

WWW Webarchivierung Wieso? Weshalb? Warum?

Dr. Astrid Schoger (BSB)
Tobias Beinert (BSB)
Katharina Schmid (BSB)

Dr. Simon Donig (Uni Passau)
Markus Eckl (Uni Passau)

nestor virtuell, 22.07.2021



Übersicht

1) Einführung

2) Aufbau von Webarchivsammlungen

~~~ Fragen ~~~

3) Zugang zu Webarchiven: Viewer, Suche, Daten

4) Forschung mit Webarchiven

~~~ Fragen ~~~



Einführung

Sie haben allgemeines Interesse oder Forschungsinteresse an Wahlkampfthemen?

Bei aktuellen Wahlkämpfen informieren Sie sich auf den Websites, auf Social Media Kanälen der Parteien, Fraktionen, Kandidat*innen, in den Medien.

Bei vergangenen Wahlkämpfen, z.B. Landtagswahl Bayern 2013,

- Website der FDP-Fraktion Bayern gleich nach der Wahl verschwunden
- Website der Partei FDP Bayern zwar noch vorhanden, jedoch Inhalte seit 2013 vollständig geändert.

Informationen gibt es nur noch in so genannten Webarchivsammlungen in Gedächtniseinrichtungen

Erste historische Dokumente...

The screenshot shows a Mozilla Firefox browser window displaying the website <http://www.fdp-fraktion-bayern.de/>. The browser's address bar shows the URL and the date 2013-09-23. The website header includes the BSB Bayerische Staatsbibliothek logo and a navigation menu with links for POLITIK, ÜBER UNS, SERVICE, and PRESSE. A search bar is located in the top right corner. A red circle highlights a 'DOMAIN ERWERBEN' notification in the browser's address bar area, which reads: 'Sie können die Domain fdp-fraktion-bayern.de kaufen!'. Below the notification, there are social media icons for Facebook, Twitter, YouTube, and RSS. The main content area features a large image of the Bayerisches Landtag building with a blue overlay containing the text 'FDP FRAKTION IM BAYERISCHEN LANDTAG' and 'FDP im Bayerischen Landtag'. A 'KONTAKTFORMULAR ZUR BILANZBESTELLUNG' is visible on the right side of the page.

Website
<http://www.fdp-fraktion-bayern.de>
gecrawlt am 23.09.2013

DFG-Projekt: Anwendung von DH-Methoden auf Webarchive

Projektpartner: Bayerische Staatsbibliothek
Universität Passau: Lehrstuhl für Digital Humanities und
Jean-Monnet-Lehrstuhl für Europäische Politik

Ziel: Experimentelle Anwendung von DH-Methoden und Werkzeugen auf Webarchive



Gefördert durch



Grundbegriffe der Webarchivierung

- **Webarchivierung:** *Sammlung, Archivierung* und *Bereitstellung* von Websites, d.h. komplexen Angeboten im Word Wide Web, die sich inhaltlich und technisch voneinander abgrenzen lassen
- **Websites** bestehen aus mehreren in Beziehung zueinander stehenden Einzeldateien, sogenannten **Webpages**, identifiziert durch URLs und sind in Webbrowsern darstell- bzw. abspielbar
- **Harvesting/Crawling:** Verfolgen der (internen Links) und lokale Speicherung der einzelnen Dateien sowie Überführung in eine Struktur bzw. ein Format, die den Gesamtkontext der Website erhält
- Ziel: *Zukünftige Nutzung* durch Wissenschaft, Forschung und breite Öffentlichkeit

