Practical Experiences and New Challenges in Web Crawling

Christopher Schröder, Martin Potthast
Leipzig University
ASV - temir.org - webis.de

DLR Open Search Colloquium · June 8, 2021
Web Crawling for the Wortschatz
Welcome to the Leipzig Corpora Collection / Deutscher Wortschatz

a project of the Natural Language Processing Group at the Institute of Computer Science at Leipzig University.

Corpora portal

The international corpora portal offers access to more than 400 corpora of the Leipzig Corpora Collection (LCC) in more than 250 languages.

To the corpora portal

CURL portal

On this website you can contribute to corpus collection for under-resourced languages by simply entering a URL.

To the CURL portal

Words of the day

The words of the day based on a selection of newspaper and news services. Daily at 7 am and available as RSS!

RSS 2.0

To the words of the day

CLARIN corpora portal

The Wortschatz’s CLARIN corpora portal offers access to all corpora of the Leipzig Corpora Collection (LCC) that we already integrated into the CLARIN infrastructure.

To the CLARIN corpora portal

ASV Online Toolbox

The ASV Toolbox is a modular collection of tools for the exploration of written language data.

Corpus statistics

The corpus and language statistics contain analyses about various aspects of natural language based on our corpora.
Goal: Facilitate the study of contemporary language

Corpora in 252 languages  [corpora.uni-leipzig.de]

8 billion unique sentences in German

20% share of Leipzig University’s web traffic  [wortschatz.uni-leipzig.de]
Web Crawling

Web crawling is the process of systematically traversing the web. Usually, visited web sites, or at least parts thereof, are saved for the purpose of web indexing.

[Wikipedia]
Web Crawling

Web crawling is the process of systematically traversing the web. Usually, visited web sites, or at least parts thereof, are saved for the purpose of web indexing. [Wikipedia]

Uses of web crawling:

- Exploring the web
- Indexing the web
- Acquiring data
- Archiving the web
Web Crawling
Selectivity

A crawler implements a selection policy determining which pages are crawled.

- **Crawl seeds** (Where to begin?)
  Initialization of a crawler frontier with previously collected URLs, so-called seeds.

- **Crawl target** (What to crawl?)
  Simple answer: Everything. More specifically: Every document for which a search engine’s user might search (“Where was that document again?”). For web search engines, only few exceptions apply. In general, predicting universal non-usefulness of documents is difficult.

- **Crawl priority** (What first?)
  Web pages that are predictably more important to the search engine’s users than others. Web sites may be judged as a whole. In particular, pages comprising high-quality content.

- **Crawl filtering** (What to avoid?)
  “Spider traps”, and web pages from web sites whose owners harbor malicious intents toward the search engine, or its users, such as spam pages.
Web Crawling for the Wortschatz

Selection Policy

The crawler is restricted to only follow links on a specific top-level domain, or to stay within the domains of specific websites.

Web (Time Slice)

- Execution: Once a year
- Restriction: TLD (e.g., only “.de” URLs)

News (Time Slice)

- Execution: Once a year
- Restriction: Language and Seeds
  (e.g., “fra” and a list of French news sites)

News (Daily)

- Execution: Daily
- Crawls a given URL list obtained via RSS/Atom
Web Crawling for the Wortschatz
Common Challenges

Stopping criteria

- After a fixed amount of time
- If less than one WARC file (1 GB compressed) is generated per day
- If crawl logs show too many low-quality / unusable links (subjective, manual)

Crawl frontier

- Typically outgrows the crawler’s throughput
- For an unmonitored operation, a large storage network is required

“Low-value” sites

- Galleries, Shops, Calendars...
- Spider traps (e.g., dynamically creating many URLs pointing at the same site)
- Spam networks
Web Crawling for the Wortschatz

Maintenance

Manual orchestration

- Automation is tricky, e.g., when automating a stopping condition

Complaints

- Crawler crawls licensed material
- Crawler adds many items to the cart in a webshop
- Missing or misconfigured robots.txt

Hardware / Software issues

- Some things are just out of your control
- Recovery can be time-consuming
Web Archiving
Web Archiving

Web archiving is the process of collecting portions of the World Wide Web to ensure the information is preserved in an archive for future researchers, historians, and the public. [Wikipedia]
Web Archiving

Web archiving is the process of collecting portions of the World Wide Web to ensure the information is preserved in an archive for future researchers, historians, and the public. [Wikipedia]

Providers and initiatives: [Wikipedia] 96 listed

- Internet Archive: Wayback Machine
- National libraries (e.g., DNB)
- Commercial services (e.g., Google)
- Open source and activism (“archivist”)
- Societies: ICA [ica.org], IIPC [netpreserve.org]
- People: Brewster Kahle, Vint Cerf
Web Archiving

Web archiving is the process of collecting portions of the World Wide Web to ensure the information is preserved in an archive for future researchers, historians, and the public. [Wikipedia]

Providers and initiatives: [Wikipedia] (96 listed)

