Internet preservation: current situation and perspectives for the future.

Julien Masanès
European Web Archive director

Summary

- Why?
- Differences?
- How?
- Collaboration

Why archiving the web?

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

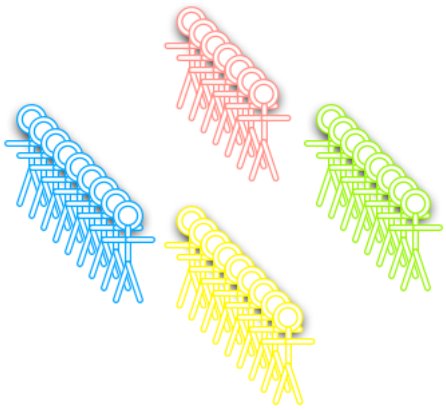
Koehler, W. (2004). A longitudinal study of Web pages continued: a consideration of document persistence. Information Research, 9(2),

(stable) & (discrete objects)

Time

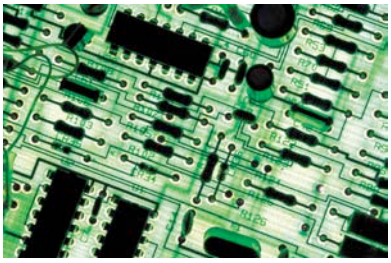
3 sec
20 pages / Min.
1 200 pages / Hr.
30 000 pages / Day

(publishers)

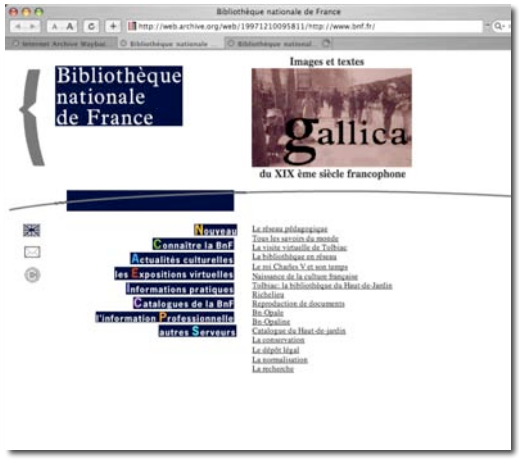


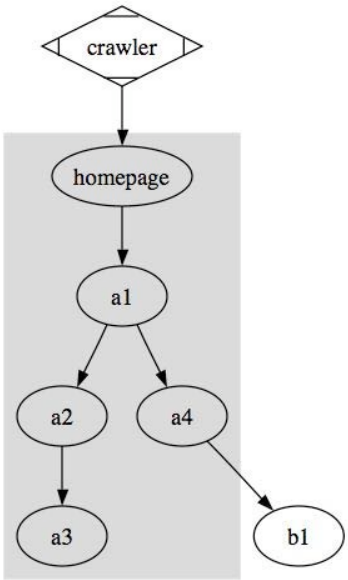
X 1000 at least

Good news



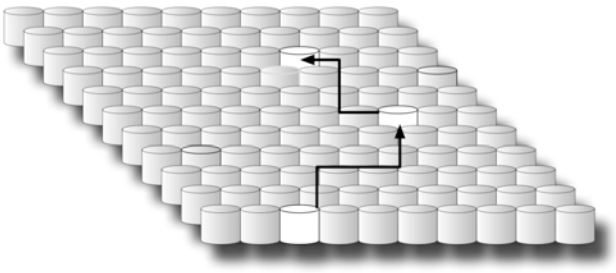
How?


Bibliothèque nationale de France
Images et textes
Bibliothèque nationale de France
gallica
du XIX ème siècle francophone
Nouveau
Connaître la BnF
Actualités culturelles
les Expositions virtuelles
Informations pratiques
Catalogues de la BnF
Information Professionnelle
autres Serveurs

