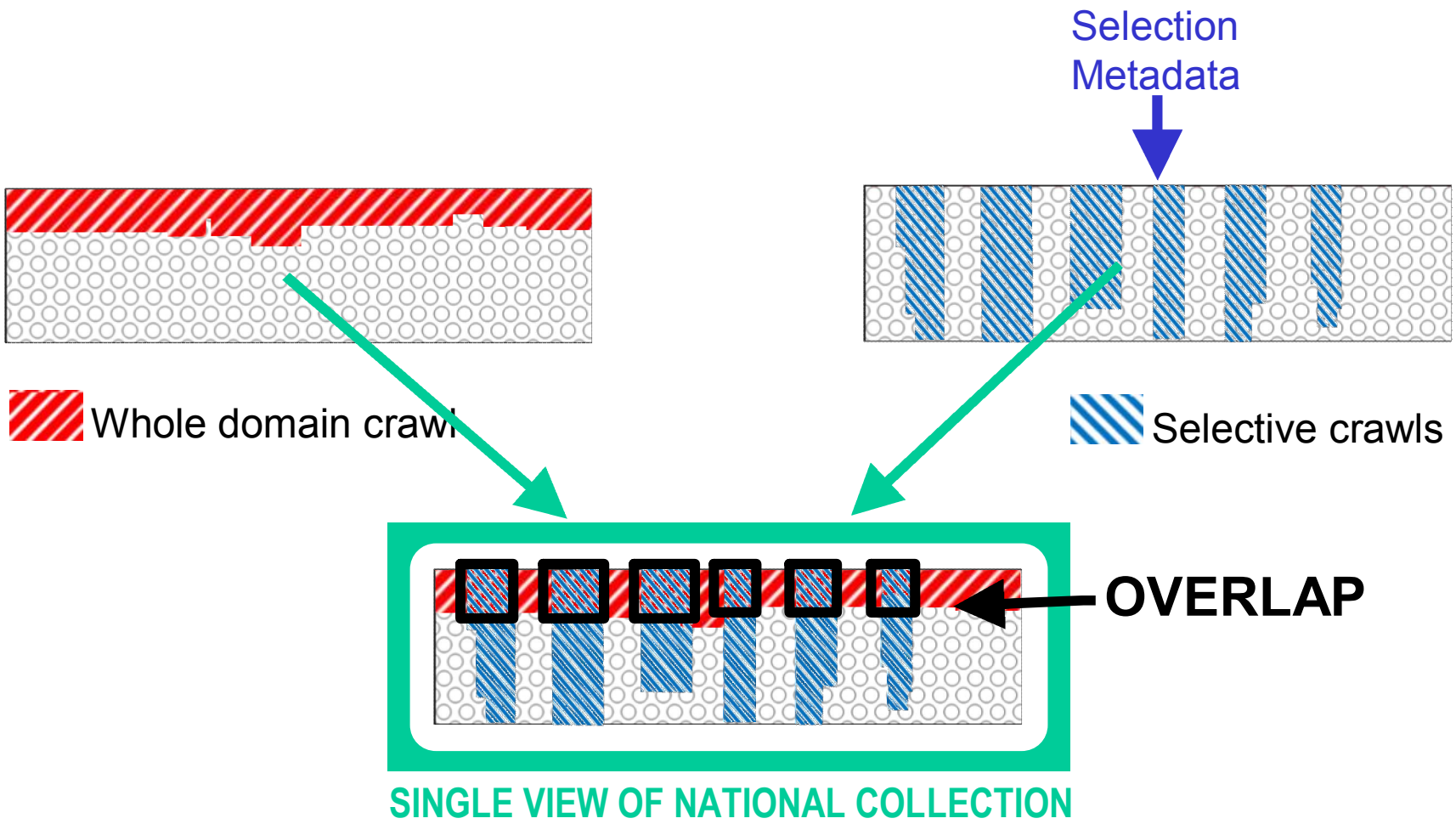
- Institutions continue to provide access to their individual collections, where appropriate, to:
  - support researchers who need access to only one sectoral or thematic collection
  - assure integrity of institutions' collections
  - allow integration with the institution's broader catalogue of other, non-web, holdings
- Institutions provide access to their individual collections in the context of the 'national' collection