- Internet Archive: Wayback Machine
- National libraries (e.g., DNB)
- Commercial services (e.g., Google)
- Open source and activism (“archivist”)  
- Societies: ICA [ica.org], IIPC [netpreserve.org]
- People: Brewster Kahle, Vint Cerf

Customers:

- Archives and Libraries
- Social science
- Humanities
- Computer science
- Business and Government
Web Archiving

Web archiving is the process of collecting portions of the World Wide Web to ensure the information is preserved in an archive for future researchers, historians, and the public.  [Wikipedia]

Providers and initiatives:  [Wikipedia]  (96 listed)

- Internet Archive:  [Wayback Machine]
- National libraries (e.g., DNB)
- Commercial services (e.g., Google)
- Open source and activism (“archivist”)
- Societies: ICA  [ica.org], IIPC  [netpreserve.org]
- People: Brewster Kahle, Vint Cerf

Customers:

- Archives and Libraries
- Social science
- Humanities
- Computer science
- Business and Government

Web Archiving
Building Blocks

Four steps of web archiving:

1. **Select.** Decision what portion of the web shall be archived.
2. **Capture.** Downloading and storing live web content.
3. **Preserve.** Checking downloads and converting them to stable format.
4. **Playback.** Enable access and interaction with archived web content.
Web Archiving
Building Blocks

Four steps of web archiving:

1. **Select.** Decision what portion of the web shall be archived.
2. **Capture.** Downloading and storing live web content.
3. **Preserve.** Checking downloads and converting them to stable format.
4. **Playback.** Enable access and interaction with archived web content.

Key performance indicators:

- Representativity of selection
- Completeness of capture
- Longevity of preservation
- Accuracy and fidelity of playback
Web Archiving

Building Blocks

Four steps of web archiving:

1. Select. Decision what portion of the web shall be archived.
2. Capture. Downloading and storing live web content.
3. Preserve. Checking downloads and converting them to stable format.
4. Playback. Enable access and interaction with archived web content.

Key performance indicators:

- Representativity of selection
- Completeness of capture
- Longevity of preservation
- Accuracy and fidelity of playback

A web page is a complex piece of client-server software.
Web Archiving

Example: Personalization

What you see
Web Archiving
Example: Personalization

What you see

What someone else sees
Web Archiving
Example: Capturing HTML only

Original

![Pink blouses](image-url)
Web Archiving
Example: Capturing HTML only

Original

Archive
Web Archiving
Example: Capturing via a Browser’s Save Page Feature
Web Archiving

Example: Capturing via a Browser’s Save Page Feature
Web Archiving
Example: Capturing via a Browser’s Save Page Feature
Web Archiving

Example: Capturing via a Browser’s Save Page Feature

Original

Archive
Web Archiving

Example: Capturing via a Browser’s Save Page Feature
Web Archiving

Example: Capturing via a Browser’s Save Page Feature

Original

Archive
Web Archiving
Example: Capturing via a Browser’s Save Page Feature

Original

Archive
Web Archiving

Example: Capturing via a Browser’s Save Page Feature
Web Archiving
Example: Capturing via a Browser’s Save Page Feature

Original

Archive
Web Archiving
Example: Capturing via a Browser’s Save Page Feature
Web Archiving
Example: Capturing via a Browser’s Save Page Feature
Web Archiving

Example: Capturing via a Browser’s Save Page Feature
Web Archiving
Example: Capturing via the Wayback Machine

Look up page in / save page using the Internet Archive (web.archive.org)
Web Archiving
Example: Capturing via the Wayback Machine

Archive
Web Archiving
Example: Capturing via the Wayback Machine
Web Archiving
Example: Capturing via the Wayback Machine
Web Archiving

Webis Web Archiver [github.com] [hub.docker.com]

- Scriptable user-page interactions
- Reproducibility of web corpora, user experience, user behavior
- Compatibility with other web archiving software
- Automatic archive quality assessment

Capturing and Preservation:
Web Archiving
Webis Web Archiver [github.com] [hub.docker.com]

- Scriptable user-page interactions
- Reproducibility of web corpora, user experience, user behavior
- Compatibility with other web archiving software
- Automatic archive quality assessment

Playback:
Remarks:

- A readily executable configuration of the archiver is encapsulated in a Docker image. Docker also ensures the reproducibility of our archiver’s execution environment by fixing the versions of each software library and especially the browser. Moreover, all of 2 GB worth of fonts available to Ubuntu are installed.

- The virtual screen software xvfb is used to run the browser without requiring a physical screen, allowing for server-side execution, and with some additions even recording videos.

- The Selenium browser automation software serves as an interface between the simulation script and the browser.

- We employ a current version of Google Chromium, but others are supported as well.

- During playback, the FakeTime Preload library is used to pretend to the browser that it runs at the time of archiving, which affects all JavaScript calls that use the current date.

- The browser is set up to communicate with an instance of BrowserMob Proxy, which is used by the script to learn when network traffic ceases.

