WebArchiv - Czech Web Archive revisited

Petr Žabíčka
Moravian Library in Brno
zabak@mzk.cz

Overview

- WebArchiv – update on current activities
- Future plans
- Web Culture Heritage project
- Project partners
- Project results
WebArchiv – overview

- Long-term preservation of valuable bohemical online resources (since 2000)
- National Library in Prague in co-operation with the Masaryk University in Brno and Moravian Library in Brno.
- Started with NEDLIB and Nordic Metadata project tools, now using IIPC tools with locally developed additions.
- Currently, WebArchiv development depends on financing from R&D projects but plans are to include this activity into the National Library budget.
- New copyright law allows us to give users access to the archive from within the library.
- Public access given only to documents covered by agreements with publishers.

http://www.webarchiv.cz

WebArchiv – Workflows

- Prague:
  - Resource selection
  - Cataloguing for the National Bibliography (MARC21)
  - Providing Dublin Core metadata for interested publishers
  - Making agreements with publishers
- Brno:
  - Running WebArchiv hardware
  - Software localization, maintenance and development
  - Pre-harvesting resource analysis
  - Harvesting, indexing, access
- Results so far:
  - 3.6 TB archive with 95 million files

http://www.webarchiv.cz
WebArchiv – Infrastructure

- Software tools:
  - Web Based Dublin Core metadata creator
  - National Bibliography Number (NBN) generator
  - Heritrix crawler
  - NutchWAX, WERA – full text indexing & public archive access
  - wa-cz – locally developed infrastructure
  - WayBack – Wayback Machine like interface for whole archive

- Hardware:
  - 3 HP ProLiant servers
  - 5.8 TB SATA disc array
  - awaiting transfer of the archive files to National Library’s central storage facility (25+ TB, mirrored, FC+SATA) later this year

http://www.webarchiv.cz

IWAW’06

WebArchiv – Infrastructure

A1 new crawl; A2 end crawl -> index; A3 update fulltext; A4 update host list

http://www.webarchiv.cz

IWAW’06
WebArchiv - Future Work

- Workflow management application
- Harvesting of bohemal resources outside the .cz domain
  - language analysis
  - feedback from Heritrix about dropped URLs from .cz crawl
- Adaptive incremental harvesting, incremental indexing
- Selective harvesting on demand
- Fulltext indexing of the whole archive
- Identification of similar documents
- Permanent linking into the archive (permanent ID)
- Integration of the archive into planned National Digital Library
  (selection of software 2008)
- Long-term preservation (via NDL system)
- Implementation of digital library standards: OAI-PMH, METS, SRU/SRW
2005/2006 – Culture 2000 project

- Project partners:
  - National Library of the Czech Republic – project coordinator
  - National Library of Estonia
  - National and University Library, Ljubljana, Slovenia
  - University Library in Bratislava, Slovakia

- Project goals:
  - Analysis of existing methods of selection policies in other countries
  - Mapping and analysis of software tools and workflows
  - Selection criteria recommendation
  - Adopting and adjustment of chosen software tools
  - Exchange of experience

http://www.webarchiv.cz  IWAwards'06

Slovenia

- harvesting .si domain since 2004
- using harvester WebBird, developed by Jožef Stefan Institute
- depend on the Intstitute with any web archiving activity
- good product but dismal service provider

- Plans
  - Transfer control of web harvesting to the library

- Contact person
  - Alenka Kavčič – Čolič (Alenka.Kavcic@nuk.uni-lj.si)
Estonia
- using htrack to harvest selected files from the net since 2003

Plans
- start using Heritrix to collect selected websites next year

Contact Persons
- Kairi Felt (Kairi.Felt@nlib.ee)
- Liina Enok (Liina.Enok@nlib.ee)

Slovakia
- no previous web-archiving activity

Project result
- established web archiving infrastructure in Bratislava (linked to Slovak ISSN agency)
- based on WebArchiv infrastructure
- first harvest of electronic serials registered by the ISSN agency

Plan
- increase storage capacity

Contact persons
- Alojz Androvič (androvic@ulib.sk)
- Katarína Kinčiková (kincikova@ulib.sk)
Project results

