

# Internet memory

Julien Masanès  
European Archive



# Why archiving the web

- today the web is the main publishing medium (tenth of billions of pages)
- all aspect of cultural, scientific, mundane production have traces on the web
- it is a unique source of information on modern societies



# ephemeral

- the half life of a page is less than 2 years
- even academic publishing on the web is not stable

Study	Resource type	Resource half-life
Koehler (1999 and 2002)	Random Web pages	about 2.0 years
Nelson and Allen (2002)	Digital Library Object	about 24.5 years
Harter and Kim (1996)	Scholarly Article Citations	about 1.5 years
Rumsey (2002)	Legal Citations	about 1.4 years
Markwell and Brooks (2002)	Biological Science Education Resources	about 4.6 years
Spinellis (2003)	Computer Science Citations	about 4.0 years (p. 74)



# What has changed?

- (stable) & (discrete objects)
- time
- (publishers)
- technique
  
- online
- ready for automatic processing



# Time

- Update frequency (serial)
- Notification (rss atom etc.)



Internet Archive Wayback Machine

http://web.archive.org/web/\*/http://www.colbud.hu/ Google

Internet Archive Wayb... Collegium Budapest - ... Collegium Budapest - ... Collegium Budapest - ... Collegium Budapest - ...

INTERNET ARCHIVE  
**WayBackMachine**

Enter Web Address:  All  [Adv. Search](#) [Compare Archive Pages](#)

Searched for <http://www.colbud.hu/> **56 Results**

\* denotes when site was updated.

**Search Results for Jan 01, 1996 - Jun 22, 2005**

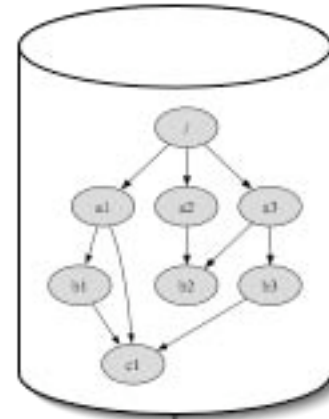
1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
0 pages	0 pages	3 pages	3 pages	3 pages	4 pages	5 pages	19 pages	19 pages	0 pages
		<a href="#">May 19, 1998</a> * <a href="#">Dec 05, 1998</a> * <a href="#">Dec 12, 1998</a>	<a href="#">Feb 08, 1999</a> <a href="#">Feb 18, 1999</a> <a href="#">Apr 29, 1999</a>	<a href="#">May 20, 2000</a> <a href="#">Aug 29, 2000</a> <a href="#">Oct 21, 2000</a>	<a href="#">May 15, 2001</a> * <a href="#">Jul 20, 2001</a> <a href="#">Sep 17, 2001</a> <a href="#">Sep 25, 2001</a>	<a href="#">Jan 21, 2002</a> * <a href="#">May 28, 2002</a> <a href="#">Sep 20, 2002</a> <a href="#">Sep 26, 2002</a> <a href="#">Nov 25, 2002</a> *	<a href="#">Jan 30, 2003</a> <a href="#">Feb 10, 2003</a> <a href="#">Mar 24, 2003</a> <a href="#">Apr 04, 2003</a> <a href="#">Apr 25, 2003</a> <a href="#">May 06, 2003</a> <a href="#">May 24, 2003</a> <a href="#">May 25, 2003</a> <a href="#">May 28, 2003</a> <a href="#">Aug 01, 2003</a> * <a href="#">Aug 07, 2003</a> <a href="#">Sep 23, 2003</a> * <a href="#">Oct 08, 2003</a> <a href="#">Oct 09, 2003</a> <a href="#">Nov 10, 2003</a> <a href="#">Nov 19, 2003</a> * <a href="#">Nov 23, 2003</a> <a href="#">Nov 28, 2003</a> <a href="#">Dec 16, 2003</a>	<a href="#">Jan 01, 2004</a> <a href="#">Jan 19, 2004</a> * <a href="#">Feb 25, 2004</a> * <a href="#">Mar 25, 2004</a> * <a href="#">Mar 26, 2004</a> <a href="#">Jun 03, 2004</a> * <a href="#">Jun 06, 2004</a> <a href="#">Jun 23, 2004</a> <a href="#">Jun 24, 2004</a> <a href="#">Jun 27, 2004</a> <a href="#">Jun 29, 2004</a> <a href="#">Jul 01, 2004</a> <a href="#">Jul 14, 2004</a> <a href="#">Oct 26, 2004</a> <a href="#">Nov 02, 2004</a> <a href="#">Nov 06, 2004</a> <a href="#">Nov 07, 2004</a> <a href="#">Nov 14, 2004</a> <a href="#">Nov 25, 2004</a>	

