Archiving the Internet: Challenges, Projects, and the Austrian Perspective

Andreas Rauber¹, Andreas Aschenbrenner¹, Alfred Schmidt²

¹Department of Software Technology (ifs)
Vienna University of Technology
http://www.ifstuwien.ac.at

²Austrian National Library
http://www.onb.ac.at
Overview

• Motivation

• Technical challenges
  • Selection
  • Archiving
  • Preservation
  • Access

• Related projects
  • Internet Archive
  • Kulturaw3
  • Nedlib

• AOLA: Austria On-Line Archive
Motivation

Why Should We Archive the Internet?

- Collection of sex & crime
- Masses of useless and/or wrong information
- Incredibly huge
- Only 0.00x % of all information is actually being looked at
- Who is interested in some fellow’s homepage?
- Important information is published in “real” media anyway
Motivation

The Invention of the Press

- Internet often compared to invention of the printing press
- Explosion of printed information
- Quality much lower than manually crafted codices
- Not to be considered important?

- Letters more interesting than books
- Ads, posters, and snippets tell more about a society than “high-quality” information sources
- What if only codices had been preserved?
Motivation

Some Considerations

- Increasing masses of information published electronically
- Volatility of Internet resources
- Social and cultural dimension - modern cultural heritage!
- Need to preserve the Internet
  - information / content
  - look-and-feel
- The early days of the Internet are already lost!
Motivation

Challenges

• Legal challenges
  • copyright issues
  • authenticity

• Technical challenges
  • what to archive
  • how to archive
  • how to keep the archive in good condition
  • how to provide access to the archive

• Financial Challenges
  • who is willing to pay
  • what do we gain (earn?) from it
Overview

• Motivation

• Technical challenges
  • Selection
  • Archiving
  • Preservation
  • Access

• Related Projects
  • Internet Archive
  • Kulturaw3
  • Nedlib

• AOLA: Austria On-Line Archive
1. Selection

- Building a complete archive is technically impossible
  - enormous amount of data
  - no central catalogue
  - high dynamics and volatility

- Manual Selection:
  select specific sites plus archiving frequency

- Automatic harvesting:
  automatically crawl hyperlinks to download sites

- Which sites to archive: *.at, .com, .cc, Austriaca, ...

- Questions of liability: who is responsible for content?
2. Archiving

- Selection of suitable storage media
  - high capacity
  - long durability
  - stable technology

- Migration to new storage media
  - when reaching lifetime of storage medium
  - when storage technology becomes obsolete
  - no “museum” of old devices
  - automatic transfer

- Media of choice currently
  - harddisk arrays
  - tapes
Technical Challenges

3. Preservation

• Digital objects have to be “interpreted”
• Software required for access
• Software needs specific hardware platform
• Ensure, that access to documents is possible in the future
• “Museum” of old hardware impossible to sustain

• 2 approaches
  • Conversion:
    converting to “standard file formats”
  • Emulation:
    emulating obsolete hardware on new systems
3.1 Preservation: Conversion

- Files are converted into (few) selected standard file formats (z.B.: text, (series of) image(s), sound, ...)

  + Access via a few file formats -> small set of access software
  + flexible and cheap, especially for immediate access
  + When standard file format becomes obsolete, converters will be around due to critical mass of existing files

- Loss of information at conversion (functionality, looks-and-feel)
- not suitable for all materials (e.g. interactive art)
- constantly maintain all data
3.2 Preservation: Emulation

- Storing description of system environment required for executing access software (metadata)
- Emulators for hardware platforms are created as the necessity arises
- Intermediate representation language
  + Theoretically most stable model
  + Conceptually clean solution
  - Very expensive (development of specialized emulators)
  - Not useful for quick, casual access
  - Information required for emulator development might not be known
  - Applicability has so far only been demonstrated on some selected examples, several open questions
4. Access

- Mostly legal issues
- Technical issues
  - provide access to large data stores within reasonable time frames
  - navigating the archive:
    * by content within a time frame
    * browsing through time (evolution of websites)
  - providing transparent access through emulators or migrated file formats
Overview