- Information on web archiving activities in other countries
- Crawl of 120 servers (30 from each country) using Heritrix, htrack and WebBird
- Comparison of the three crawls
- Archive analysis to find best file-type based harvesting limits
- Selection criteria recommendation

- All documents available on project website:


Selection criteria

1. National domain, other domain only if resource meets...
2. ...national aspect: author’s nationality; national language; about nation/country
3. Original content; cultural, scientific or research value
4. Freely available access, or access by special agreement
5. Commonly used file formats
6. Digitally born documents
7. Resource types:
   + periodicals, monographs, conference proceedings, research and other reports, academic works, government materials
   - games, intranet resources, personal blogs, portals without intellectual content, datasets, broadcasting
### Crawler comparison I.

<table>
<thead>
<tr>
<th>Fetched documents</th>
<th>Heritrix</th>
<th>Hittrack</th>
<th>WebBird</th>
<th>Uncompressed size(GB)</th>
<th>Heritrix</th>
<th>Hittrack</th>
<th>WebBird</th>
</tr>
</thead>
<tbody>
<tr>
<td>cz domain:</td>
<td>104 659</td>
<td>16 719</td>
<td>74 217</td>
<td></td>
<td>3 5</td>
<td>0 7</td>
<td>3 4</td>
</tr>
<tr>
<td>sk domain:</td>
<td>138 300</td>
<td>28 544</td>
<td>155 487</td>
<td></td>
<td>3 5</td>
<td>1 7</td>
<td>8 6</td>
</tr>
<tr>
<td>si domain:</td>
<td>102 910</td>
<td>18 852</td>
<td>76 838</td>
<td></td>
<td>5 9</td>
<td>1 6</td>
<td>4 6</td>
</tr>
<tr>
<td>Total:</td>
<td>349 969</td>
<td>83 897</td>
<td>256 534</td>
<td></td>
<td>16 6</td>
<td>5 6</td>
<td>24 4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fetched per host</th>
<th>Heritrix</th>
<th>Hittrack</th>
<th>WebBird</th>
<th>Size per host(MB)</th>
<th>Heritrix</th>
<th>Hittrack</th>
<th>WebBird</th>
</tr>
</thead>
<tbody>
<tr>
<td>kvetova.cz</td>
<td>5 192</td>
<td>319</td>
<td>75</td>
<td></td>
<td>67 2</td>
<td>2 2</td>
<td>0 7</td>
</tr>
<tr>
<td>mak.kosmo.cz</td>
<td>5 099</td>
<td>1 738</td>
<td>2 260</td>
<td></td>
<td>14 0</td>
<td>1 6</td>
<td>3 2</td>
</tr>
<tr>
<td>natura.baf.cz</td>
<td>1 313</td>
<td>62</td>
<td>574</td>
<td></td>
<td>32 5</td>
<td>0 4</td>
<td>1 9</td>
</tr>
<tr>
<td>nb.vse.cz</td>
<td>635</td>
<td>24</td>
<td>523</td>
<td></td>
<td>32 0</td>
<td>0 9</td>
<td>3 3</td>
</tr>
<tr>
<td>portalk.med.muni.cz</td>
<td>599</td>
<td>240</td>
<td>223</td>
<td></td>
<td>8 2</td>
<td>3 0</td>
<td>8 6</td>
</tr>
<tr>
<td>rostliny.nikode.cz</td>
<td>5 005</td>
<td>2 625</td>
<td>5 528</td>
<td></td>
<td>102 4</td>
<td>6 0</td>
<td>19 0</td>
</tr>
<tr>
<td>vanastro.cz</td>
<td>203</td>
<td>117</td>
<td>211</td>
<td></td>
<td>24 3</td>
<td>2 1</td>
<td>0 0</td>
</tr>
<tr>
<td><a href="http://www.nasepole.sk">www.nasepole.sk</a></td>
<td>1 357</td>
<td>121</td>
<td>1 100</td>
<td></td>
<td>14 9</td>
<td>1 2</td>
<td>1 3</td>
</tr>
<tr>
<td><a href="http://www.natrologa.sk">www.natrologa.sk</a></td>
<td>1 387</td>
<td>608</td>
<td>4 387</td>
<td></td>
<td>34 8</td>
<td>18 8</td>
<td>173 0</td>
</tr>
<tr>
<td><a href="http://www.abcan.sk">www.abcan.sk</a></td>
<td>2 444</td>
<td>1 057</td>
<td>4 813</td>
<td></td>
<td>78 9</td>
<td>1 6</td>
<td>0 0</td>
</tr>
<tr>
<td><a href="http://www.pu.army.sk">www.pu.army.sk</a></td>
<td>1 103</td>
<td>1 42</td>
<td>58</td>
<td></td>
<td>24 2</td>
<td>3 1</td>
<td>0 0</td>
</tr>
<tr>
<td><a href="http://www.xkske.sk">www.xkske.sk</a></td>
<td>830</td>
<td>39</td>
<td>2 437</td>
<td></td>
<td>21 8</td>
<td>0 3</td>
<td>3 4</td>
</tr>
<tr>
<td><a href="http://www.akase.sk">www.akase.sk</a></td>
<td>52</td>
<td>0</td>
<td>50</td>
<td></td>
<td>6 4</td>
<td>0 0</td>
<td>8 0</td>
</tr>
</tbody>
</table>

