

Practical Experiences and New Challenges in Web Crawling

Christopher Schröder, Martin Potthast
Leipzig University
[ASV - temir.org](http://ASV-temir.org) - webis.de

DLR Open Search Colloquium · June 8, 2021

Web Crawling for the Wortschatz

Search in more than 46 million sentences of German newspaper material

Enter a word



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](#).

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.

Search in more than 46 million sentences of German newspaper material

Enter a word



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.

CURL portal

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

Words of the day

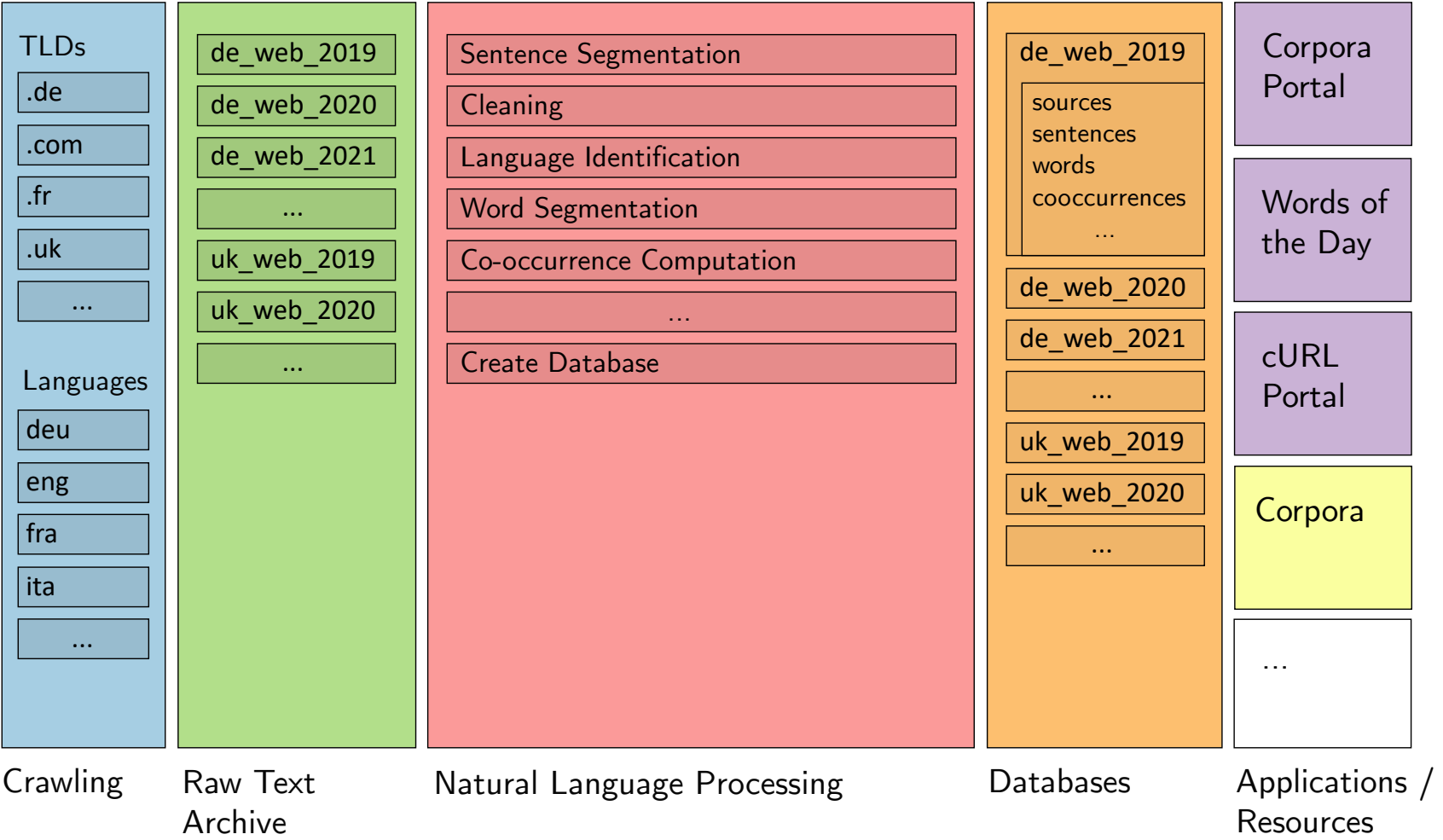
The words of the day based on a selection of newspaper and news services. Daily at 7 am and available as RSS!

[RSS 2.0](#) 

- ❑ 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]

Projekt Deutscher Wortschatz

Architecture



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

Use cases: Artificial intelligence. Automatic HTML generation. Automatic mobile device adjustment. Change detection. Clutter reduction. Content extraction. Design optimization. Discussion analysis. Duplicate detection. Entity mining. Error page detection. Genre detection. Indexing. Language analysis. Link analysis. Person disambiguation. Phishing detection. Profiling. Ranking. Record mining. Screen reading. Snippet generation. Spam detection. Summarization. Template detection. Text reuse analysis. Trend prediction. Usage mining. Wrapper induction.

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

Pinterest Principali utilizzi e cookie per avere la migliore esperienza possibile. OK

Registrali **Accedi**

Camicette rosa

Petali di rosa **Camicetta da lavoro** **Camicia con ruffe** **Camicette** **Abbigliamento a fantasia floreali** **Stile primavera** **Peplum dress** **Giacca impermeabile**

2016 t-shirt per le donne moda rosa top camicetta dalle signore dell'ufficio Casual lavoro tops stili...

Borsa stampata con colletto Rosa Camella - Camicette - Bimba - Obabbi

28 idee da mettere in pratica: Come ridare NUO...
ESKI PENYEVİ ŞİK ELDESEYE DÖNÜŞTÜRMEK...
[https://rendinyapmani.com/veste/...](https://rendinyapmani.com/veste/)

Crochet Bag, Little Bag, Little Girl Crochet Purse...
Borsa Little Little Girl ardetretto borsa di laamobdyCrochet

Camicetta a due strati
Camicetta a due strati - Rosa chiaro - DONNA | H&M IT

2016 t-shirt per le donne moda rosa top camicetta dalle signore dell'ufficio Casual lavoro tops stili...

Cheap Trasporto libero 2016 di Estate Donna New Fashion Casual Solid Slash collo Chiffon Camic...

T-shirt da donna raglan stampa "3 gatti" rosa

Camicette, Top & Tuniche - T-SHIRT DA DONNA RAGLAN STAMPA "3 GATTI" ROSA...