[Home](#) | [Help](#)

Copyright © 2001, [Internet Archive](#) | [Terms of Use](#) | [Privacy Policy](#)

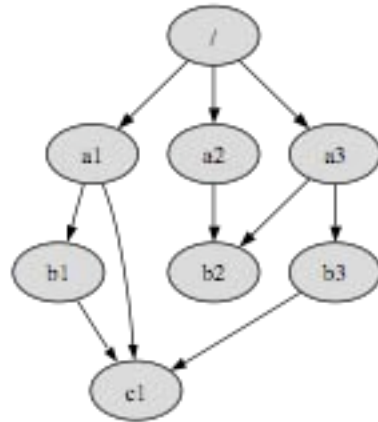


Crawler



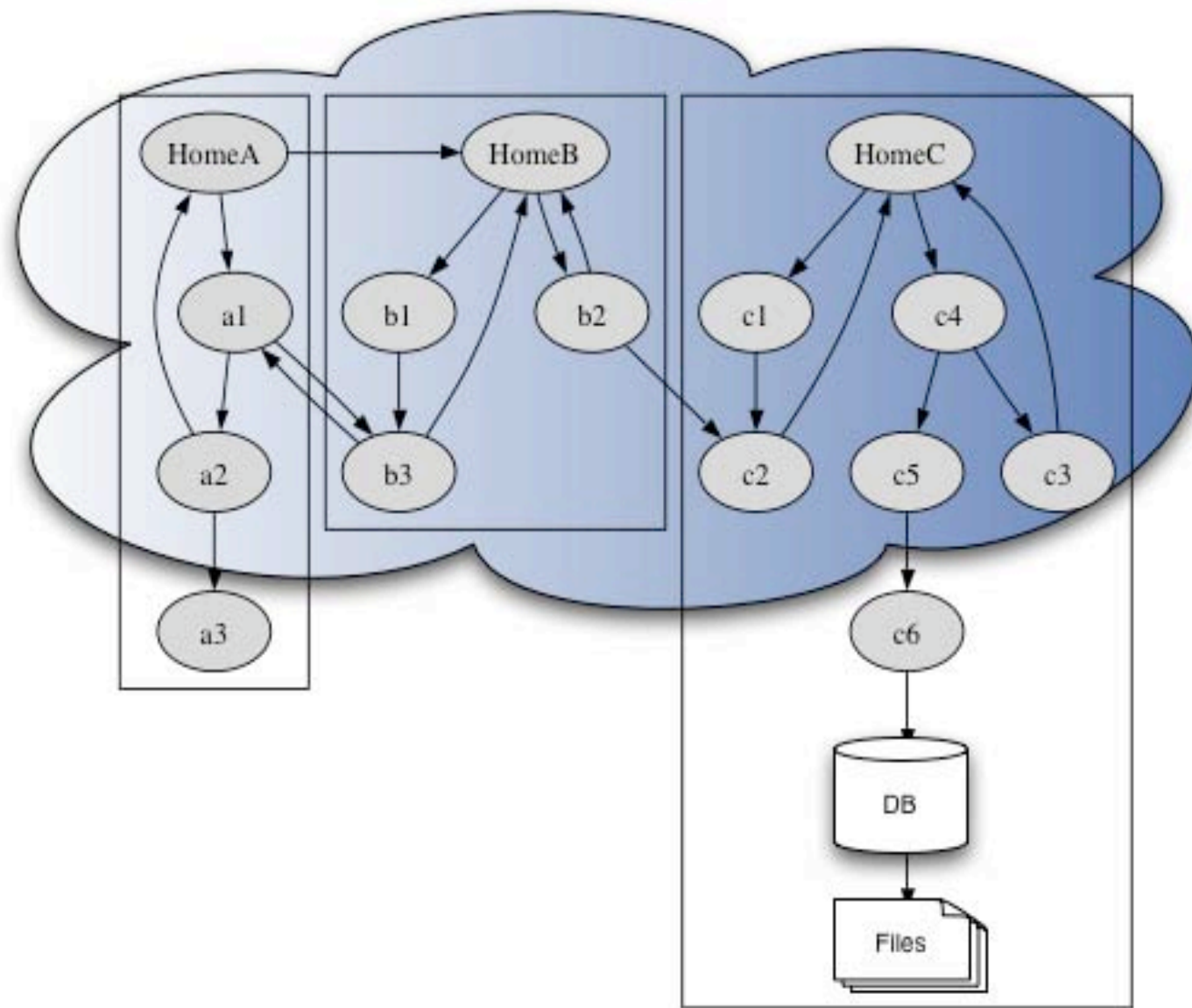
Archive

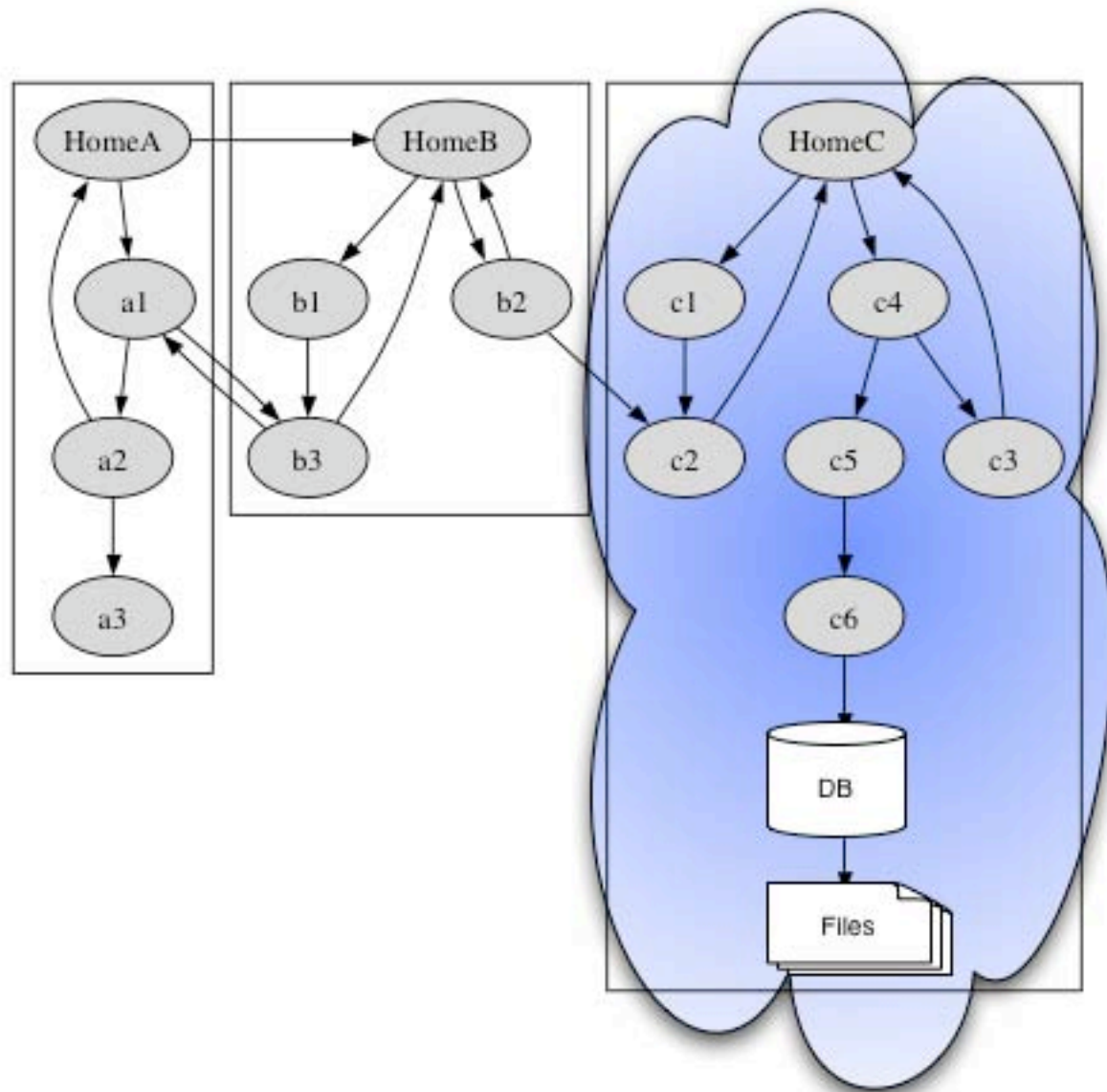
Web

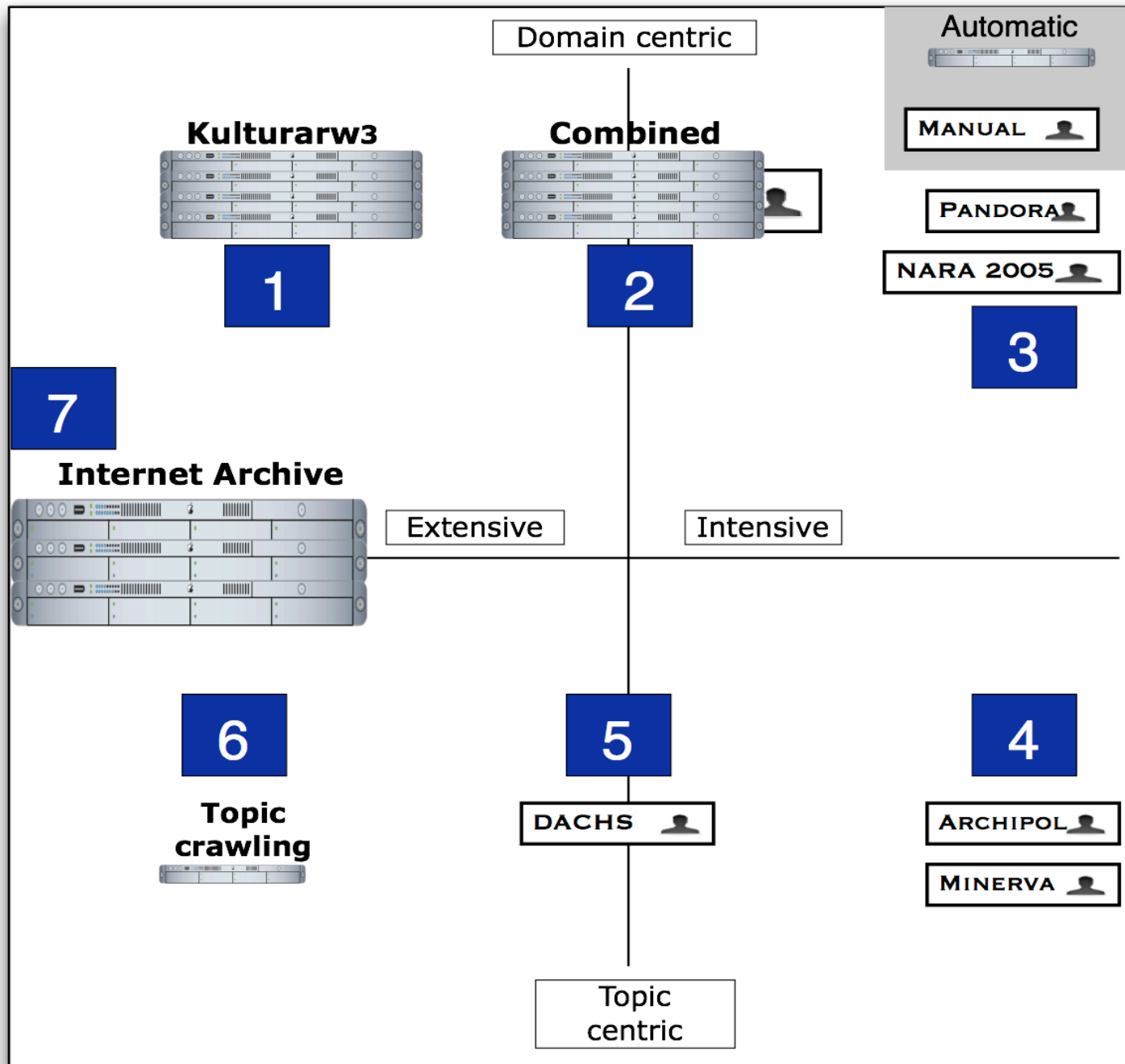


User

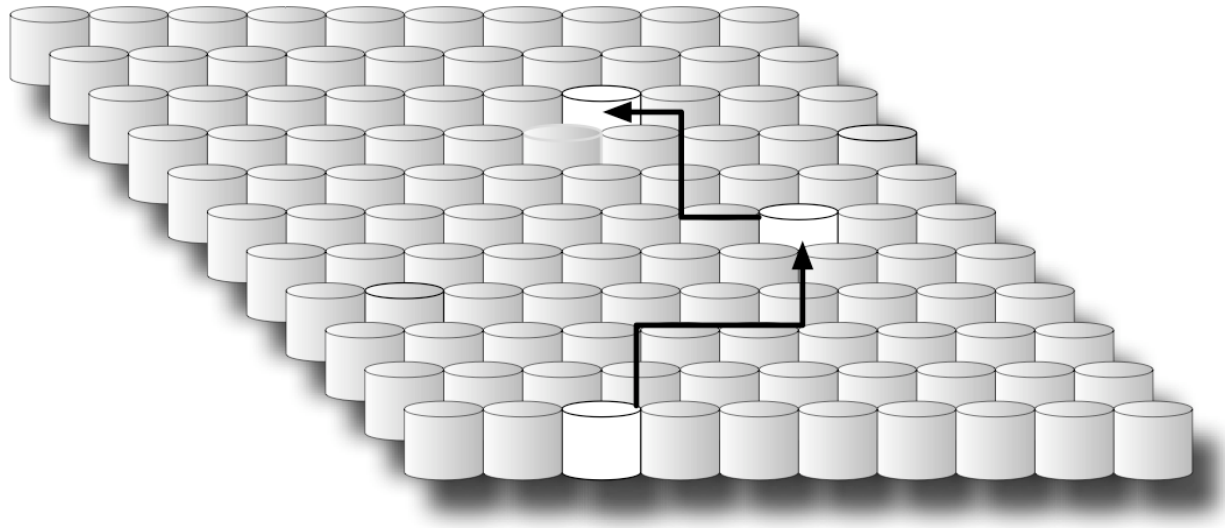






