Creating a web archive for the EU institutions' websites
Achievements and challenges

Els Breedstraet
Coordinator web preservation
Publications Office of the EU

Creating the archive

- Harvesting
- Quality control
- Description
  - Access
  - Long term preservation
# Selection

<table>
<thead>
<tr>
<th>Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal: Publications Office / Historical Archives of the EU</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seed list of ca 70 europa.eu URLs</td>
</tr>
<tr>
<td>Covers mainly sites of EU institutions, bodies and agencies</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection policy</td>
</tr>
<tr>
<td>To be limited to europa.eu domain and subdomains?</td>
</tr>
<tr>
<td>Which seeds to add? Which selection criteria?</td>
</tr>
<tr>
<td>Who decides to add or not?</td>
</tr>
<tr>
<td>=&gt; need for policies + curator and/or selection board</td>
</tr>
<tr>
<td>Seed management:</td>
</tr>
<tr>
<td>How to propose new seeds =&gt; need for selection tool</td>
</tr>
<tr>
<td>How to keep list up to date?</td>
</tr>
</tbody>
</table>
# Harvesting

## Outsourced (Internet Memory Research)

- Quarterly crawling of seed list
- Using Heritrix
- Runs successful since end of 2013
- Till mid 2016: ARC, afterwards WARC

## Questions

- Dynamic content
- Social media
- Multilingual harvesting
- Ad hoc demands/campaigns
  - How to decide to do or not?
  - Planning and budgeting
## Quality control

| | Automatic: outsourced  
Manual: website owners, coordinated by the Publications Office |
|---|---|
| ![Checkmark] | • Automatic checks + manual / visual control  
• Network of quality controllers  
• Structured follow up of feedback => long term improvement harvests |
| ![Question Mark] | • Automatic checks: more? Reporting?  
• Manual control  
  • Need for quality control tool  
  • Need for crawl reports  
• Awareness raising -> webmasters to be encouraged to make "well preservable" sites |
## Description / metadata

<table>
<thead>
<tr>
<th>Icon</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Info" /></td>
<td><strong>Automatic (up to now)</strong></td>
</tr>
</tbody>
</table>
| ![Check](image) | - Automatic capturing of descriptive and technical MD at the source  
  - Proposal for minimal set of MD (descriptive, technical and provenance) ready |
| ![Question](image) | - Awareness raising -> webmasters to be encouraged to add more and/or more structured MD at the source  
  - Provenance MD  
  - Post-crawl manual description?  
  - How to exploit MD to improve access? |
## Access

<table>
<thead>
<tr>
<th>Outsourced (Internet Memory Research)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Fully open archive: <a href="http://collections.internetmemory.org/haeu">http://collections.internetmemory.org/haeu</a></td>
</tr>
<tr>
<td>• Basic search and browse options</td>
</tr>
</tbody>
</table>

- Audience / use cases?
- Usage statistics
- Branding / customisation
- Data protection, copyright, etc.
- Multilingual access
- Promotion and communication
- Centralised access to EU related web archives? Cooperation?

---

Access

<table>
<thead>
<tr>
<th>Outsourced (Internet Memory Research)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Fully open archive: <a href="http://collections.internetmemory.org/haeu">http://collections.internetmemory.org/haeu</a></td>
</tr>
<tr>
<td>• Basic search and browse options</td>
</tr>
</tbody>
</table>

- Audience / use cases?
- Usage statistics
- Branding / customisation
- Data protection, copyright, etc.
- Multilingual access
- Promotion and communication
- Centralised access to EU related web archives? Cooperation?
## Long term preservation

<table>
<thead>
<tr>
<th>![Icon]</th>
<th><strong>Internal: digital repository at Publications Office</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>![Checkmark]</td>
<td>- First tests started with ingestion in internal trusted long term digital preservation repository</td>
</tr>
</tbody>
</table>
| ![Question Mark] | - Adaptation information architecture  
- Provenance metadata  
- Trustworthy Digital Repository Certification (ISO16363) |
Conclusions

- Solid basis for an EU institutional web archive is in place, accessible to everyone
- Many challenges, a lot of work ahead
- Many dreams and ideas for the future
  - One step at the time
  - Cooperation?
More information

Have a look at our web archive:
http://collections.internetmemory.org/haeu

Questions or suggestions? Contact us:
op-web-preservation@publications.europa.eu
Thanks for your attention!

Questions?