From Web Measurement to Archiving
*(Technical Development of Croatian Web Archive Over the Past 15 Years)*

Miroslav Milinović, Draženko Celjak
SRCE - University of Zagreb, University Computing Centre

*Web Archiving Conference, 6 - 7 June 2019, Zagreb, Croatia*
Contents

• About SRCE
• HAW
  • History
  • Current status
  • Future challenges
SRCE
University of Zagreb, University Computing Centre

• main computing centre and the architect of the e-infrastructure
• providing services for the whole research and high education system (200+ institutions)
SRCE
Data services, digital repositories, archives ...

HAW
Hrvatski arhiv weba
Croatian Web Archive
NUL & Srce since 2004

hrčak
since 2006

dabar
DIGITAL ACADEMIC ARCHIVES AND REPOSITORIES
since 2014
Our experience

- **Croatian Web Measurement Project (MWP)**
  - 2002 - 2008 (6 measurements)
  - MWP SW

- **Croatian Web Archive (HAW)**
  - in cooperation with NUL
  - cooperation started in 2003
  - [http://haw.nsk.hr](http://haw.nsk.hr)
  - DAMP SW

- **Digital Archive of the Web Resources of the Republic of Croatia**
  - in cooperation with Central State Office for the Development of the Digital Society
  - harvesting public authorities’ web resources since 2004
  - AMD SW (modified DAMP)
Croatian Web Measurement Project (MWP)

- **motivation:** cooperation with NUL on project NISKA – building digital library
- **goals:** estimate the size and complexity of Croatian Web, acquire basic info about the content
- **target:** resources accessible via HTTP / HTTPS at servers in .hr
- **we collected:** size, formats (MIME), metadata, links between servers/sites
- **conclusion:**
  - harvesting / archiving is possible
  - with certain limits
  - minimum: site’s “screenshot”
  - (surface) Web is (still) simple
  - small numbers of formats mainly used
  - authors do not care much about standards and archiving
  - challenges: new technologies, dynamic content, embedded resources, innovative but non-standard way of using technology
Croatian Web Archive (HAW)

• **2004 - DAMP**
  • Digital Archive System for Harvesting and Archiving of Legal Deposit of Croatian Web Publications
  • selective capturing of web resources
  • interaction with the NUL’s Catalogue
  • DAMP SW (continuous maintenance & development)
  • special feature: site’s “screenshot”
  • 63000+ captures / site copies

• **2011 - domain and thematic harvesting**
  • Heritrix + (Open)Wayback
  • 8 domain harvests
  • 10 thematic harvests
HAW: architecture

- HAW 2019
  - $\approx$ 54 TB
    - selective capturing: 17 TB
    - .hr harvesting: 36.7 TB
    - thematic harvesting: 300 GB
- OpenSource
  - DAMP 4.2.2
  - Heritrix 3.3
  - OpenWayback 2.3.2
- features
  - OAI PMH
  - URN: NBN
- 4 servers
- hosted in SRCE
DAMP: functional model
HAW: national domain and thematic harvesting

**WEB**
- HTTP / HTTPS

**harvesting system** *(HERITRIX)*

**DATA STORAGE**
- ARCHIVE *(WARC format)*
- (WAYBACK) index

**ADMINISTRATOR**

**USERS**
- HTTP

**Web interface** *(WAYBACK)*

**access to archive** *(WAYBACK)*

**WEB**

**USERS**
## DAMP vs. Heritrix

<table>
<thead>
<tr>
<th>Heritrix</th>
<th>DAMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>needs Wayback to access the archived web sites</td>
<td>archived web site can be directly accessed with any browser (captured copy is modified)</td>
</tr>
<tr>
<td>uses warc format (saves disk space; easier maintenance)</td>
<td>saves a copy of the site (lot of files on the disk)</td>
</tr>
<tr>
<td>flexible, number of options</td>
<td>simple to use, tailored to the need (includes curator tool)</td>
</tr>
<tr>
<td>saves HTTP headers</td>
<td>screenshot feature</td>
</tr>
</tbody>
</table>
Challenges in harvesting

• **planning phase:**
  • resource selection (protocol, domain, seed, depth, ...)
  • handling embedded resources (media, frames, ads, …)
  • redirects
  • social media
  • robots.txt

• **execution phase:**
  • CMS features / configurations
  • “endless” sites
  • online catalogues / sales
  • interactive content
  • (wrong) link detection
Before conclusion: MWP1 vs .hr harvesting 2016

- MWP1 (29.03.2002 - 07.05.2002)
  - 4,667,920 resources
  - estimated size > 300 GB
  - share in total number of resources:
    - HTML 67%
    - picture formats 23%

- .hr harvesting 2016 (25.12.2016 - 02.01.2017)
  - 77+ millions of files
  - total size ≈ 7.0 TB
  - share in total number of resources:
    - HTML 51.3%
    - pictures in JPEG format 33.8%
Conclusion (How we see the future)

• harvesting / archiving is (still) possible
  • with certain limits
  • minimum: “screenshot” of site’s homepage

• (surface) Web is (still) simple
  • small numbers of format mainly used

• authors do not care much about standards and archiving

• challenges:
  • new technologies, dynamic content, embedded resources
  • innovative but non-standard ways of using technology

• future work:
  • address recognised challenges
  • cater for new web technologies
  • better handling of embedded resources
  • archive indexing (full text)
  • social media harvesting
  • tools for archive admins
  • interoperability
According to the Open Access Policy, Srce ensures that all research data made by Srce is accessible and free to use by the general public, especially educational and professional information and content derived from the actions and work of Srce.

This material is available under the International Creative Commons License 4.0 Attribution-NonCommercial.

http://haw.nsk.hr/
damp@srce.hr