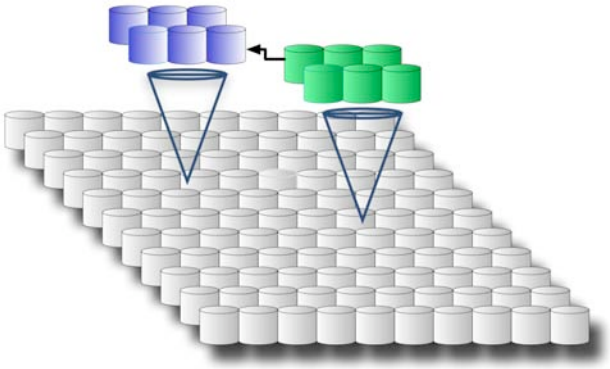

crawler
homepage
a1
a2
a4
a3
b1

Collaboration

www as a grid of servers



Web archives grid

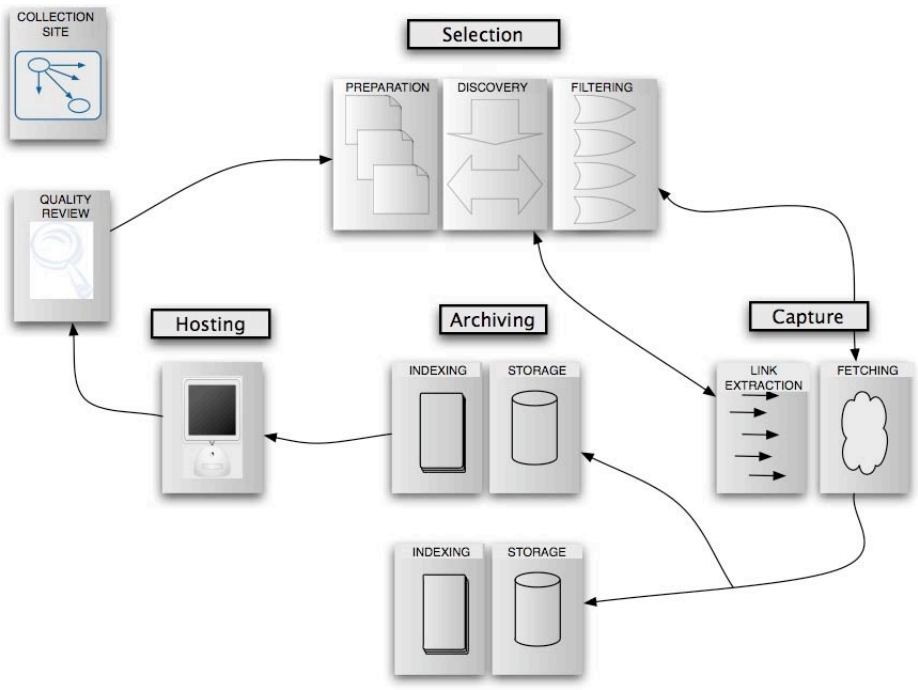


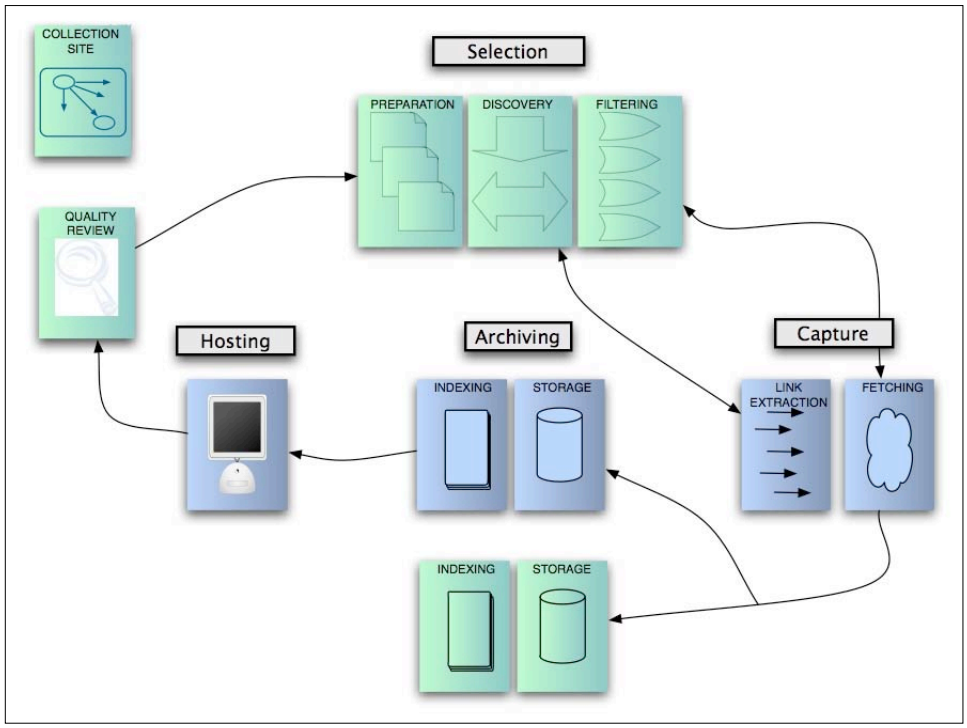
IIPC: standards and tools for the Web archives grid