Borsa stampata con colletto Rosa Camella - Camicette - Bimba - Obabbi

28 idee da mettere in pratica: Come ridare NUO...

Crochet Bag, Little Bag, Little Girl Crochet Purse...

Tunica da donna "gatti e cuori" rosa
Camicette, Top & Tuniche - TUNICA DA DONNA "GATTI E CUORI" ROSA - un...

Tunica da donna stampa "miaoo righe" rosa
Camicette, Top & Tuniche - TUNICA DA DONNA STAMPA "MIAOO RIGHE" ROSA - un...

Web Archiving

Example: Personalization

What you see

Pinterest utilizza i cookie per offrirti la migliore esperienza possibile. OK

Registrati Accedi

Camicette rosa

Petali di rosa Camicetta da lavoro Camicia con ruffe Camicette Abbigliamento a fantasia floreali Stile primavera Pepum dress Giacca impermeabile

2016 t-shirt per le donne moda rosa top camicette dalle signore dell'ufficio Casual lavoro tops stili...

Blusa stampata con colletto Rosa Camella - Camicette - Blinle - Obabli

28 idee da mettere in pratica: Come ridare NUO...
ESKI PENYEVİ ŞİK ELDESEYE DÖNÜŞTÜRMEK...
https://www.diyapiana.com/veste/...

Crochet Bag, Little Bag, Little Girl Crochet Purse, ...
Borsa Little Little Girl (sindretto borsa di lalalayCrochet

Camicetta a due strati
Camicetta a due strati - Rosa chiaro - DONNA | H&M IT

Cheap Trasporto libero 2016 di Estate Donna New Fashion Casual Solid Slash collo Chiffon Camice...

T-shirt da donna raglan stampa "3 gatti" rosa

Camicette, Top & Tuniche - T-SHIRT DA DONNA RAGLAN STAMPA "3 GATTI" ROSA...

Camicette, Top & Tuniche - TUNICA DA DONNA "GATTI e CUORI" ROSA

Tunica da donna stampa "miaoo righe" rosa

Camicette, Top & Tuniche - TUNICA DA DONNA STAMPA "MIAO RIGHE" ROSA - un...

What someone else sees

Pinterest uses cookies to help give you the best experience we can. OK

Sign up Log in

Women's fashion > Pink blouses

Pink blouses

Blouse designs silk Pink saree blouse Silk saree blouse designs Blouse designs catalogue Saree blouse designs 2017 Women's pink style

Chloe Pink Contrast Painted Collar Blouse (\$22) Liked on Polyvore

Milly all the-shoulder blouses and pink top See More

M Missoni Silk Tie Neck Blouse (8,665 HNL) Liked on Polyvore

Pierced V Neck Cold
Elegant Navy Blue Chanderi Saree with Bright Pink Blouse Matel See More

Elegant Navy Blue Chanderi Saree with Bright Pink Blouse Matel See More

Chowishi Kiss Me Bow Top in Candy Pink See More

Lena Hoschek and suspenders, tie.

Web Archiving

Example: Capturing HTML only

Original

The screenshot shows a Pinterest search results page for 'Pink blouses'. At the top, there is a dark navigation bar with the Pinterest logo, the word 'Pinterest', and buttons for 'Sign up' and 'Log in'. Below this, a breadcrumb trail reads 'Women's fashion > Pink blouses'. The main heading is 'Pink blouses', followed by a row of filter tabs: 'Blouse designs silk', 'Pink saree blouse', 'Silk saree blouse designs', 'Blouse designs catalogue', 'Saree blouse designs 2017', and 'Women's pink style'. The page displays a grid of image pins. Each pin includes a thumbnail image of a blouse, a title, a description, and a 'See More' link. Some pins also show the number of likes and the user's profile picture. The pins feature various styles of pink blouses, including sleeveless, off-the-shoulder, long-sleeved, and saree blouses. The layout is clean and organized, typical of a search results page on a social media platform.

Web Archiving

Example: Capturing HTML only

Original

Archive

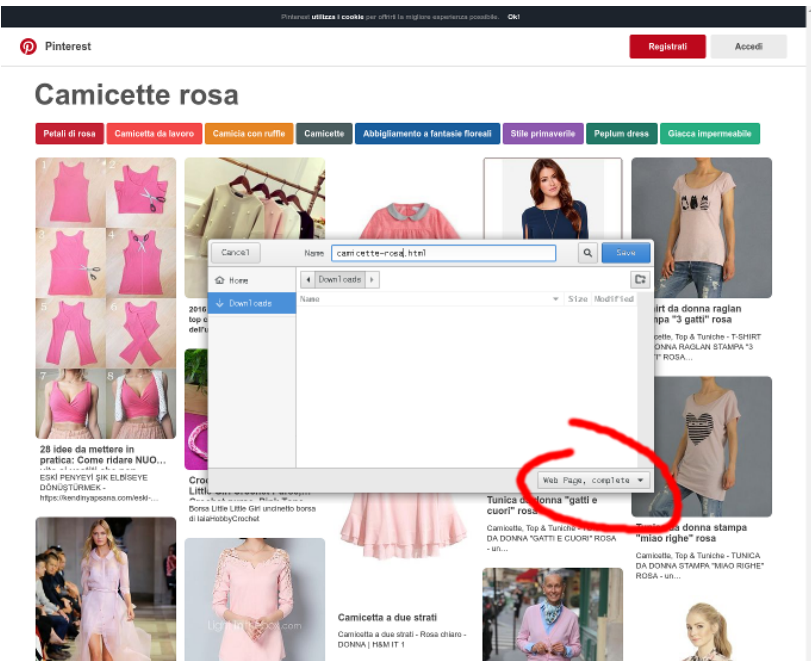
The screenshot shows the original Pinterest page for 'Pink blouses'. At the top, there is a navigation bar with the Pinterest logo, a 'Sign up' button, and a 'Log in' button. Below the navigation bar, the page title 'Pink blouses' is displayed. A horizontal menu contains several filter categories: 'Blouse designs silk', 'Pink saree blouse', 'Silk saree blouse designs', 'Blouse designs catalogue', 'Saree blouse designs 2017', and 'Women's pink style'. The main content area features a grid of image pins. Each pin includes a thumbnail image of a blouse, a title, a description, and a 'See More' link. For example, one pin is titled 'Chloee Pink Contrast Painted Collar Blouse (32)' and another is 'Milly off the-shoulder blouses and pink top'. The page layout is clean and modern, with a white background and clear typography.

