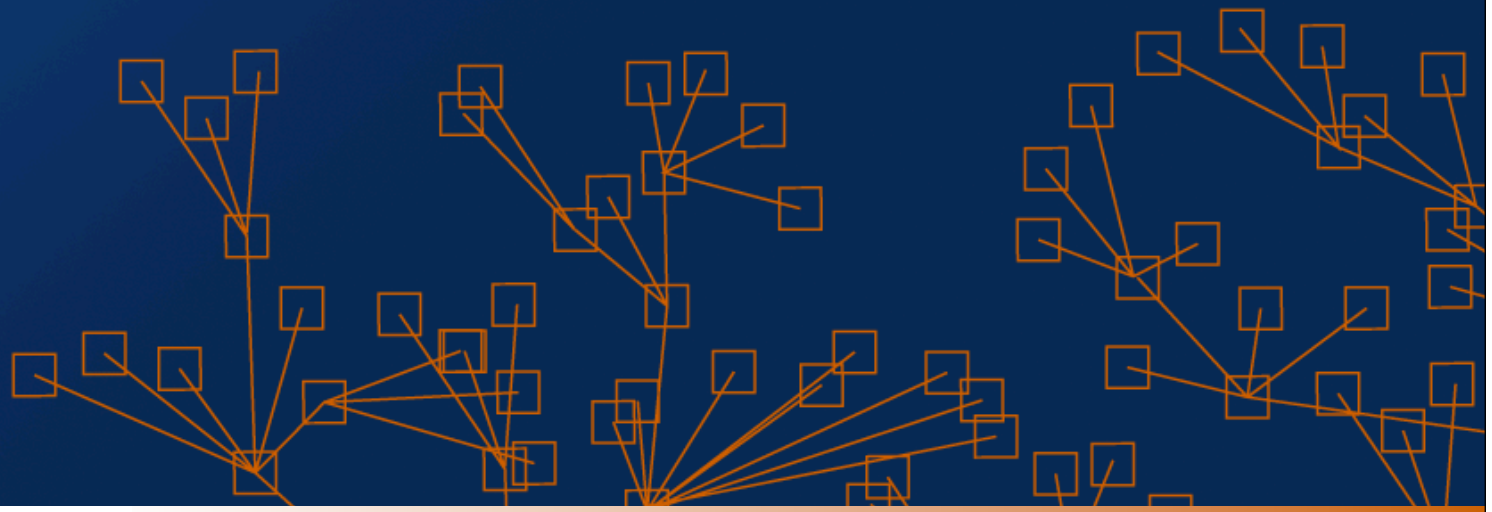
IIPC
Web Archiving Toolset

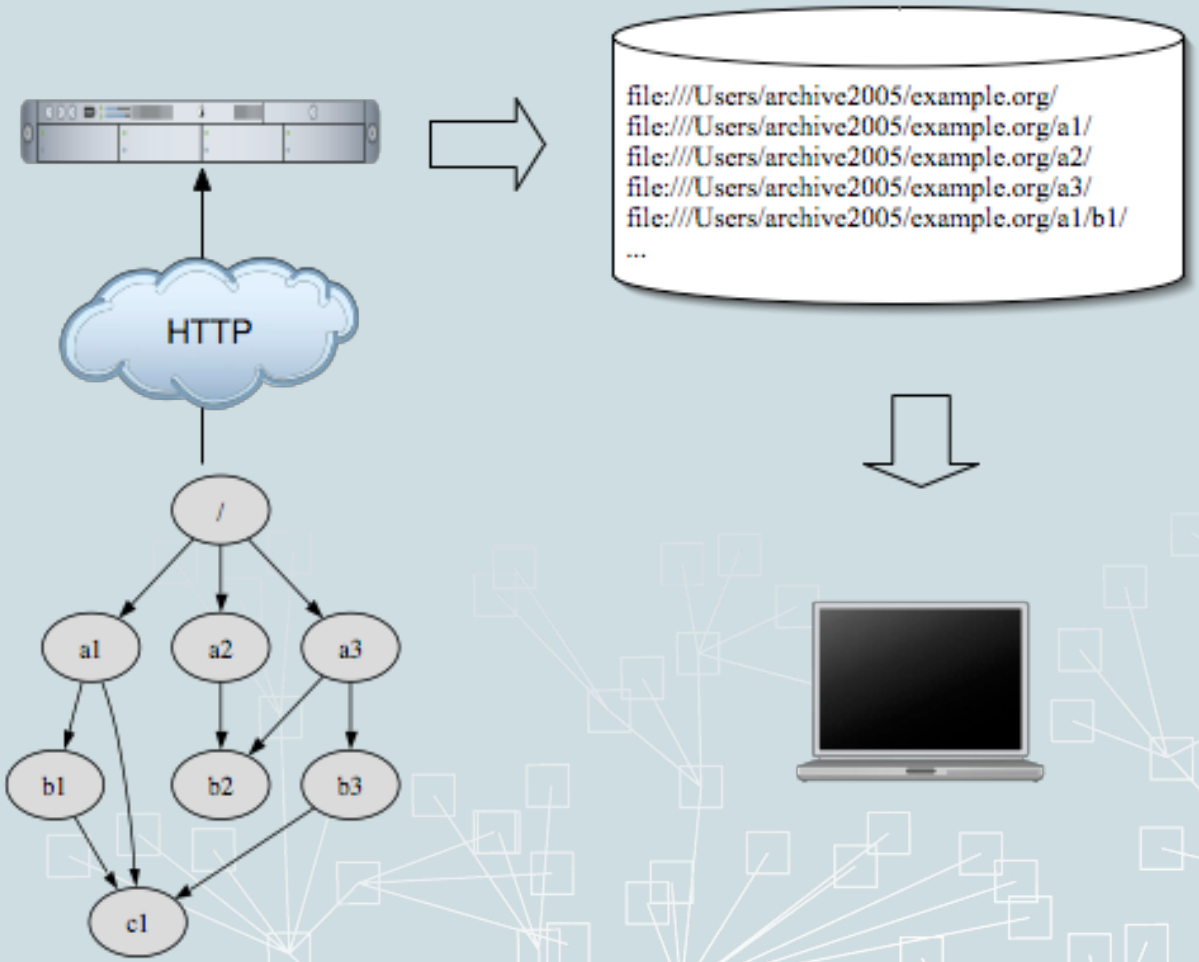


netpreserve.org
international internet preservation consortium

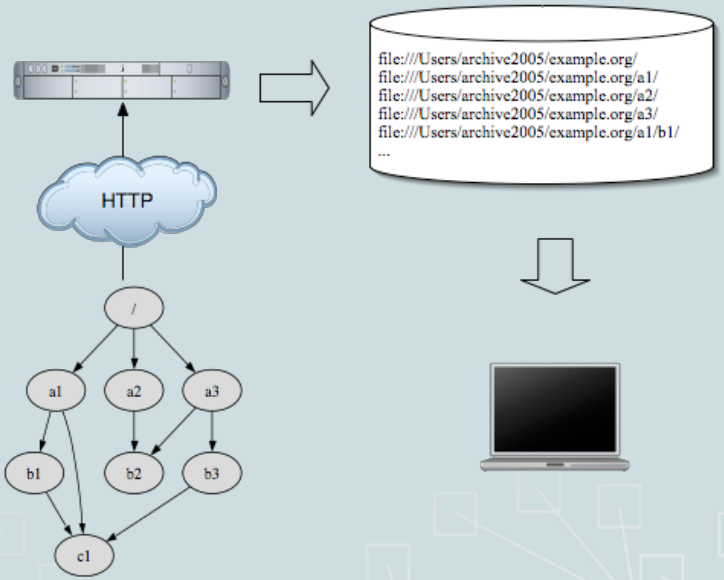
3 types of Web Archives

- Local File System
- Web Served Archives
