Web Archiving: a collaborative effort in progress
What has changed?

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

• ready before mid-2006:

• Robust & scalable up to the global web

• Implement IIPC standards (WARC, metadata, API...)

• Easy to install and use for advanced user (web archiving engineers)

• Open source and available for the all community of web archives
HERITRIX

HARVESTER

SOAP

Ingest

HTTP

BATCH

WEB

ARCHIVE

INDEX

feed

HTTP

SOAP

Search

ENGINE

SOAP

DIP

HTTP

REQUEST

WERA

RESULT

SET (XML)

ACCESS

TOOL

Focus
crawler

Curator
tool
Functional collaboration

• Mutualisation of resources for some tasks?
• Do all heritage institutions have to cope with all the technical difficulties?
• Can we all keep up with a permanently fast changing environment?
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
  – 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