The screenshot shows the archived version of the Pinterest page for 'Pink blouses'. The layout is identical to the original page, but the content is significantly degraded. The navigation bar and filter menu are present, but the main content area is mostly blank or contains very low-resolution, pixelated images. The text is also rendered in a way that is difficult to read, appearing as a series of small, disconnected characters and words. This illustrates the challenges of web archiving when only HTML is captured, as the visual and interactive elements of the page are lost.

Web Archiving

Example: Capturing via a Browser's Save Page Feature

Original



Web Archiving

Example: Capturing via a Browser's Save Page Feature

Original

Archive

This screenshot shows the original Pinterest search results for 'Camicette rosa'. The page features a navigation bar with the Pinterest logo, a search bar, and buttons for 'Registrati' and 'Accedi'. Below the search bar, the title 'Camicette rosa' is displayed, followed by a row of filter buttons: 'Petali di rosa', 'Camicetta da lavoro', 'Camicia con ruffe', 'Camicette', 'Abbigliamento a fantasia floreali', 'Stile primaverale', 'Pepum dress', and 'Giacca impermeabile'. The main content area is a grid of image thumbnails, each with a small caption. The thumbnails show various styles of pink blouses, including long-sleeved and short-sleeved options, some with ruffles or prints. The captions are truncated, showing only the beginning of the text for each item.

This screenshot shows the archived version of the Pinterest page for 'Camicette rosa'. The layout is identical to the original page, including the navigation bar, search bar, and filter buttons. The main content area is a grid of image thumbnails, each with a small caption. The thumbnails show various styles of pink blouses, including long-sleeved and short-sleeved options, some with ruffles or prints. The captions are truncated, showing only the beginning of the text for each item. The overall appearance is a faithful reproduction of the original page's content and structure.

Web Archiving

Example: Capturing via a Browser's Save Page Feature

Original

Original webpage content:

- Camicetta a due strati**
Camicetta a due strati - Rosa chiaro - DONNA | H&M IT 1
- Camicia Da donna Per uscire Formali Ufficio...**
Camicia Da donna Per uscire Formali Ufficio Serenale Moda 2019 Sforzicato Per tutte le stagioni...
- Camicetta Incrociata Scollo V Maniche A Campana...**
- Camicetta senza maniche**
Camicetta senza maniche | H&M
- Il colore e' poeta dell'anima**
da il colore e' poeta dell'anima
- COLROVE Donna Camicetta Coreano 58% di Moles Camicette Donne Streetwear Rosa Mock Ne...**
- Camicetta, Top & Tuniche - TUNICA, DA DONNA STAMPA "MAGO RICHI" ROSA - ur...**
- Come vestirei dopo i 50 anni: idee look.. strapitosi!**
- Camicia In georgette di seta, nero rosa - Diffusione Tessile**
- Occasioni! Lotto di 13 pezzi da donna ragazza primavera estate taglia S M Canottiere t-shirt vari...**
- SALVE RAGAZZE in questi giorni sole tepido belle giornate , i colori dei fiori rallegrano le nostre...**
- Sfilata Gucci Milano - Collezioni Primavera Estate 2013 - Vogue**

Archive

Archived webpage content:

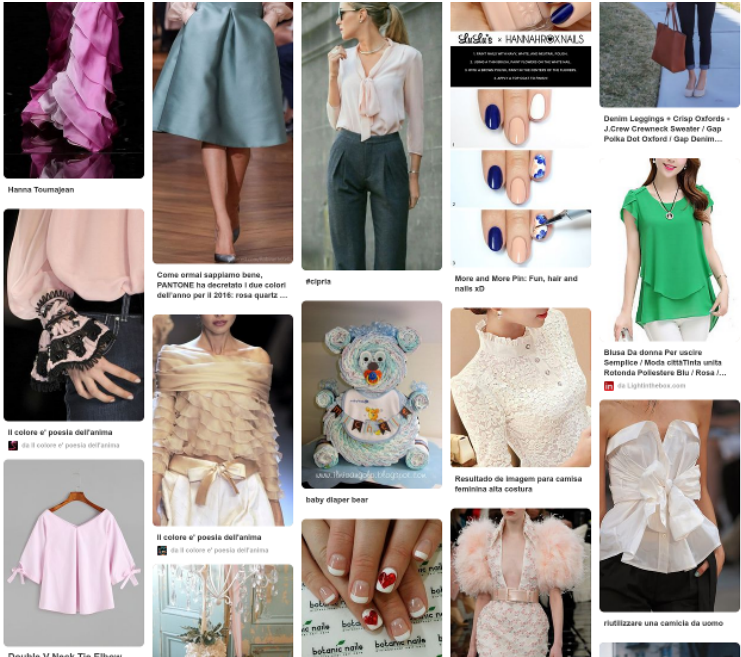
- Camicetta a due strati**
Camicetta a due strati - Rosa chiaro - DONNA | H&M IT 1
- Camicia Da donna Per uscire Formali Ufficio...**
Camicia Da donna Per uscire Formali Ufficio Serenale Moda 2019 Sforzicato Per tutte le stagioni...
- Camicetta Incrociata Scollo V Maniche A Campana...**
- Camicetta senza maniche**
Camicetta senza maniche | H&M
- Il colore e' poeta dell'anima**
da il colore e' poeta dell'anima
- COLROVE Donna Camicetta Coreano 58% di Moles Camicette Donne Streetwear Rosa Mock Ne...**
- Camicetta, Top & Tuniche - TUNICA, DA DONNA STAMPA "MAGO RICHI" ROSA - ur...**
- Come vestirei dopo i 50 anni: idee look.. strapitosi!**
- Camicia In georgette di seta, nero rosa - Diffusione Tessile**
- Occasioni! Lotto di 13 pezzi da donna ragazza primavera estate taglia S M Canottiere t-shirt vari...**
- SALVE RAGAZZE in questi giorni sole tepido belle giornate , i colori dei fiori rallegrano le nostre...**
- Sfilata Gucci Milano - Collezioni Primavera Estate 2013 - Vogue**

