Internet preservation: current situation and perspectives for the future.

Julien Masanès
European Web Archive
director

Summary

• Why?
• Differences?
• How?
• Collaboration
Why archiving the web?

<table>
<thead>
<tr>
<th>Study</th>
<th>Resource type</th>
<th>Resource half-life</th>
</tr>
</thead>
<tbody>
<tr>
<td>Koehler (1999 and 2002)</td>
<td>Random Web pages</td>
<td>about 2.0 years</td>
</tr>
<tr>
<td>Harter and Kim (1996)</td>
<td>Scholarly Article Citations</td>
<td>about 1.5 years</td>
</tr>
<tr>
<td>Rumsey (2002)</td>
<td>Legal Citations</td>
<td>about 1.4 years</td>
</tr>
<tr>
<td>Markwell and Brooks (2002)</td>
<td>Biological Science Education</td>
<td>about 4.6 years</td>
</tr>
<tr>
<td>Spinellis (2003)</td>
<td>Computer Science Citations</td>
<td>about 4.0 years</td>
</tr>
</tbody>
</table>

(stable) & (discrete objects)

Time

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

\( \times 1000 \) at least

Good news
How?
Collaboration

www as a grid of servers
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
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