- Non Web Archives

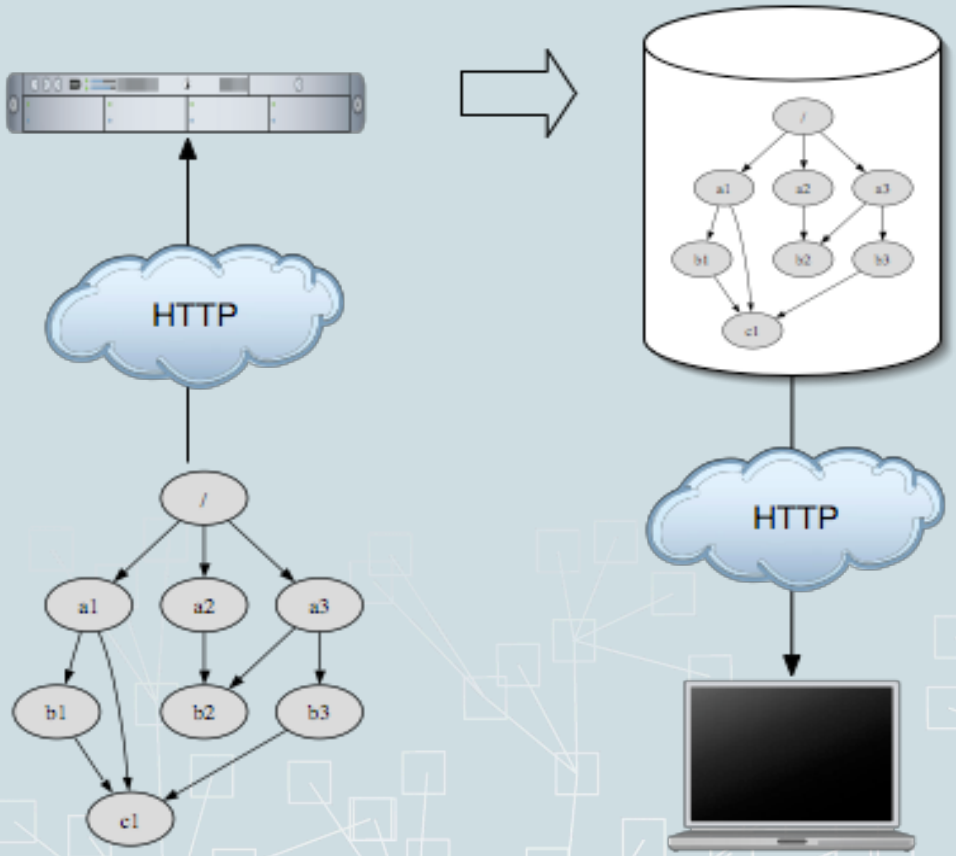
Local File System



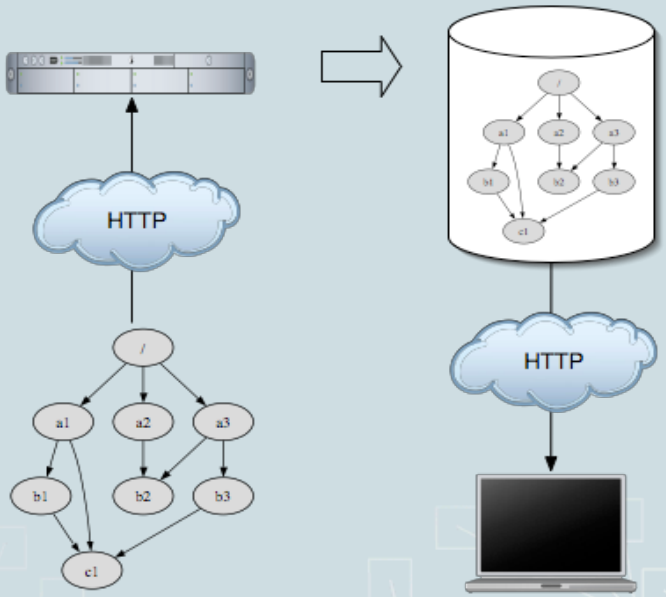
- How
  - All links are converted into relative ones.
  - Hypertext Navigation is done directly on the local file system.
- Application
  - Single site archiving and small and middle scale archiving.
  - Tools: Web site copier (like HTTrack)
- Comments
  - Simple to implement but does not scale up.
  - Requires renaming and limited re-organization of content for navigation.
  - Need a file system level management of archived collection and versions.