http://www.webarchiv.cz

IWAW'06

---

### Crawler comparison II.

<table>
<thead>
<tr>
<th>Mimetype</th>
<th>Heritrix</th>
<th>Hittrack</th>
</tr>
</thead>
<tbody>
<tr>
<td>text/html</td>
<td>283 294</td>
<td>66 840</td>
</tr>
<tr>
<td>image/jpeg</td>
<td>52 972</td>
<td>10 252</td>
</tr>
<tr>
<td>image/gif</td>
<td>40 633</td>
<td>6 206</td>
</tr>
<tr>
<td>application/pdf</td>
<td>9 158</td>
<td>2 970</td>
</tr>
<tr>
<td>application/x-javascript</td>
<td>4 999</td>
<td>0</td>
</tr>
<tr>
<td>application/msword</td>
<td>4 275</td>
<td>15</td>
</tr>
<tr>
<td>image/png</td>
<td>2 898</td>
<td>231</td>
</tr>
<tr>
<td>text/css</td>
<td>2 229</td>
<td>0</td>
</tr>
<tr>
<td>text/plain</td>
<td>1 632</td>
<td>0</td>
</tr>
<tr>
<td>text/x-csv</td>
<td>1 863</td>
<td>405</td>
</tr>
<tr>
<td>application/x-javascript</td>
<td>1 441</td>
<td>218</td>
</tr>
<tr>
<td>application/vnd.ms-excel</td>
<td>1 248</td>
<td>177</td>
</tr>
<tr>
<td>application/rtf</td>
<td>1 216</td>
<td>133</td>
</tr>
<tr>
<td>text/rdf</td>
<td>1 194</td>
<td>15</td>
</tr>
<tr>
<td>text/plain</td>
<td>1 046</td>
<td>58</td>
</tr>
<tr>
<td>application/octet-stream</td>
<td>1 000</td>
<td>0</td>
</tr>
<tr>
<td>application/vnd-zip-xgzt</td>
<td>578</td>
<td>377</td>
</tr>
</tbody>
</table>

http://www.webarchiv.cz

IWAW'06
Crawler comparison III.
Number of files downloaded from any given pathname

Archive analysis
- If we limit the size of files in the archive, what percentage of web content (number of files, size of archive) do we lose?
- Performed on two 1TB harvests
  - cz2004: no limit on filesize
  - cz2006: max. 100 MB per file
- File type groups: text (incl. pdf, doc...!), image, application, video, audio, container
Web Cultural Heritage

http://www.webarchiv.cz

IWAW’06

Web Cultural Heritage

http://www.webarchiv.cz

IWAW’06
Web Cultural Heritage

We are ready to [co|x]operate!

Project management: Ludmila Celbová
Petr Žabička
Harvesting: Adam Brokeš
Archive access: Lukáš Matějka

mail@webarchiv.cz