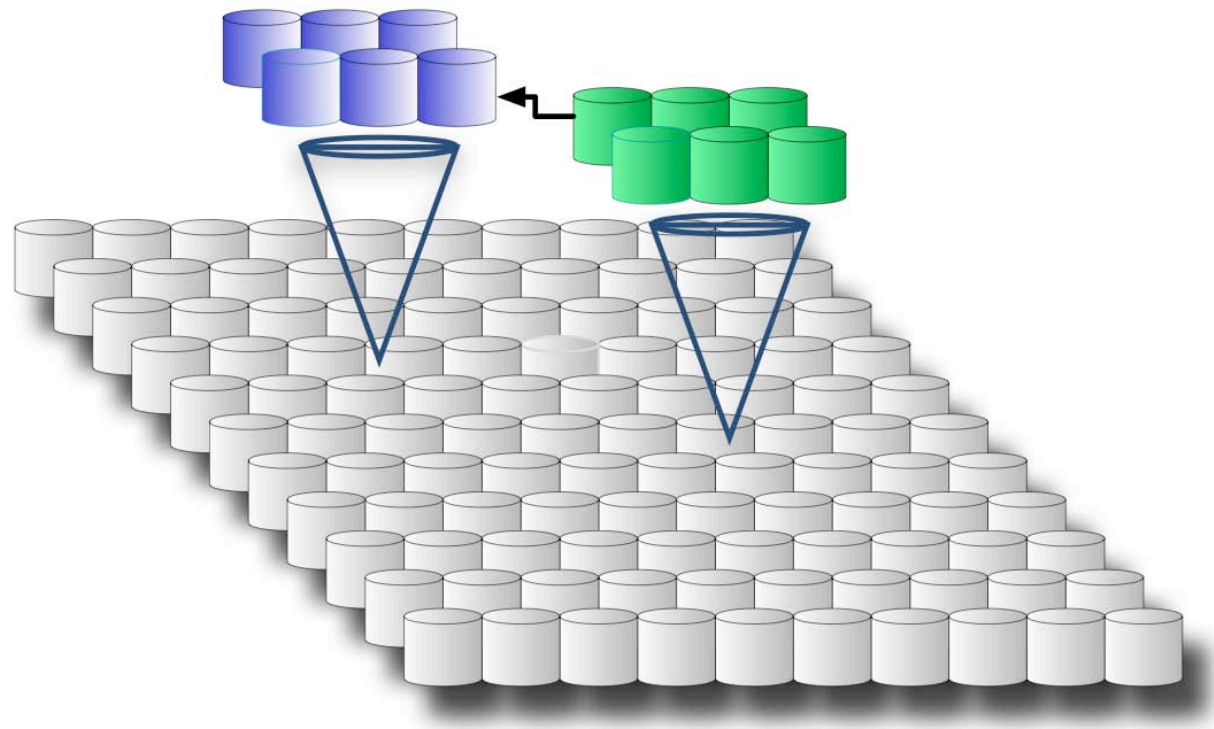


# www as a grid of servers



# Web archives grid

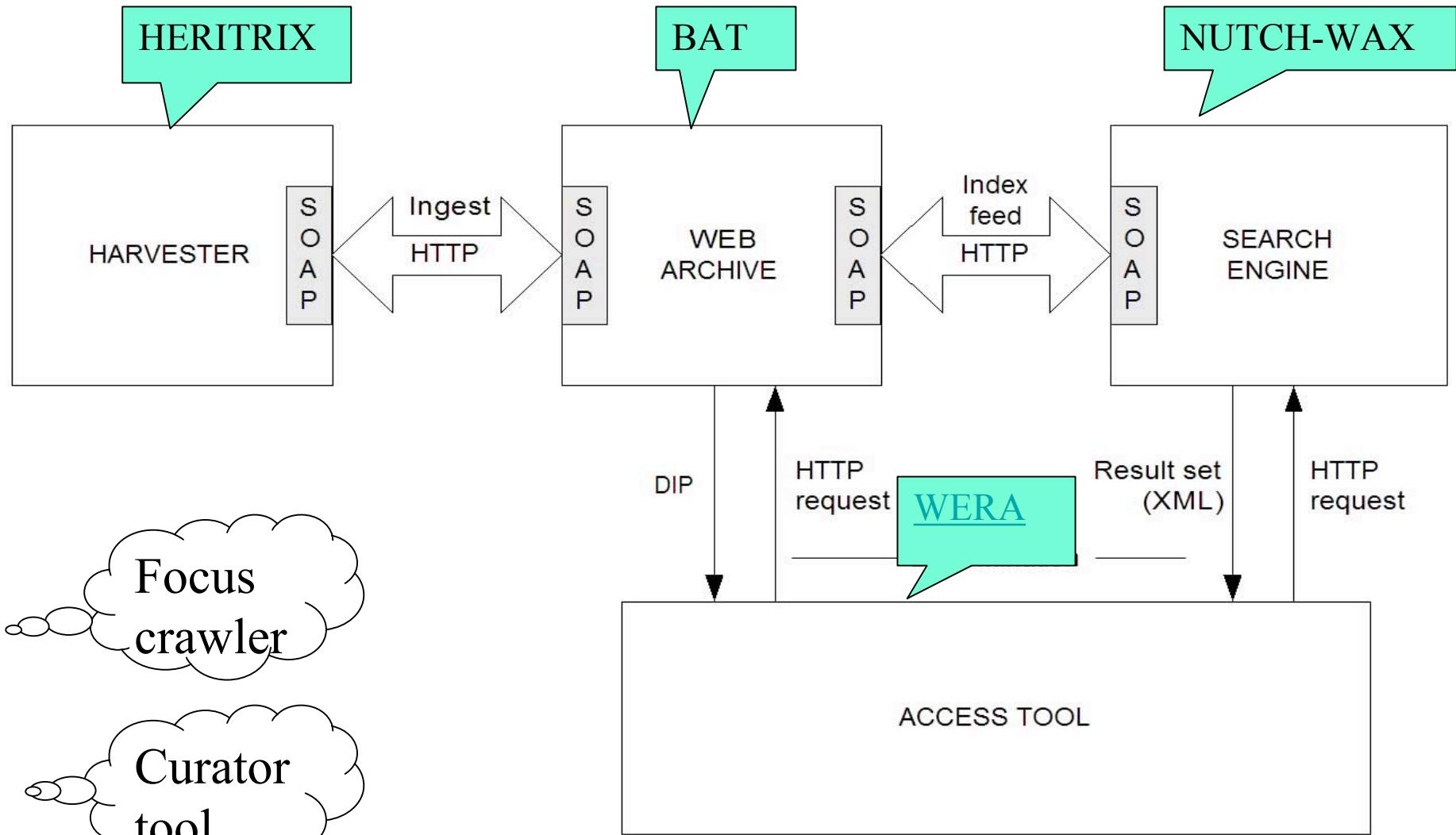
(by redirection)



# IIPC: standards and tools for the Web archives grid

- Standards
  - Architecture
  - Storage format (WARC)
  - Metadata





Focus crawler

Curator tool



# The Internet Archive

- *“Universal Access to Human Knowledge”*