• Motivation

• Technical issues
  • selection
  • archiving
  • preservation
  • access

• Related projects
  • Internet Archive
  • Kulturaw3
  • Nedlib

• AOLA: Austria On-Line Archive
Related Projects

Internet Archive

- Since 1996, URL: www.archive.org
- Set of Linux-Systems with harddisk arrays

Archives “donated” data collections
- Mostly based on free harvesting (Alexa)
- Initially text-only, now all data types
Related Projects

Internet Archive (2)

- March 2001: approx. 43 TB (March 2000: 14 TB)
- Daily growth: up to 100 GB per day
- Redundancy: distributed across several sites
- Automatic migration onto new storage media
- Collecting existing emulators
- Access limited to research institutions
- “Programming skills” required for using the archive
Related Projects

Kulturaw3

- Project of the Swedish National Library, since 1996
- Sun Sparc Stations with tape robot archive
- Uses modified indexer (Combine) for harvesting
- Snapshots of the swedish web (.se, .nu, special domains)
- Preservation: originals plus possibly standard file formats
- 5 snapshots so far, last crawl:
  15 Mio. URLs from 58,400 websites, total approx. 280 GB data
- Tendency: dramatically increasing (incomplete 6. crawl: 360 GB !)
- Access tool under development
- Plan to use hierarchical storage media
Related Projects

NEDLIB

- URL: http://www.kb.nl/coop/nedlib/
- Networked European Deposit Library
- EU-Project January 1998 till January 2001
- Framework for Archiving on-line media
- Open for all concepts
- Guidelines, technical standards, “best-practice” models
- Nedlib harvester for archiving issues
- Platforms: Linux, SUN, ...
- Further tools under development
- No large-scale experiments so far
Related Projects

Further Projects:

- Nordic Web Archive - http://nwa.nb.no
- CAMiLEON - http://www.si.umich.edu/CAMILEON/
- CEDARS - http://www.leeds.ac.uk/cedars/
- Prism - http://prism.cornell.edu/PrismWeb/
- LOCKSS - http://lockss.stanford.edu/
- InterPARES - http://www.interpares.org/
Overview

• Technical issues
  • selection
  • archiving
  • preservation
  • access

• Related Projects
  • Internet Archive
  • Kulturaw3
  • Nedlib

• AOLA: Austria On-Line Archive
Austria On-line Archive

- URL: http://www.ifs.tuwien.ac.at/~aola/
- Cooperation between the Austrian National Library and the Department of Software Technology, Vienna Univ. of Technology
- Pilot study: preparations since 1999, 1. phase since March 2001
- Linux-System with 240 GB harddisk plus 6-fold tapedrive
- Open source approach to ensure independent access
- Initially: Nedlib harvester (incl. modifications and expansions)
- Goal: snapshot of the Austrian webspace
Austria On-line Archive (2)

- between May 7 and May 16 2001 approx. 10 crawler parallel
- Download during pilot phase: approx. 1GB per day
- at-domain as well as selected subdomains, esp. *.cc, *.com, *.edu, etc.

- Statistics May 7. - 16.:  
  - about 666,000 unique URLs harvested  
  - 1,210 sites accessed  
  - total of 8.3 GB of data stored  
  - numerous problems with Nedlib harvester encountered
Results of Pilot Phase

• Basically, the setup works!

• Archiving system problems:
  • XFS file system for Linux still unstable (pre-release)

• Nedlib Harvester
  • problems with mal-formatted links in html pages
  • communication problems within system
  • several pages downloaded numerous times
  • still in development phase

• --> crawl needed to be stopped
## Statistics - Domains (excerpt)

