WERA
A Web Archive Collection Access Tool

Julien Masanes and Michael Stack

Vienna September 22nd, 2005

What is WERA?

- Web Archive Collection (WAC) access tool (Elsewhere termed a WAC *viewer* application*)*.
  - Like the Internet Archive Wayback Machine but...
    - Supports full text search of Web Archive Collections.
      - Time dimension displaying search results.
    - Open Source, distributable.
- Development sponsored by IIPC.
- Subset of the NWA Toolset.
  - Collection of tools that facilitates web collection access.
  - NWA = Nordic Web Archive, Nordic National Libraries forum; exchange of experience harvesting and archiving web content.
- First release August, 2005.

IIPC
international internet preservation consortium

Example: Query

WERA

Match: All words | Query: henrik ibsen | Search | Help

Year (from - to)

Total number of versions found : **578**. Displaying URL's **1-10**

1. Portrett av Henrik Ibsen (http://www.nb.no/html/portrett_av_henrik_ibsen.html)
(**...** Portrett av **Henrik Ibsen Henrik Ibsen** Konservering av Catilina Mange av oss forestiller oss gjerne **Henrik Ibsen** som en gammel mann. I 1849 da **Ibsen** skrev sitt debutstykke var han ennå ikke fylt 21 år. Dette fotografiet er et av de tidligst kjente portrettene av **...** harmonisk Udtryk." Due bemerket også "et Glimt i hans smukke Øine, der gjorde Indtryk på mig". Konserveringsatelieret Konservering av Catilina **Henrik Ibsen** Apoteket i Grimstad **...**)
Number of versions satisfying query / total number of versions : 5/5
Timeline | Overview

2. Henrik Ibsen (http://www.nb.no/html/henrik_ibsen.html)
(**...** kjendiser. Bildesamlingen eier de fleste glassplatene etter Szacinski. Disse er deponert på Oslo bymuseum. Originalen måler 16x22 cm. Signatur: Ibf2a1001. **Henrik Ibsen** på italiensk postkort fra ca. 1910. "Amalfi - Hotel de la Lune hvor **Henrik Ibsen** skrev Et Dukkehjem i 1879". Originalen måler 9x14 cm. Signatur: Ibf2a1001. **Henrik Ibsen** fotografert i Gustav Borgens (1865-1926) atelier i 1900. Bildet har påskriften "Billedet som ikke blev benyttet i Samlede Værker**Henrik ...**)
Number of versions satisfying query / total number of versions : 11/11
Timeline | Overview

Example: Result

NWA Browser - Mozilla

File Edit View Go Bookmarks Tools Window Help

Uri: http://www.nb.no/html/portrett_av_henrik_ibsen.html Go Search Go

Viewing version 1 of 5
Des. 8th 2004, 15:14

Nov. 23rd Des. 23rd

Resolution: Days Auto: Help

WERA... External links, forms, and search boxes may not function within this collection. [hide]