- During archiving, the BrowserMob Proxy communicates with the Internet via an instance of the warcprox proxy, which stores all requests and the corresponding responses that pass through it in a standard WARC archive file.

- During playback, a local server is started that pretends to be a proxy, but actually attempts to retrieve the previously recorded responses to requests made by a to-be-reproduced web page from its WARC files.

- Our archiver currently allows to choose from three different tools for reproducing web pages: warcprox, Python WayBack (pywb), and a custom implementation.
Remarks: (continued)

- Example of a user simulation script:

```java
@Override
protected void executeInteraction(final Browser browser, final String url, final Path outputDirectory)
    throws Throwable {
    WebDriver window = browser.openWindow(url);
    this.scrollDown(browser, window); \ 25 times
    this.saveSnapshot( \ PNG and HTML
        browser, window, outputDirectory);
    this.doCoolStuff( \ Your analysis
        browser, window, outputDirectory);
}
```
Web Archiving
Quality Assurance

- Capturing all content: embedded multimedia, dynamic content, mobile view...
- Content overlay detection and interactive removal
- Content error detection
- Fuzzy request-response matching during playback
- Playback accuracy and fidelity
Web Archiving
Quality Assurance

- Capturing all content: embedded multimedia, dynamic content, mobile view...
- Content overlay detection and interactive removal
- Content error detection
- Fuzzy request-response matching during playback
- Playback accuracy and fidelity

ERROR: 404
Someone's chewed through something they shouldn't have...

Quick! While no one's looking
Click this little button for TFLN texts and updates

We also send dope emails
Forget to check the site? We'll send our best texts, memes and weekly shenanigans straight to your inbox.

already like us? well thanks! click here.
Web Archiving
Quality Assurance

- Capturing all content: embedded multimedia, dynamic content, mobile view...
- Content overlay detection and interactive removal
- Content error detection
- Fuzzy request-response matching during playback
- Playback accuracy and fidelity
Web Archiving
Quality Assurance

- Capturing all content: embedded multimedia, dynamic content, mobile view...
- Content overlay detection and interactive removal
- Content error detection
- Fuzzy request-response matching during playback
- Playback accuracy and fidelity
Web Archiving
Quality Assurance

- Capturing all content: embedded multimedia, dynamic content, mobile view...
- Content overlay detection and interactive removal
- Content error detection
- Fuzzy request-response matching during playback
- Playback accuracy and fidelity
Conclusion

Summary:

- Projekt Deutscher Wortschatz: Resources to study contemporary language.
- Webis Web Archiver: Tool to capture web pages with high fidelity.
Conclusion

Summary:

- Projekt Deutscher Wortschatz: Resources to study contemporary language.
- Webis Web Archiver: Tool to capture web pages with high fidelity.

Take-away messages:

- Web crawling setups differ considerably depending on one’s goals.
- Maintaining and running crawlers comes with manual overhead.
- Web archiving presents unique challenges on top of and beyond crawling.
- For commercial search engines, crawling means archiving.
Conclusion

Summary:

- Projekt Deutscher Wortschatz: Resources to study contemporary language.
- Webis Web Archiver: Tool to capture web pages with high fidelity.

Take-away messages:

- Web crawling setups differ considerably depending on one’s goals.
- Maintaining and running crawlers comes with manual overhead.
- Web archiving presents unique challenges on top of and beyond crawling.
- For commercial search engines, crawling means archiving.

Open Search Initiative:

- Can we afford not to archive, but only to crawl?
Conclusion

Summary:

- Projekt Deutscher Wortschatz: Resources to study contemporary language.
- Webis Web Archiver: Tool to capture web pages with high fidelity.

Take-away messages:

- Web crawling setups differ considerably depending on one’s goals.
- Maintaining and running crawlers comes with manual overhead.
- Web archiving presents unique challenges on top of and beyond crawling.
- For commercial search engines, crawling means archiving.

Open Search Initiative:

- Can we afford not to archive, but only to crawl?

Thank you!
Appendix
Welcome to the Leipzig Corpora Collection / Deutscher Wortschatz

a project of the Natural Language Processing Group at the Institute of Computer Science at Leipzig University.

Corpora portal
The international corpora portal offers access to more than 400 corpora of the Leipzig Corpora Collection (LCC) in more than 250 languages.

To the corpora portal

CURL portal
On this website you can contribute to corpus collection for under-resourced languages by simply entering a URL.

To the CURL portal

Words of the day
The words of the day based on a selection of newspaper and news services. Daily at 7 am and available as RSS!

To the words of the day

CLARIN corpora portal
The Wortschatz's CLARIN corpora portal offers access to all corpora of the Leipzig Corpora Collection (LCC) that we already integrated into the CLARIN infrastructure.

To the LCC's CLARIN corpora portal

ASV Online Toolbox
The ASV Toolbox is a modular collection of tools for the exploration of written language data.

To the online toolbox

Corpus statistics
The corpus and language statistics contain analyses about various aspects of natural language based on our corpora.

To the corpus statistics