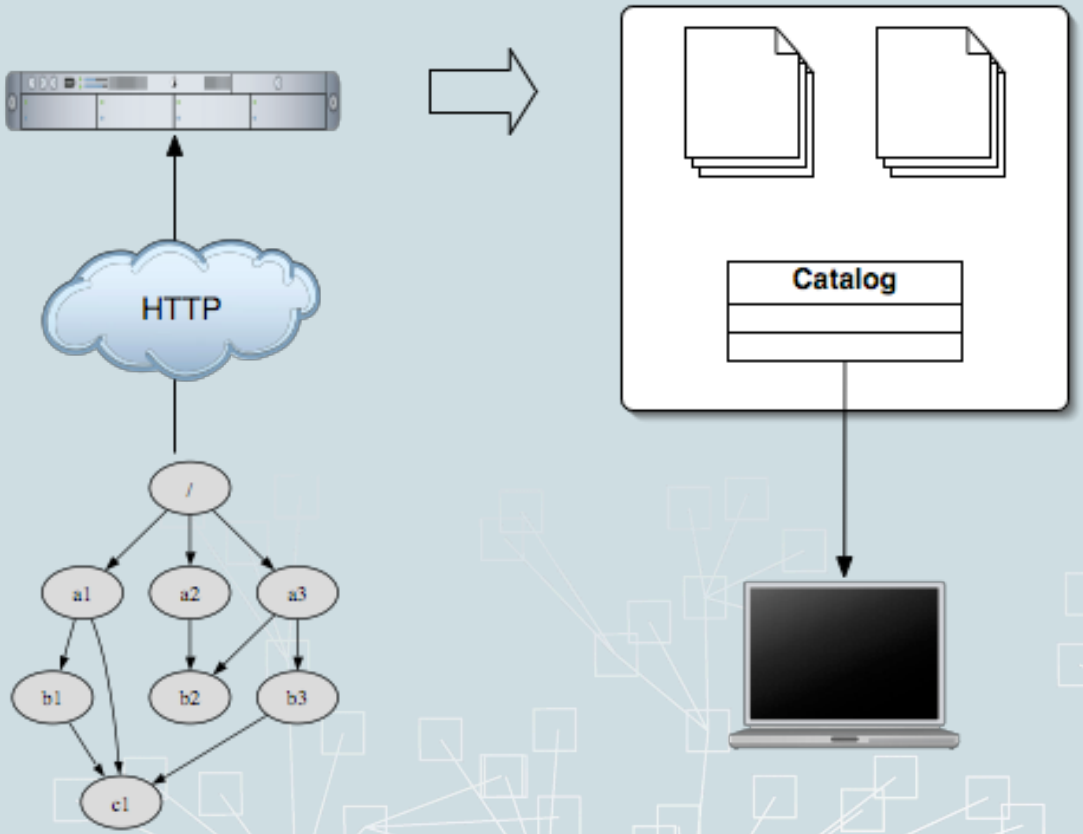
Web Served Archives



- How:
  - A Web server is set up for access through which documents are served.
  - Hypertext navigation is closed to the original one.
- Application
  - middle and large scale archiving.
  - Archiving Crawler (like heritrix) and index system for warc files.
- Comments
  - Authenticity, scalability but more difficult to implement



Non Web Archives



Non web archives

- Documents are extracted from the original hypertext content and re-organized along a different logic.
- Applies for specific (non-web) collections archiving.
- Enable integration in traditional OPAC or other local collections organizations.
- Lost of hypertext structure. Can only be applied for isolated, non-web documents.