Web Archiving

Example: Capturing via a Browser's Save Page Feature

Original

Archive

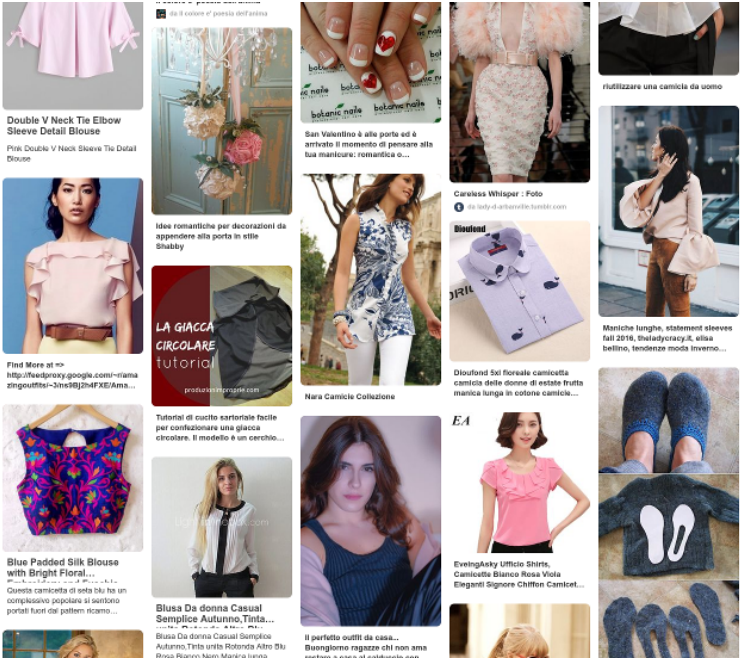


Web Archiving

Example: Capturing via a Browser's Save Page Feature

Original

Archive

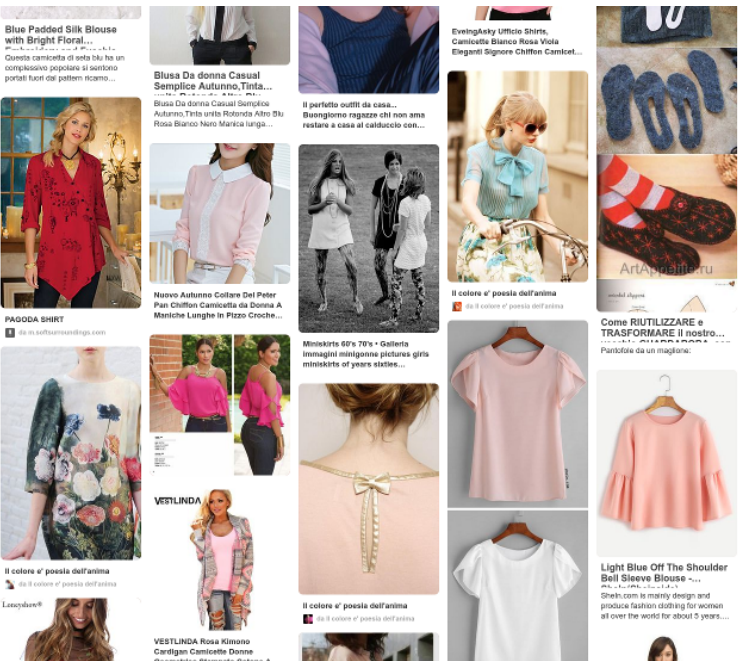


Web Archiving

Example: Capturing via a Browser's Save Page Feature

Original

Archive

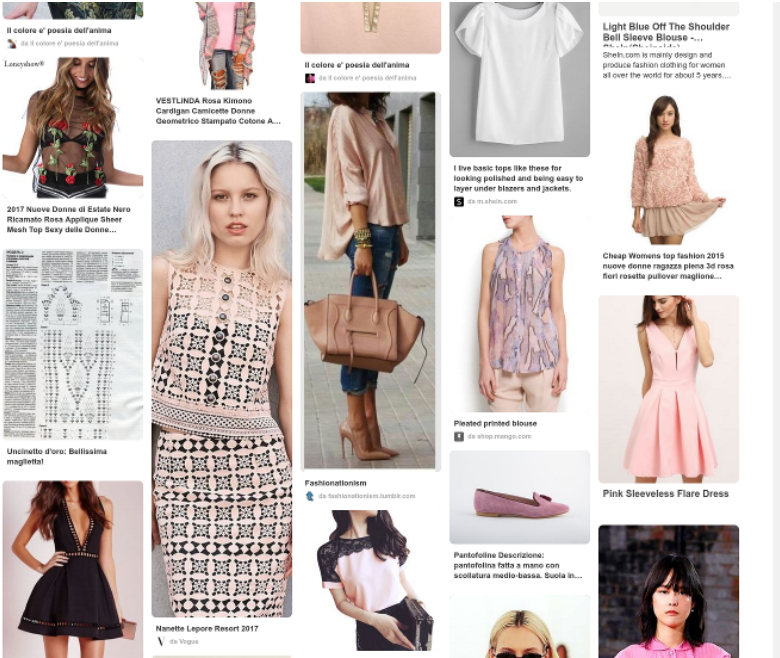


Web Archiving

Example: Capturing via a Browser's Save Page Feature

Original

Archive

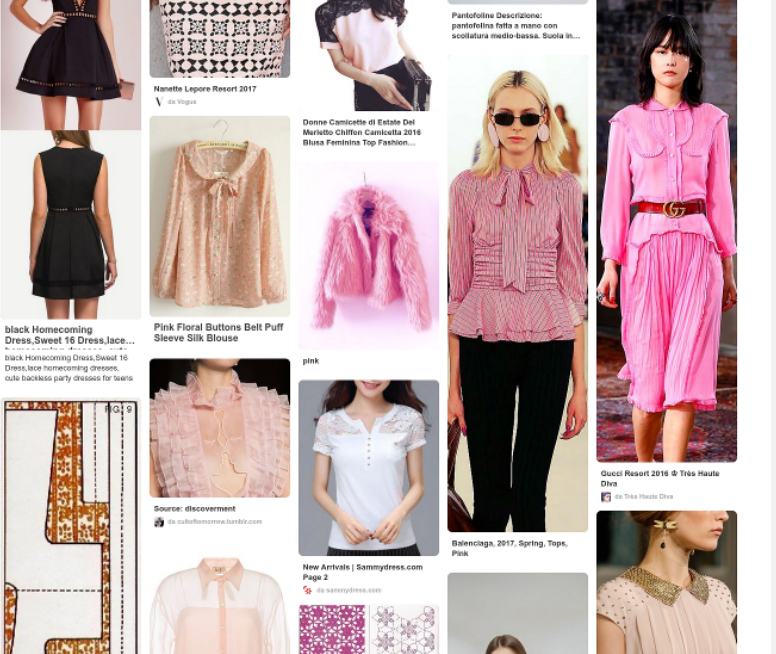


Web Archiving

Example: Capturing via a Browser's Save Page Feature

Original

Archive

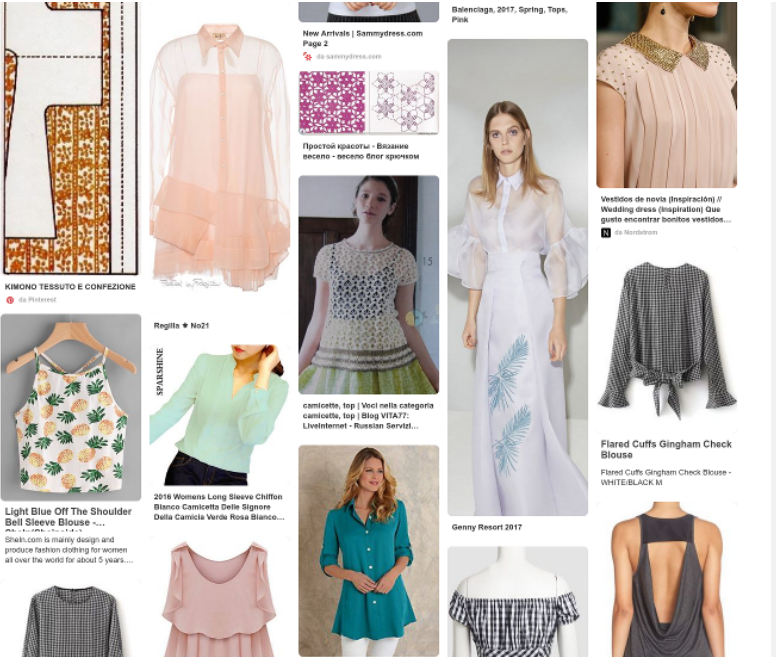


Web Archiving

Example: Capturing via a Browser's Save Page Feature

Original

Archive

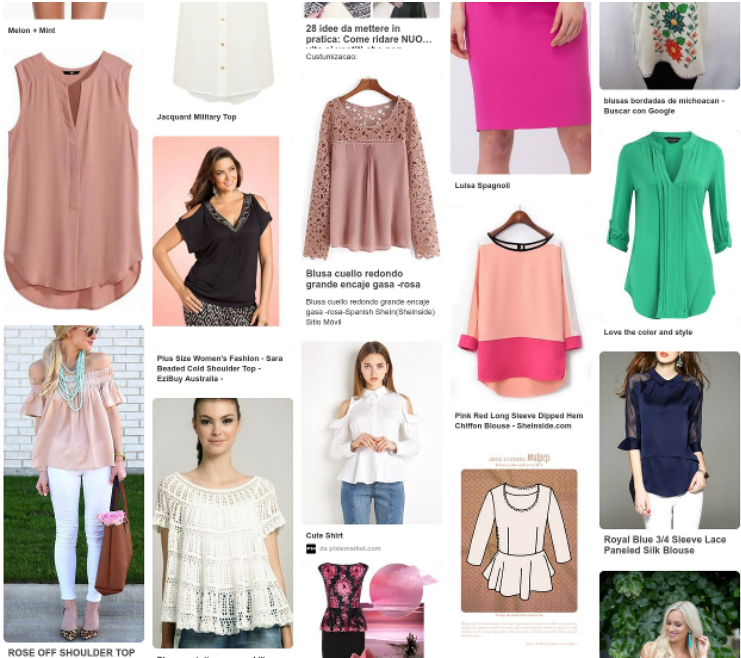


Web Archiving

Example: Capturing via a Browser's Save Page Feature

Original

Archive



Web Archiving

Example: Capturing via the Wayback Machine

Archive

The screenshot shows a browser window with the Wayback Machine interface. The address bar contains the URL <https://www.pinterest.com/waybackmachine/>. The date and time are set to May 08, 2016, at 10:17. The page content is a Pinterest