The National Archives

Recommendations: Eliminate overlap

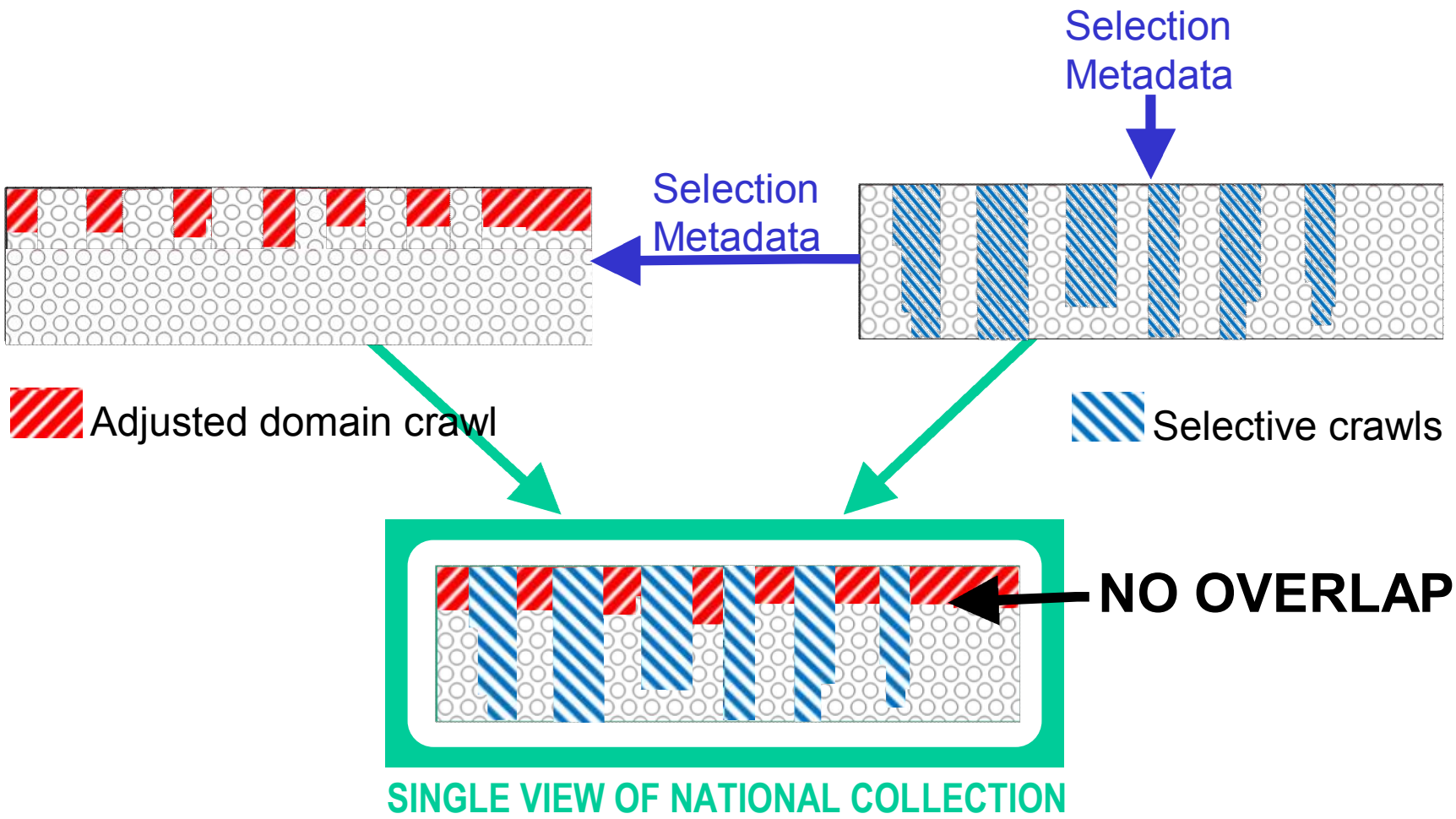
- Overlap in collecting arises due to:
  - thematic, legal and geographic remits
  - combining selective harvesting with whole domain harvesting
- Overlap produces duplication
  - Duplication is good for preservation, especially with separate repositories
  - Duplication is not good for users within a single repository (or a single view of several repositories)
- …so eliminate overlap within the ‘single view’ of the national collection