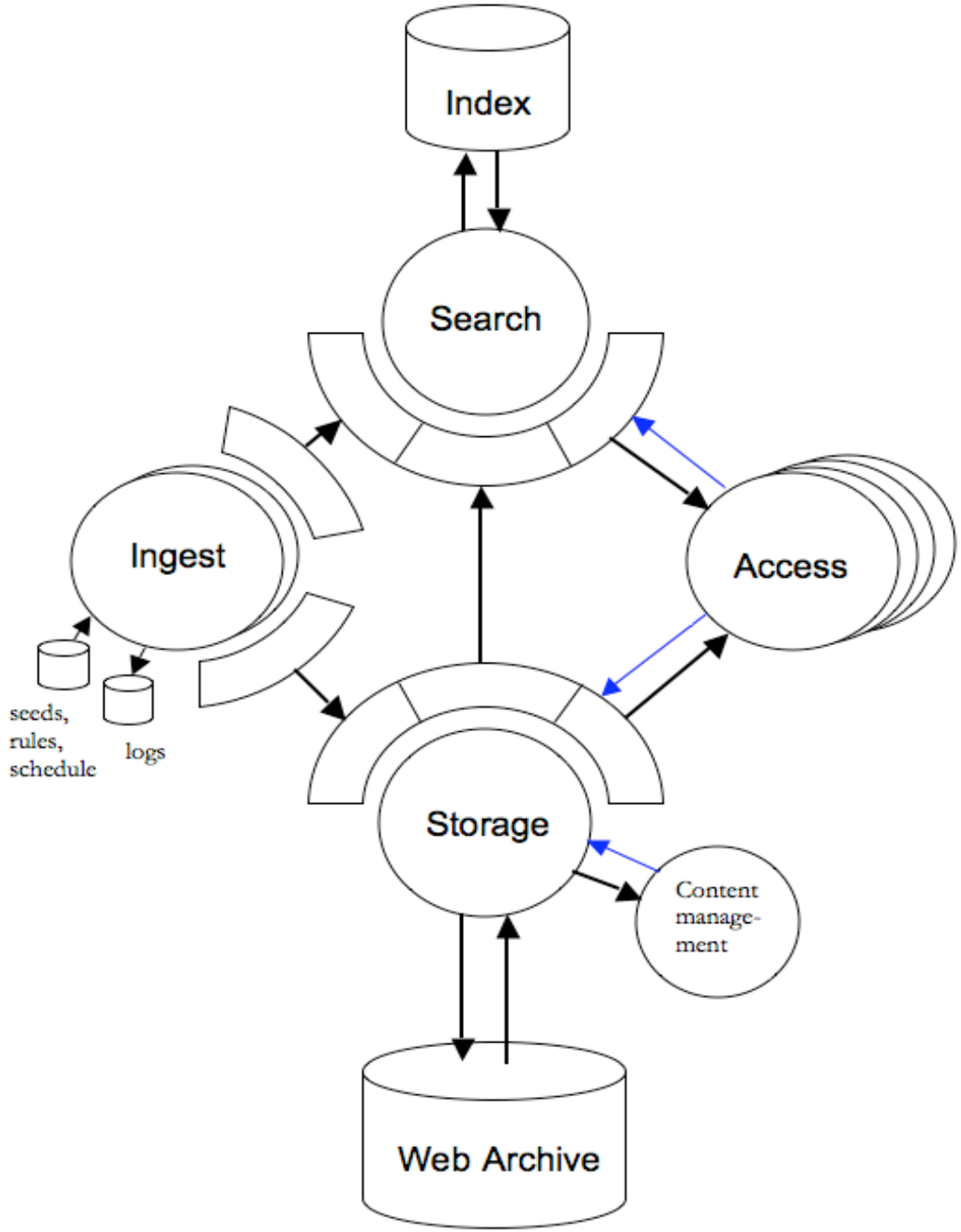
IIPC Toolset

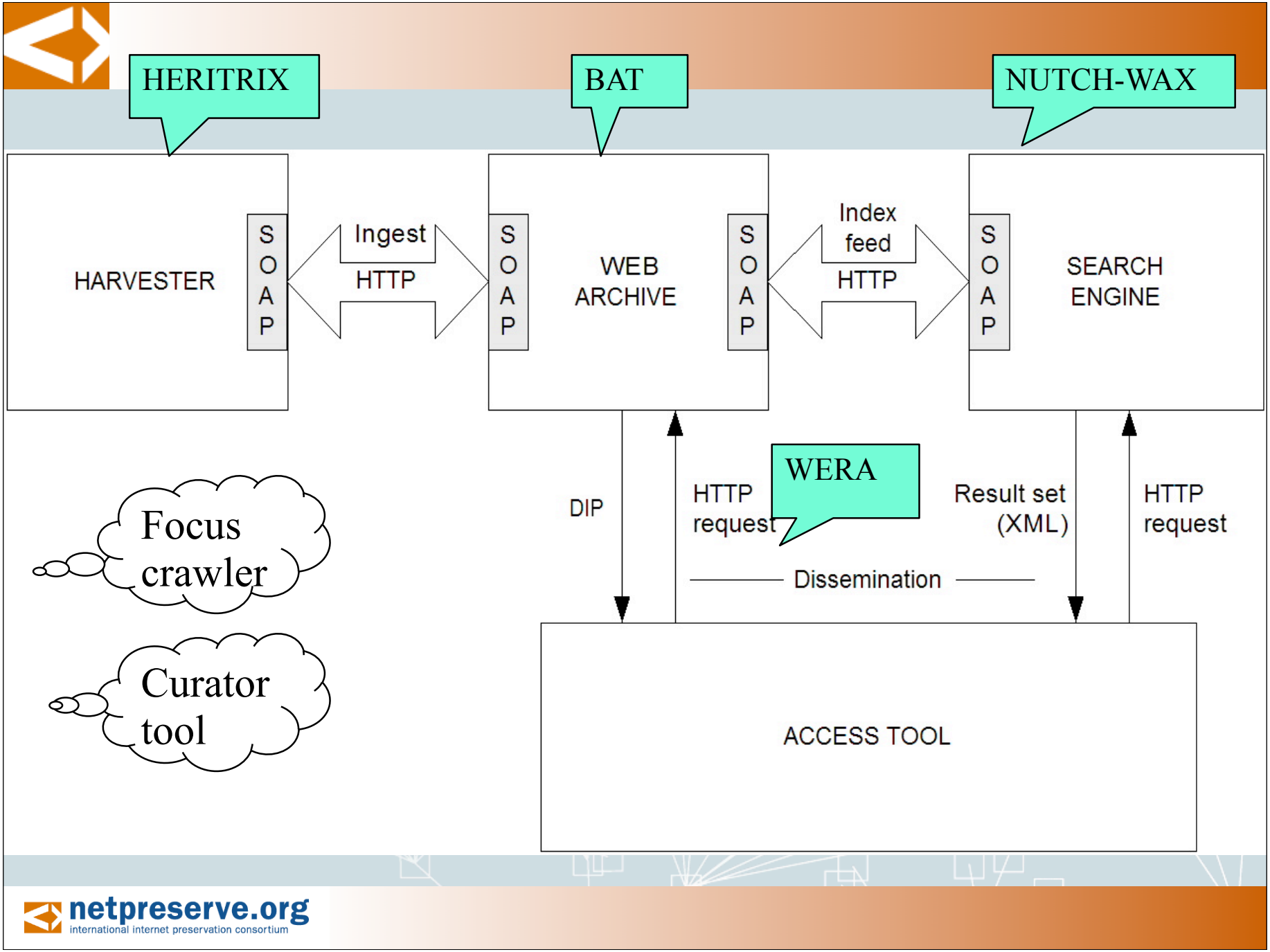
- For middle and large scale archiving
  - Web served archives model
- To implement IIPC standards
- Ensure compatibility and 'plugability' of the resulting collection
  (global cross-access to collections)

How?

- Common specifications
- IIPC joint project (at least 2 partners)
- Open source (GPL)

Architecture of tools for Web Archives




HERITRIX
BAT
NUTCH-WAX
HARVESTER
SOAP
Ingest
HTTP
SOAP
WEB ARCHIVE
SOAP
Index feed
HTTP
SOAP
SEARCH ENGINE
DIP
HTTP request
WERA
Result set (XML)
HTTP request
Dissemination
Focus crawler
Curator tool
ACCESS TOOL

Enabling the web archives grid!

IIPC site: www.netpreserve.org

Web Archive information list: webarchive@cru.fr

julien@netpreserve.org