board titled "Camicette rosa" (Pink Blouses). The board features a grid of images and descriptions for various pink blouses and tops. The navigation bar includes categories like "Petali di rosa", "Camicetta da lavoro", "Camicia con nuffe", "Camicette", "Abbigliamento a fantasia floreale", "Stile primaverile", "Pigiama donna", and "Giacca impermeabile".

Camicette rosa

2016 t-shirt per le donne moda rosa top camicette della signora dell'ufficio Casual lavoro top stili...

Blusa stampata con colletto Rosa Camicetta - Camicetta - Bimba - Giacca

Camicetta a due strati

T-shirt da donna raglan stampa "3 gatti" rosa Camicetta, Top & Tuniche - T SHIRT DA DONNA RAGLAN STAMPA "3 GATTI" ROSA...

Tunica da donna stampa "miao righe" rosa Camicetta, Top & Tuniche - TUNICA DA DONNA STAMPA "MIAO RIGHE" ROSA - it...

28 idee da mettere in pratica: Come ridere NUO... EDU PENNY DI ELDOVE DONATUNEX... <https://rendityapoco.com/est/>

Crochet Bag, Little Bag, Little Girl Crochet Purse... Donna Little Girl con anchetto borsa di laaniseyCrochet

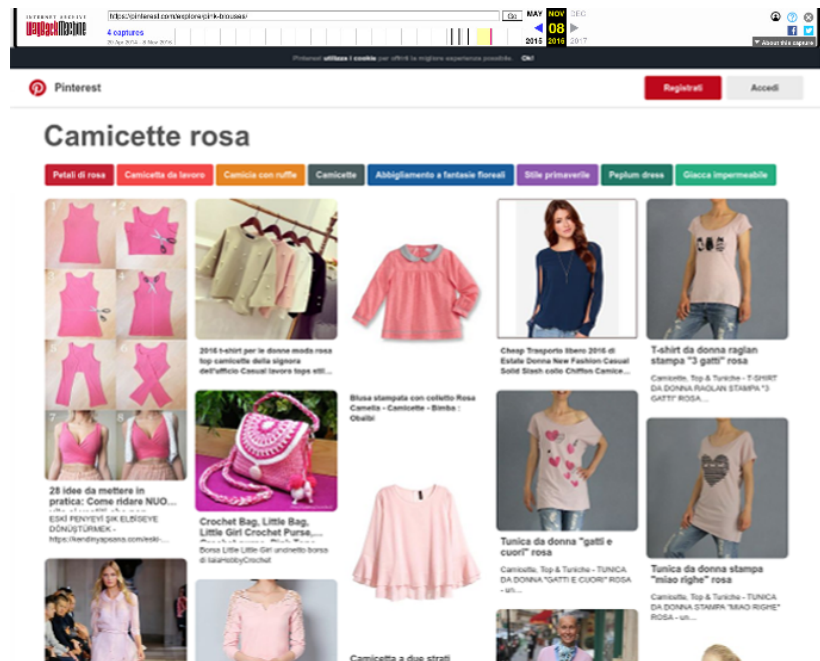
Chap Transporte Item 2016 di Estate Donna New Fashion Casual Solid Strain colte Clotho Camisa...