Nasjonalbiblioteket

Søk Innhold Kontakt Nyheter Databaser

Samlinger
Tjenester
Ut i verden
Utstillinger
Om NB

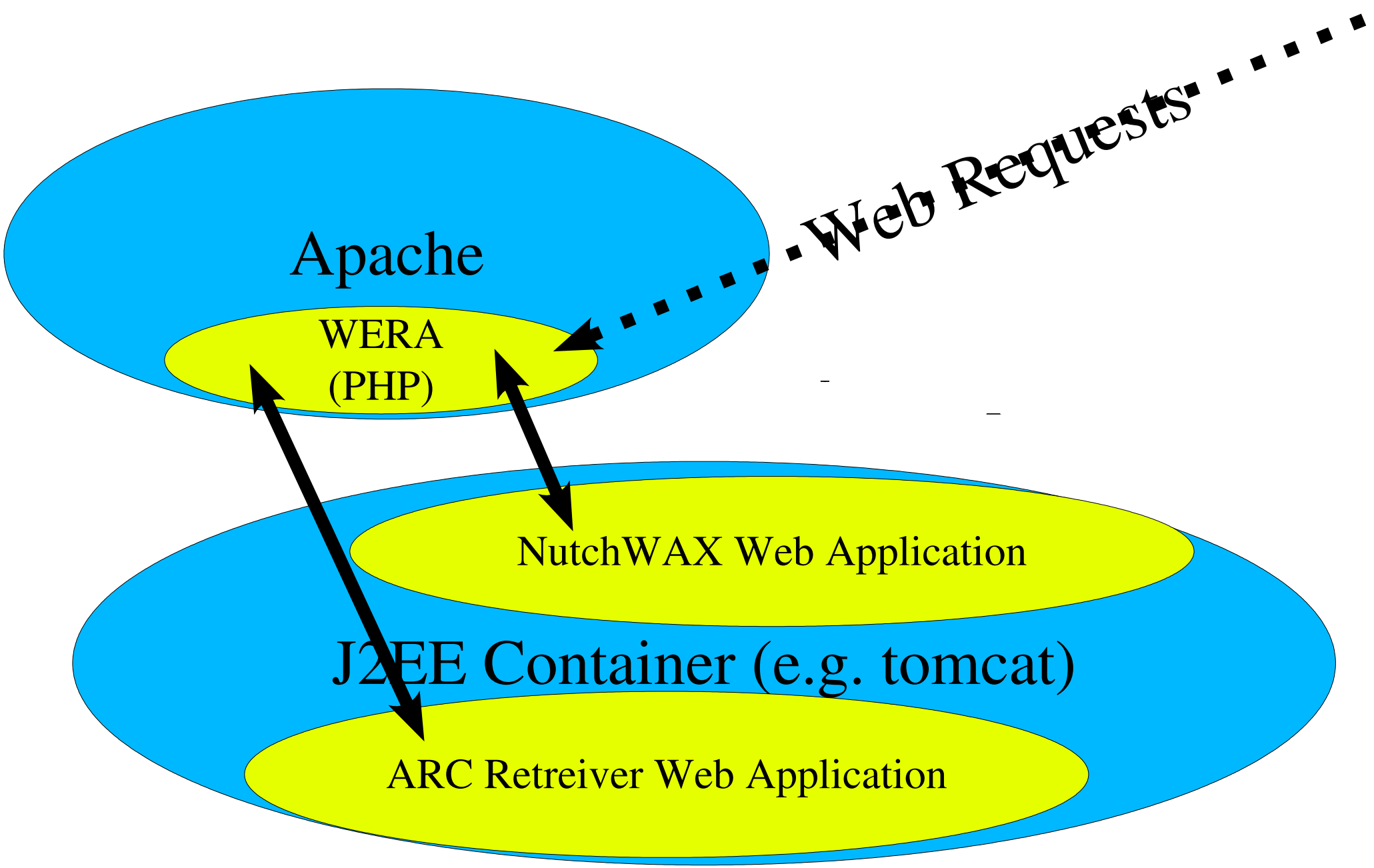
Henrik Ibsen

Konservering av Catilina

Konserveringsatelieret

Konservering av Catilina
Henrik Ibsen
Apoteket i Grimstad
Henrik Ibsens losji
Fra idé til ferdig verk
Annonse i Morgenbladet
Første utgave trykt i 1850
Catilina innkjøpt av UB i 1901
2 skrivehefter
3 forskjellige papirkvaliteter
Tilstandsvurdering
Visuelle og kjemiske undersøkelser
Nedbrytningsforløp og - faktorer
Kjemiske inngrep
Innlimte tillegg
Identifisering av klebestoffet
Fjerning av klebestoffet

Architecture



WERA Detail

- Built against search engine abstraction layer.
  - Previously, implementations for FAST and Lucene.
  - A Nutch implementation was added.
    - Or, to be more precise, NutchWAX (Nutch + Web Archive eXtensions).
      - http://archive-access.sourceforge.net/projects/nutch
  - Builds query request (An URL) passed to NutchWAX.
  - Transforms the Nutch OpenSearch XML result set into WERA internal format (PHP arrays).
  - WERA + NutchWAX = complete WAC access solution
    - Currently, for small collections only (< 1-10
    - million document repositories and < 1 hit a second)
- Page links rewritten using IA's client side javascript.
- Java based (very) user friendly GUI installer:
  - Using AntInstaller: http://antinstaller.sourceforge.net/

Future Development

- Fix Bugs and Implement Feature Requests.
  - See list at archive-access.sourceforge.net/projects/nutch. Includes:
    - ARC Retreiver to go against distributed repository (Currently expects ARCs to be local).
    - Improved link rewriting (Just as for IA Wayback, currently imperfect).
    - Support for proxied requests.
    - Implement advanced search screens to go against NutchWAX.
  - Next release synchronized with NutchWAX release: October, 2005.
- Long term:
  - Redo WERA frontend in Java as a web application (No need of Apache, matches ARC Retreiver and NutchWAX web applications).
  - Address larger scales (Higher hit rates, larger collections).