Using web archives for researchers

15:15 - 16:30

Sally Chambers – GhentCDH and DARIAH
Eveline Vlassenroot – imec-mict-UGent
Jesse de Vos – Beeld & Geluid
Valérie Schafer – C2DH
Patricia Blanco – MSc in Digital Humanities (KU Leuven)
Introduction

Sally Chambers - GhentCDH and DARIAH
User requirements for a web archive

Eveline Vlassenroot – imec-mict-UGent
Research question:

What are the requirements and needs of potential users of web archives?

How do they, or do they want to, access, use and consult web archives?
Methodology

- Survey with a local, regional, national and international focus
- Ran 2 months (April 2018 - end of May 2018)
- 3 languages: Dutch, French and English
- Total: 587 respondents (after cleaning: 399)
Target groups

1. **Research**: students, academics, anyone involved in research more broadly

2. **Archives, libraries and governmental institutions**

3. **General public**
Socio demographics (n=399)

Researchers: 36.3%
Archivists, librarians: 32.6%
Broad public: 31.1%

Educational attainment:
- Elementary school: 0.8%
- Secondary school: 13.0%
- Bachelor’s degree: 20.1%
- Master’s degree: 49.1%
- PhD: 17.0%
Are you aware of the existence of web archives? n=397

- Researchers: 28.2% Yes, 8.3% No
- Archivists, librarians: 28.0% Yes, 4.3% No
- Broad public: 23.4% Yes, 7.8% No
Can you provide an example of a web archive?
Have you ever used a web archive? n=399

<table>
<thead>
<tr>
<th>Group</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researchers</td>
<td>15.5%</td>
<td>20.8%</td>
</tr>
<tr>
<td>Archivists, librarians</td>
<td>15.0%</td>
<td>17.5%</td>
</tr>
<tr>
<td>Broad public</td>
<td>16.5%</td>
<td>14.5%</td>
</tr>
</tbody>
</table>
Why are you not using web archives? n=180

- I didn’t know of their existence. 45.6%
- No relevant data available in the web archive for my topic. 18.3%
- Access is (too) restricted. 11.1%
- Too difficult to work with (lack of research infrastructure). 5.0%
- Incomplete or inaccurate archived materials. 1.7%
- Other 18.3%
Which search functionalities do you deem most important for the Belgian web archive? (max. 2 options)

- Search engine for URLs
  - A: 21%

- Search engine for images
  - B: 17%

- Search engine for videos
  - C: 4%

- Search engine for keywords (full-text)
  - D: 46%

- Search engine that allows searching for keywords in two or more specific domain names
  - E: 13%
Select the three most important web archive functionalities in your opinion.

n=290

(1)

- Search engine for keywords: 29.0%
- Search engine for images: 11.7%
- Collections of archived websites specific to my research area: 10.1%
- The possibility to download data in multiple data formats: 9.2%
- Search engine for URLs: 7.0%
- The possibility to download data in multiple ways: 5.2%
- Tools that help with the discovery of patterns, trends or relationships in data: 4.6%
Select the three most important web archive functionalities in your opinion.

(2)

- A tool that enables the creation and extraction of a corpus from a larger collection
- Overview of personal search history
- Tools that help me analyse the data
- Side-by-side visualization of a website throughout the years
- Search engine that allows searching for keywords in two or more specific URLs
- Search engine for videos
- Auto-completion of search fields in search engine
How satisfied are you with this or these webarchive(s)?

n=179

- Very dissatisfied: 1.7%
- Moderately dissatisfied: 2.8%
- Slightly dissatisfied: 2.2%
- Neutral: 6.1%
- Slightly satisfied: 7.8%
- Moderately satisfied: 14.0%
- Very satisfied: 4.5%
Which other functionalities would you like to see in the Belgian web archive? (max. 3 options)

- The possibility to download data in multiple data formats: 4%
- The possibility to create and extract a corpus from a larger collection: 25%
- Overview of personal search history: 4%
- Tools that help me analyse the data: 17%
- Tools that help with the discovery of patterns, trends or relationships in data: 21%
- Side-by-side visualisation of a website throughout the years: 13%
- API access (Application Programming Interface): 17%
In working with data from web archives, what challenges have you faced? \( n=292 \)

(1)

Challenges in accessing the data (e.g. web archive data was only available in the reading room).

A lack of appropriate descriptive information about the web archives (e.g. lack of metadata).

Difficulty in finding relevant materials (e.g. limited search options).

Lack of content / collections in my research area.
In working with data from web archives, what challenges have you faced? n=292

Data is in such a form that it is hard to process (messiness).

Lack of transparency and clarity regarding intellectual property rights.

The size of the data is too large.

Challenges conducting data analysis (e.g. lack of resources or tools to analyze the data).
Other interesting topics:

- Use of data from web archives
- Most important methods of analysis
- Most helpful metadata
- Subjects or topics of most interest
- Contribution to the selection of material
- Format to download the data
- Receiving guidance
- Collecting own web archives
- Quality control
Online dashboard:

bit.ly/surveypromise
“Clearly Google is the norm for the participants. They expected a search engine to work on full text, just like Google.”

“Access and use (...) remains a perceived area of need. Likewise, metadata is identified as an area that would benefit from ongoing knowledge-sharing around best practices. Social media and quality assurance continue to be recognized as areas for which better and more accessible tools are needed.”

“(...) better discoverability options for the archived content, data selection and management, as well as better access to more ways of analysing the data is needed.”
Coordination of Webarchiving in the Netherlands

Jesse de Vos – Beeld & Geluid
Using Web archives for researchers: experience, challenges, issues

Valérie Schafer – C2DH
Saving the Belgian Web

Web archiving practices, research opportunities and limitations

Patricia Blanco – MSc in Digital Humanities (KU Leuven)
Web Archives?

http://bit.do/webarchiving
ACCESS & USE:
Testing tools to research in web archives

SELECTION:
Creation of thematic collections

PRESERVATION

QUALITY ASSURANCE:
manual comparison

HARVEST: collection of websites using Heritrix

Technical
Legal
Obstacles
SELECTION: Creating a thematic collection
Portugalnet

Hispagenda

Italiansnet

http://bit.do/promise3
<table>
<thead>
<tr>
<th></th>
<th>Spanish</th>
<th>Italian</th>
<th>Portuguese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial seed list</td>
<td>41</td>
<td>40</td>
<td>28</td>
</tr>
<tr>
<td>URLs downloaded</td>
<td>1,772,554</td>
<td>2,610,369</td>
<td>2,604,372</td>
</tr>
<tr>
<td>URLs queued</td>
<td>87,438</td>
<td>7,938,266</td>
<td>14,166,405</td>
</tr>
<tr>
<td>Size</td>
<td>81 GB</td>
<td>163 GB</td>
<td>865 GB</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Spanish</th>
<th>Italian</th>
<th>Portuguese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of WARC files</td>
<td>22</td>
<td>11</td>
<td>36</td>
</tr>
<tr>
<td>Total size(^2)</td>
<td>~70 GB</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Wayback Machine

Please try our Beta Wayback Machine - Now with Site Search!

DONATE

Explore more than 279 billion web pages saved over time

Tools

Wayback Machine Availability API
Build your own tools.

WordPress Broken Link Checker
Banish broken links from your blog.

404 Handler for Webmasters
Help users get where they were going.

Subscription Service

Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. Visit Archive-It to build and browse the collections.

Save Page Now

http://

Save Page

Capture a web page as it appears now for use as a trusted citation in the future.

Only available for sites that allow crawlers.
Thank you

Questions?
P.BLANCO.NUNEZ@GMAIL.COM
Questions, wrap-up and discussion