Overlap: illustration



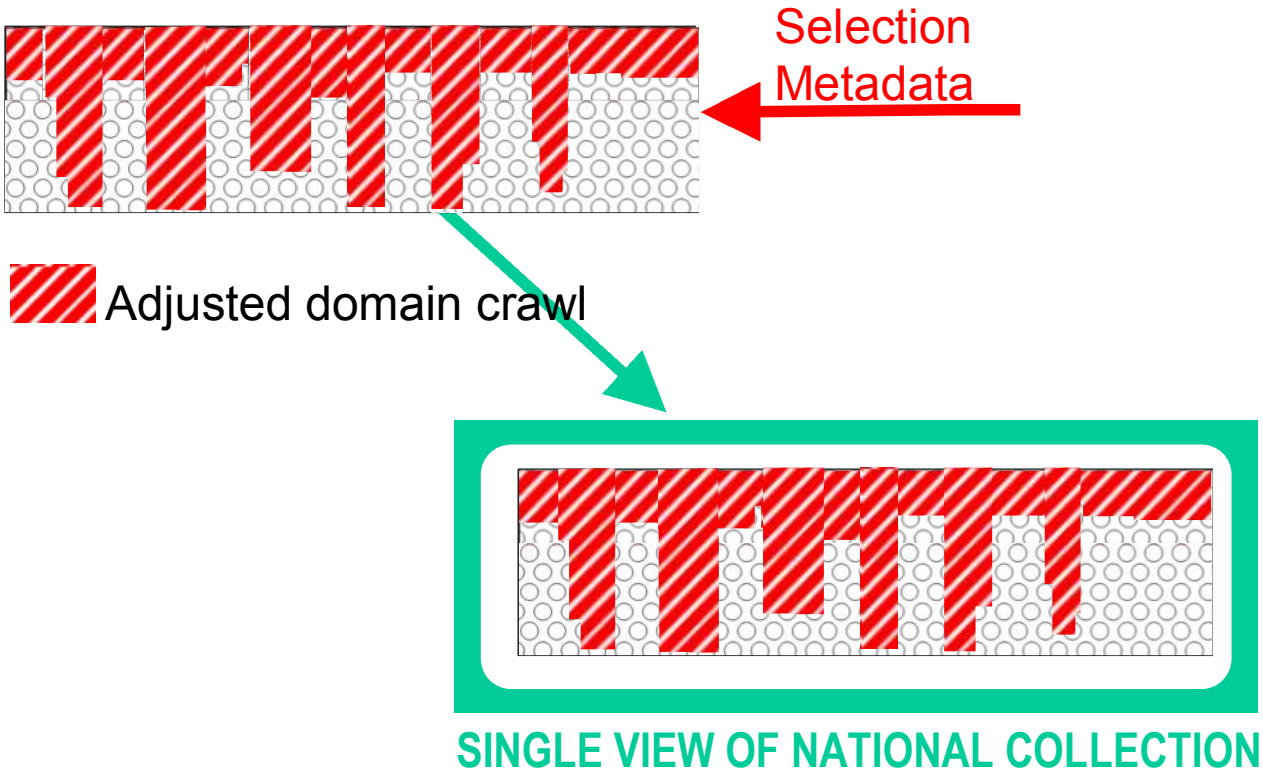
The National Archives

Overlap elimination option 1: Subtractive crawl



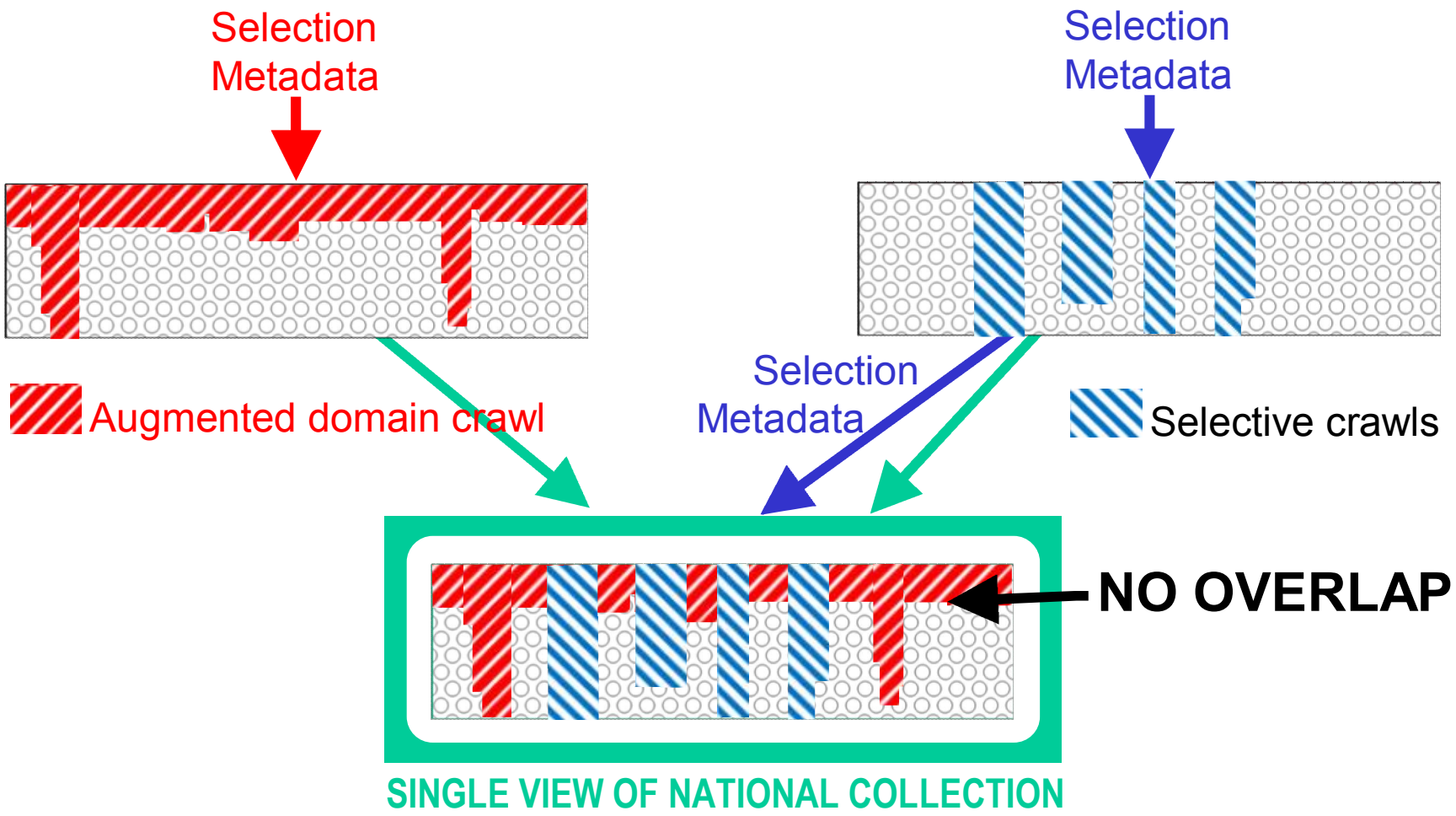
The National Archives

Overlap elimination option 2: Shared crawl



The National Archives

Overlap elimination option 3: Subtractive view



Recommendations: Coordination

- Need to agree the way forward and select the option(s)
  - Basis for decision could be clearer early-2010
- Whichever option is chosen, there will be a need for coordination to include:
  - Agreeing inter-institutional protocols for sharing collection development policies
  - Defining an agreed minimum metadata standard
  - Developing technical interfaces
- UKWAC should continue to provide national coordination
  - Success dependent on outcome from current discussions on repositioning the Consortium

Summary

- UKWAC has come a long way in a short time
- We cannot continue to scale up the current selective web archiving activities – we need regulations that permit whole domain harvesting as well
- There is a good argument for a ‘single view’, or what we can call a National Collection of Websites – and we can see possible ways to achieve that goal