- Digital repository accessible on the Internet
- Started in 1996 to archive web pages
- Best known for Wayback Machine
- Expanded to include movies, books, concerts, and other audio



# Web Archive

- Contents:
  - Snapshot every 2 months since 1996
  - Currently collect 4 billion pages per crawl (50TB/mo)
  - Collecting done by Alexa Internet, web cataloguing company
  - Storage, access and preservation done by Internet Archive



# The European digital Archive

- Was incorporated in 2004 as a non-profit foundation in Amsterdam with public and private support
- Technological and collection peering agreement with the Internet Archive



EA's 200 Tb  
data center in  
Amsterdam



# Our role

- Open archive for the public
- Technology partner for cultural institutions wishing to do web collections
- Focus and domain Crawl
- Access via online interface and search
- Quality assurance and reporting on collections
- Hosting and delivery of content
- Preservation and backup



- Current or recent Web projects
  - EU referendum
  - British elections with British Library
  - German election with DDB
  - Pilot study on archiving of TV and Radio website with the Netherlands Audiovisual Archive (BeelendGeleid)



- IIPC: <http://netpreserve.org>
- European Archive: <http://europarchive.org>  
(end of 2005)
- Web Archive information  
list: <http://listes.cru.fr/sympa/info/web-archive>
- International Web Archiving Workshop  
(IWAW): <http://iwaw.net>