<table>
<thead>
<tr>
<th>Domain (47)</th>
<th>Size</th>
<th>#Docs</th>
<th>#Hosts</th>
</tr>
</thead>
<tbody>
<tr>
<td>at</td>
<td>4.345.098.283</td>
<td>239.821</td>
<td>8.740</td>
</tr>
<tr>
<td>ac.at</td>
<td>454.072.064</td>
<td>19.248</td>
<td>676</td>
</tr>
<tr>
<td>co.at</td>
<td>138.067.628</td>
<td>12.557</td>
<td>427</td>
</tr>
<tr>
<td>gv.at</td>
<td>75.164.569</td>
<td>4.584</td>
<td>234</td>
</tr>
<tr>
<td>or.at</td>
<td>55.349.125</td>
<td>5.576</td>
<td>197</td>
</tr>
<tr>
<td>com</td>
<td>331.110.660</td>
<td>18.419</td>
<td>813</td>
</tr>
<tr>
<td>edu</td>
<td>737.588</td>
<td>24</td>
<td>9</td>
</tr>
<tr>
<td>int</td>
<td>1.183.712</td>
<td>80</td>
<td>1</td>
</tr>
<tr>
<td>net</td>
<td>202.837.209</td>
<td>13.108</td>
<td>457</td>
</tr>
<tr>
<td>org</td>
<td>45.412.967</td>
<td>1.908</td>
<td>93</td>
</tr>
<tr>
<td>cc</td>
<td>402.520.513</td>
<td>13.513</td>
<td>119</td>
</tr>
<tr>
<td>de</td>
<td>32.043.054</td>
<td>2.233</td>
<td>250</td>
</tr>
<tr>
<td>hu</td>
<td>516.579</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>tw</td>
<td>43</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
## Statistics - Extensions (excerpt)

<table>
<thead>
<tr>
<th>Extension</th>
<th>Size</th>
<th>#Docs</th>
<th>Extension</th>
<th>Size</th>
<th>#Docs</th>
</tr>
</thead>
<tbody>
<tr>
<td>htm</td>
<td>483.073.276</td>
<td>58.554</td>
<td>exe</td>
<td>266.082.863</td>
<td>513</td>
</tr>
<tr>
<td>html</td>
<td>504.579.811</td>
<td>43.815</td>
<td>bin</td>
<td>240.768</td>
<td>1</td>
</tr>
<tr>
<td>txt</td>
<td>5.499.481</td>
<td>1.452</td>
<td>cgi</td>
<td>49.386.025</td>
<td>3.671</td>
</tr>
<tr>
<td>wav</td>
<td>38.212.215</td>
<td>107</td>
<td>java</td>
<td>7.489</td>
<td>1</td>
</tr>
<tr>
<td>mp3</td>
<td>216.255.942</td>
<td>169</td>
<td>jsp</td>
<td>18.854.236</td>
<td>684</td>
</tr>
<tr>
<td>avi</td>
<td>8.955.594</td>
<td>12</td>
<td>asp</td>
<td>848.447.298</td>
<td>26.527</td>
</tr>
<tr>
<td>mpg</td>
<td>179.078.751</td>
<td>19</td>
<td>php</td>
<td>160.913.685</td>
<td>7.881</td>
</tr>
<tr>
<td>jpeg</td>
<td>2.090.006</td>
<td>133</td>
<td>xls</td>
<td>4.000.256</td>
<td>28</td>
</tr>
<tr>
<td>jpg</td>
<td>549.196.700</td>
<td>35.298</td>
<td>doc</td>
<td>41.327.637</td>
<td>328</td>
</tr>
<tr>
<td>gif</td>
<td>388.089.230</td>
<td>76.498</td>
<td>f94</td>
<td>957</td>
<td>1</td>
</tr>
<tr>
<td>zip</td>
<td>184.489.636</td>
<td></td>
<td>fangan</td>
<td>10.022</td>
<td>3</td>
</tr>
<tr>
<td>gz</td>
<td>3.091.367</td>
<td>9</td>
<td>woa</td>
<td>4.046</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>346a</td>
<td>43</td>
<td>1</td>
</tr>
</tbody>
</table>
AOLA - Next Steps

• Currently switching to Combine harvester for the next crawl
• Transform pilot study into permanent institution
• Archiving frequent snapshots of the Austrian webspace
• Develop long-term strategies for preservation
• Combination of conversion and emulation approaches
• Setting up technical and personnel infrastructure
Conclusions

• **Goal**: Preservation of (modern) cultural heritage

• **Selection**: Combination of manual selection and free harvesting

• **Archiving**: Migration of (hierarchical) storage media

• **Preservation**: Emulation and conversion approaches

• **Access**: Interfaces and legal aspects

• **Urgency**: We have to start **NOW**!
AOLA Project-Homepage:

http://www.ifs.tuwien.ac.at/~aola