Tunica da donna "gatti e cuori" rosa Camicetta, Top & Tuniche - TUNICA DA DONNA "GATTI E CUORI" ROSA - it...

Web Archiving

Example: Capturing via the Wayback Machine

Archive



Timestamps

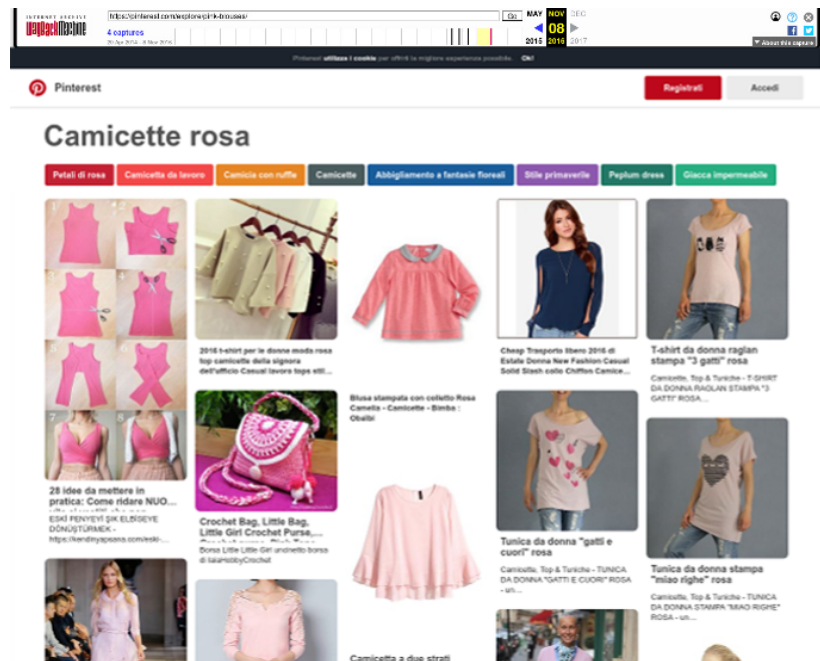


Web Archiving

Example: Capturing via the Wayback Machine

Archive

Timestamps

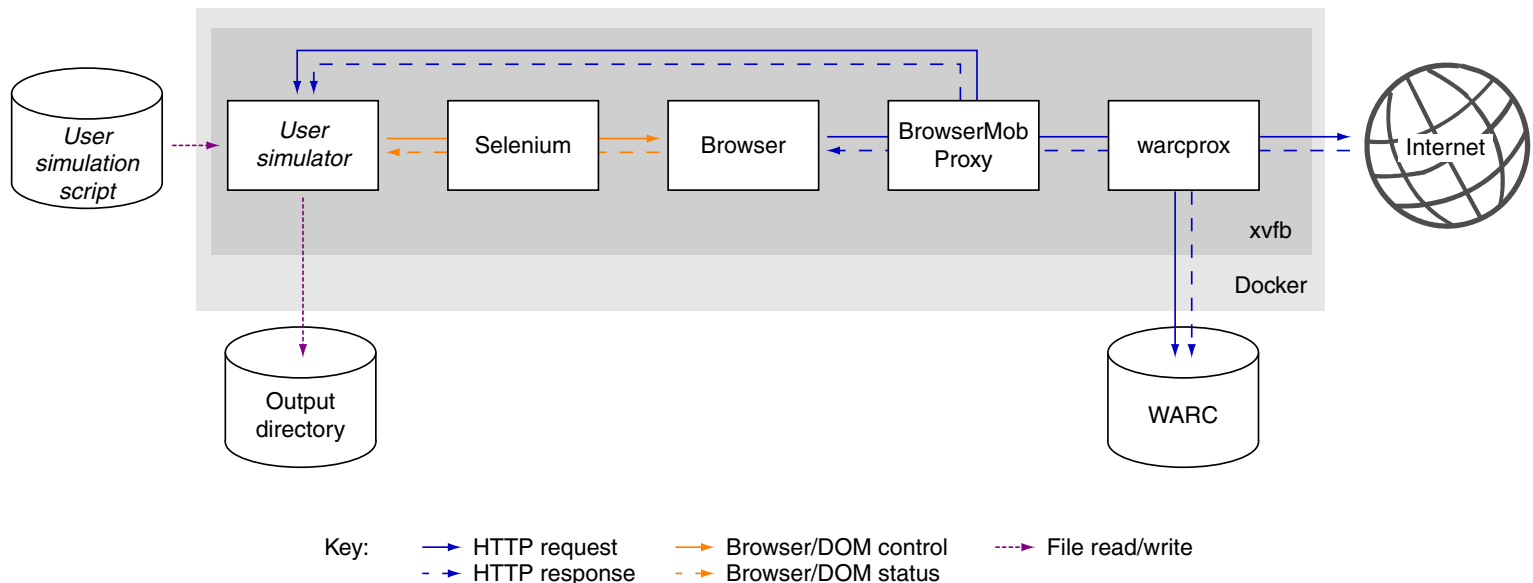


Web Archiving

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:

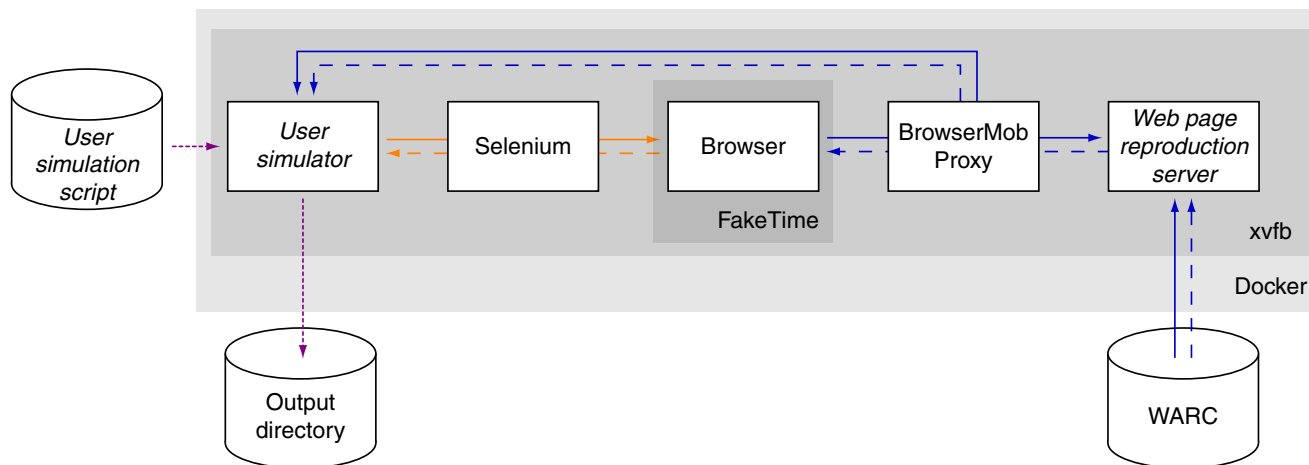


Web Archiving

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:



Key: → HTTP request → Browser/DOM control -.- File read/write
 - - HTTP response - - Browser/DOM status