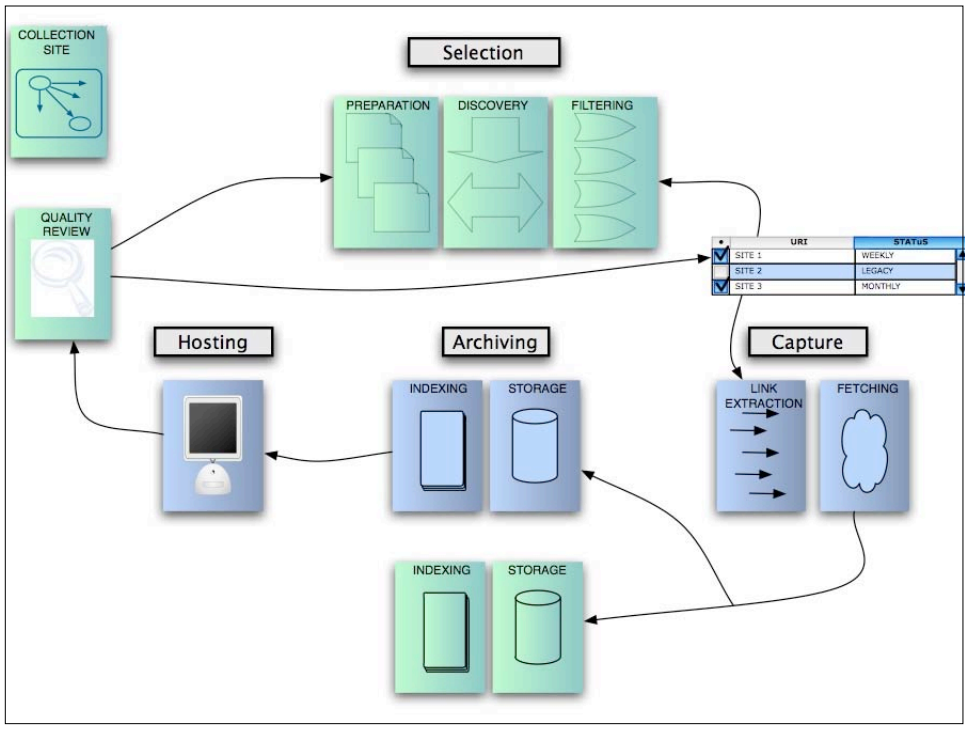
- Standards
  - Architecture
  - Storage format (WARC)
  - Metadata
- Tools for web archiving

Functional collaboration

- Do all heritage institutions have to cope with all the technical difficulties?
- Can we all keep up with a permanently fast changing environment?
- Mutualisation of resources for some tasks
  - Crawling
  - Mapping
  - Online access
  - Storage
  - Preservation




COLLECTION SITE
Selection
PREPARATION
DISCOVERY
FILTERING
QUALITY REVIEW
Hosting
Archiving
Capture
INDEXING
STORAGE
LINK EXTRACTION
FETCHING
INDEXING
STORAGE


COLLECTION SITE
Selection
PREPARATION
DISCOVERY
FILTERING
QUALITY REVIEW
URI
STATUS
SITE 1
WEEKLY
SITE 2
LEGACY
SITE 3
MONTHLY
Hosting
Archiving
Capture
INDEXING
STORAGE
LINK EXTRACTION
FETCHING
INDEXING
STORAGE

The European Web 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
  - UK gov archive with TNA
  - 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
- Web Archive information list:http://listes.cru.fr/sympa/info/web-archive
- International Web Archiving Workshop (IWAW): http://iwaw.net
- julien@europarchive.org