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:

```
@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

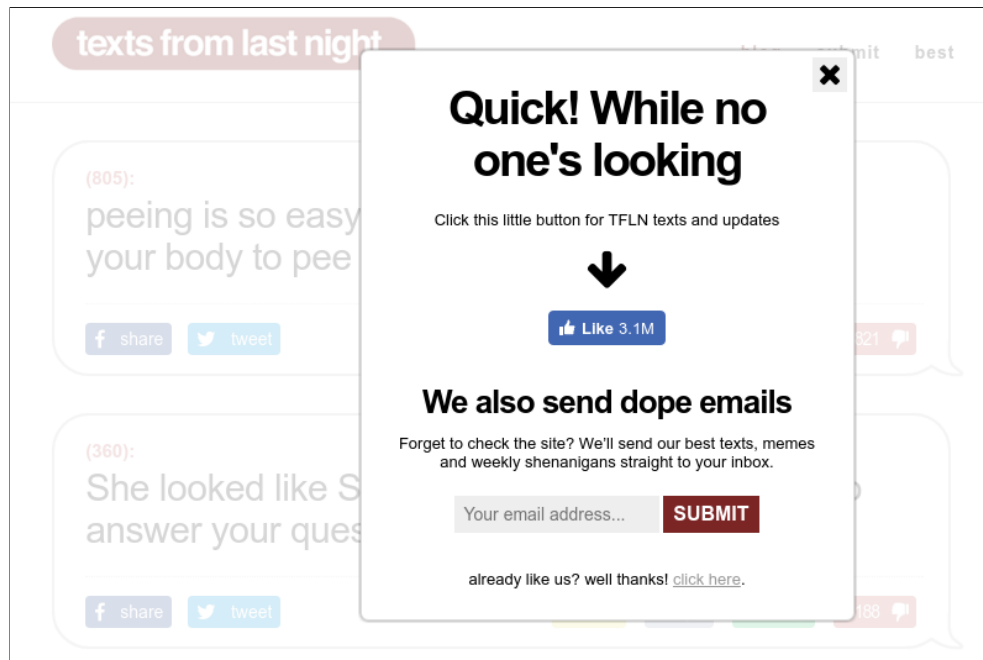
Quality Assurance

- ❑ 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



The screenshot shows a website interface with a modal overlay. The background content includes a header "texts from last night", a text snippet "(805): peeing is so easy your body to pee", and a text snippet "(360): She looked like S answer your ques". The modal overlay has a close button (X) in the top right corner. The modal text reads: "Quick! While no one's looking", "Click this little button for TFLN texts and updates", a downward arrow, a "Like 3.1M" button, "We also send dope emails", "Forget to check the site? We'll send our best texts, memes and weekly shenanigans straight to your inbox.", an email input field with a "SUBMIT" button, and a link "already like us? well thanks! click here."



The screenshot shows a 404 error page. The text reads: "ERROR:404", "Someone's chewed through something they shouldn't have...", a cartoon illustration of a mouse chewing on a computer monitor, "#OutOfCheeseAgain", and a "CONTINUE SHOPPING" button.

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**

The screenshot displays a Business Insider webpage with the following content:

- Header:** Business Insider logo, navigation links (Home, Finance, Tech, Health, Entertainment, Sports, Lifestyle), and a search bar.
- Top Article:** "London's crazy house price bubble is finally losing some air" by Owen Williams. The article discusses the impact of the stamp duty tax on the UK property market, noting a 2.9% decline in house prices in London and a 1.6% decline in the UK overall.
- Market Data:** A section titled "MARKETS INSIDER" showing Dow Jones at 22,334.08 (-1.06%), S&P 500 at 2,160.08 (0.00%), and NASDAQ at 7,311.64 (+8.74%).
- Other Articles:**
 - "Why Non-Financial Data is a CEO Game Changer" by workday.
 - "17 Photos Of Melania Trump That Donald Trump Has Kept Secret".
 - "Prepare to Shape the Future of Finance" by Investopedia.
 - "Videos You May Like" featuring clips from various news segments.
 - "Sponsored Financial Content" section with several promotional articles.
- Footer:** Social media links for Facebook, Twitter, LinkedIn, and YouTube, along with a "Recommended for You" section.

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

The screenshot displays a news website with a dark blue header containing 'FINANCE' and 'BUSINESS INSIDER'. The main article is titled 'London's crazy house price bubble is finally losing some air' by Owen Williams. The article text discusses the decline in house prices in London and the impact of the stamp duty surcharge. A video player is embedded in the article, showing a man in a suit and sunglasses, identified as Henry Ford, spending millions. Below the article, there are several 'Recommended For You' sections with various news snippets, including '20 Trumps Left Out', 'Why Is Donald Trump's Approval Rating So Low?', and 'How Many People Are in the White House?'. The page also features a 'Markets Insider' section with a 'SPONSORED FINANCIAL CONTENT' banner and a 'BI Intelligence Exclusive' section.

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

Search in more than 46 million sentences of German newspaper material

🔍 ?

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 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](#)

Search in 431 Corpus-Based Monolingual Dictionaries for 252 Languages.

Q ?

Corpus: English (eng_news_2016)

English news corpus based on material from 2016 with 156,934,303 sentences.

[i](#) [Statistics](#) [Downloads](#)

[German](#) [English](#) [French](#) [Arabic](#) [Russian](#) [all...](#)

Random words:

[minimum](#) [academic](#) [hour](#) [assist](#) [considered](#)

Corpus

Search corpus...

A Abkhazian Acoli Afrikaans ▾ Akan ▾ Albanian ▾ Amharic ▾ Anaang
Arabic ▾ Aragonese Armenian ▾ Assamese ▾ Assyrian Neo-Aramaic Asturian
 Avaric Aymara Azerbaijani

B Bambara Banjar ▾ Bashkir ▾ Basque Bavarian Belarusian Bengali ▾
 Betawi Bicol ▾ Bishnupriya Bosnian Breton **Bulgarian** Buriat

C Catalan Cebuano ▾ Central Bicol ▾ Central Khmer Chavacano Chechen ▾
 Cherokee **Chinese** ▾ Chuvash ▾ Classical Nahuatl Cornish Corsican Crimean Tatar
Czech

D **Danish** ▾ Dari Dhivehi Dimli ▾ **Dutch**

E Eastern Mari Eastern Yiddish Egyptian Arabic Emiliano-Romagnolo **English** ✓ ▾
 Erzya Esperanto Estonian Ewe Extremaduran

F Faroese Fiji Hindi Fijian **Finnish** **French** ▾ Friulian Fulah

G Gagauz Galician ▾ Gan Chinese Ganda ▾ Georgian ▾ **German** ▾ Gilaki ▾
 Goan Konkani ▾ Guarani ▾ Gujarati

H Haitian ▾ Halh Mongolian Hausa ▾ **Hebrew** Hindi **Hungarian**

I **Icelandic** ▾ Ido Igbo Iloko ▾ Indonesian Interlingua Interlingue
 Iranian Persian Irish **Italian**

J Japanese Javanese ▾

K Kabardian ▾ Kabyle Kalaallisut Kalmyk Kannada ▾ Kara-Kalpak
 Karachay-Balkar Kashubian Kazakh Kinyarwanda Kirghiz ▾ Kituba (Congo)
 Komi Komi-Permyak Konkani ▾ **Korean** Kurdish ▾ Kölsch

L Ladino Lak Lao ▾ Latgalian Latin ▾ Latvian Ligurian Limburgan ▾

- deu_newscrawl-public_2018
- deu_typical-mixed_2018
- deu_newscrawl_2011
- deu_newscrawl-hyphen_2011
- deu_wikipedia_2018
- deu-ch_web_2002
- deu-na_newscrawl_2012
- deu_mixed_2011_dan
- deu-na_web_2019
- deu-ac_web_2017
- deu-na_web_2013

Enter a word



Word: **winter** Number of occurrences: 278,185 Rank: 1,385 Frequency class: 9 📄

See also: [Winter](#), [WINTER](#), [Winter](#)

Part of speech: Verb, Noun

Baseform of: [wintered](#), [wintering](#), [winters](#), [Winters](#), [Winter](#)

Part of: [winter sports](#), [winter wheat](#), [winter day](#), [winter solstice](#), [winter squash](#), [hard winter](#), [winter garden](#), [winter wind](#), [winter crop](#), [winter quarters](#), [winter rye](#), [winter annual](#), [winter tire](#), [winter sleep](#), [winter barley](#), [winter melon](#), [winter beer](#), [winter moth](#), [winter flounder](#), [depth of winter](#), [winter golf](#), [winter fat](#), [winter jasmine](#), [winter aconite](#), [winter wren](#), [more...](#)

Examples: 📄

- The met office are predicting rain, sleet and possibly snow in **winter**. ([www.impartialreporter.com](#), *crawled on 25/11/2016*)
- There are plans to stop there **winter** fuel Allowance. ([www.blackelectorate.com](#), *crawled on 01/11/2016*)
- Reducing high **winter** bills. ([www.eugeneweekly.com](#), *crawled on 02/11/2016*)
- The Salvation Army will continue to collect and distribute coats throughout the **winter**. ([www.tyronepa.com](#), *crawled on 07/11/2016*)
- So how exactly do fashionistas bring **winter** style to the Golden State? ([www.pcccourier.com](#), *crawled on 22/11/2016*)
- Gloves, scarves and warm socks are also important layers to add to any **winter** ensemble. ([www.kentontimes.com](#), *crawled on 12/11/2016*)
- It's the shortest east-west crossing and doesn't get closed in **winter**. ([www.newsandstar.co.uk](#), *crawled on 12/11/2016*)
- Resident Kari Simonson said she thought the need for **winter** parking change was obvious. ([www.snyderdailynews.com](#), *crawled on 22/11/2016*)
- Bald eagles use the area in **winter**. ([www.idahostatesman.com](#), *crawled on 18/11/2016*)
- Paredes got good news on his application last **winter**. ([www.modbee.com](#), *crawled on 11/11/2016*)

+10 +30 +100

Cooccurrences: 📄

Sentence

Left

Right

[cold](#) (75,757), [snow](#) (73,129), [summer](#) (71,314), [spring](#) (64,363), [weather](#) (58,502), [during](#) (51,546), [months](#) (49,240), [fall](#) (40,488), [warm](#) (34,813), [winter sports](#) (31,858), [season](#) (30,453), [the](#) (23,655), [temperatures](#) (23,572), [this](#) (22,016), [storm](#) (20,887), [ice](#) (18,205), [in](#) (17,361), [Winter](#) (16,511), [break](#) (16,139), [mild](#) (15,732), [winter wheat](#) (15,165), [winter day](#) (14,451), [wonderland](#) (13,467), [coats](#) (12,273), [warmer](#) (12,111), [storms](#) (12,110), [winter solstice](#) (11,698), [winter squash](#) (10,755), [conditions](#) (10,742), [for](#) (10,404), [last](#) (10,352), [harsh](#) (9,621), [snowfall](#) (9,404), [and](#) (9,353), [autumn](#) (9,077), [colder](#) (9,062), [solstice](#) (8,966), [ski](#) (8,700), [skiing](#) (8,270), [freezing](#) (8,154), [snowy](#) (7,850), [through](#) (7,763), [birds](#) (7,739), [coat](#) (7,543), [wet](#) (7,399), [long](#) (7,218), [heat](#) (7,048), [sports](#) (6,909), [dry](#) (6,905), [wheat](#) (6,780), [early](#) (6,742), [heating](#) (6,268), [days](#) (6,015), [the weather](#) (5,999), [coldest](#) (5,882), [tires](#) (5,881), [hard winter](#) (5,759), [sun](#) (5,682), [rain](#) (5,574), [Weather](#) (5,275)

Graph: 📄