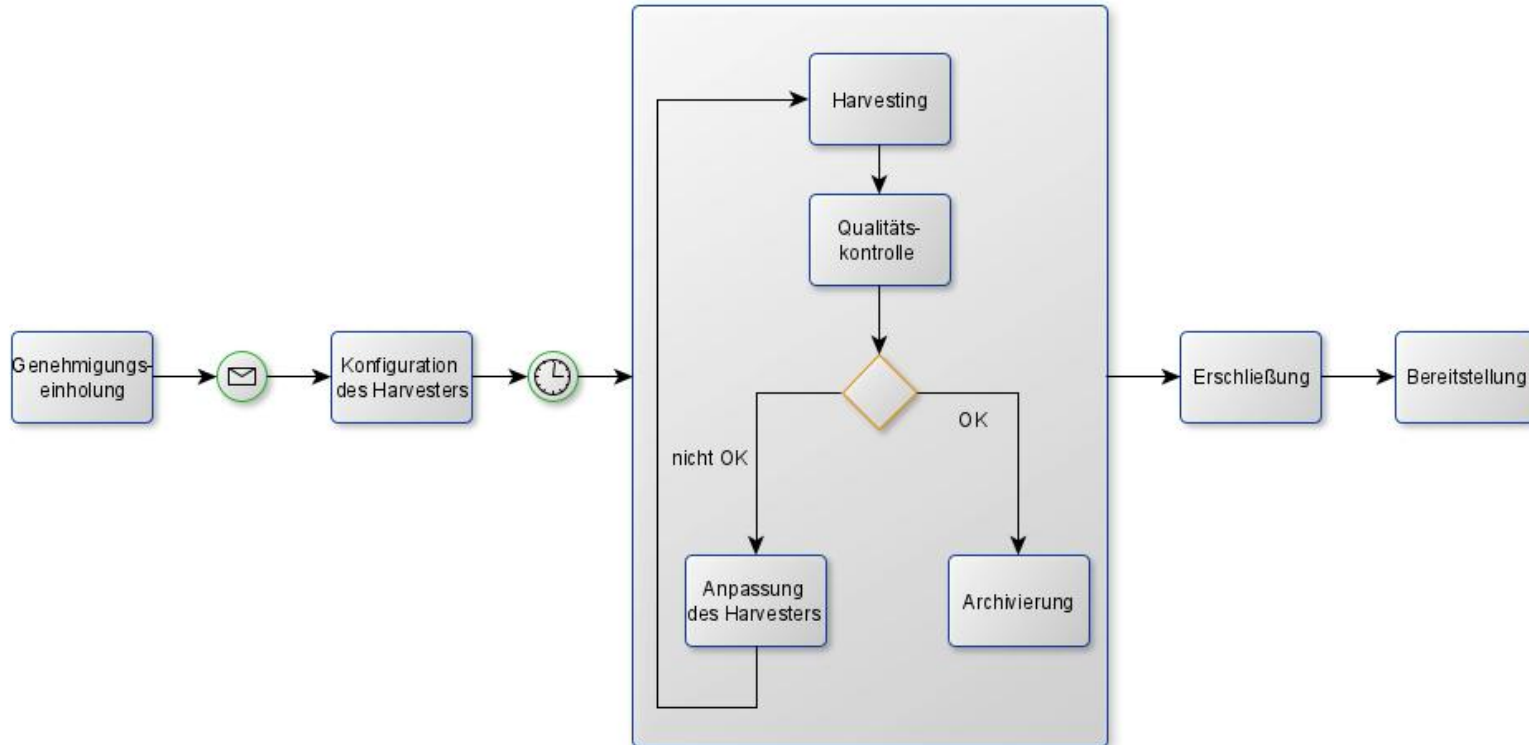
Aufbau von Webarchivsammlungen

- **Selektive Webarchivierung:** Archivierung von ausgewählten Websites mit Qualitätskontrolle (z.B. Websites von Ministerien, Behörden, Wissenschaftler*innen)
- **Event Harvesting:** Anlassbezogene, temporäre Archivierung von Websites zu gesellschaftlichen Ereignissen (z.B. Corona-Pandemie, Wahlen, Sportereignissen)
- **Domain-Crawl:** Flächendeckendes Harvesting aller Websites unter einer bestimmten Domain (z.B. .de)

Rechtliche Rahmenbedingungen

- Für Erstellung, Archivierung und Langzeiterhaltung sowie Bereitstellung wird eine rechtliche Grundlage benötigt → Vervielfältigung laut **Urheberrecht**
- Anpassung der **Pflichtgesetze** für DNB sowie der meisten Bundesländer
- Anpassung des **Gesetzes über die DNB 2018**
- Standard: Zugänglichmachung im Lesesaal
- Andere Institutionen: **Genehmigungseinholung** für Harvesting, Langzeitarchivierung und Zugänglichmachung
- **UrhG § 60d: Text und Data Mining**
→ Vervielfältigungen für Text und Data Mining für wissenschaftliche Forschung

Workflow selektive Webarchivierung



Qualität und Herausforderungen

- Vollständige Archivierung einer Website nur selten möglich
- **Qualitätskriterien zur Beurteilung einer geharvesteten Website**
 - Vollständigkeit/Konsistenz (z.B. alle Unterseiten vorhanden, keine externen Inhalte)
 - Vorhandensein der Inhalte (z.B. alle Images, PDFs vorhanden)
 - Erhalt der Funktionalität (z.B. Navigation, Browsing, Kalenderfunktionen)
 - Erhalt des Look and Feel (z.B. Schriften, Stylesheets)
- **Technische Limits:** Flash, JavaScript, Videostreaming, Datenbanken, dynamischer Content
- Wachsende **Komplexität der Daten** und Einbettung in **geschlossene Systeme**

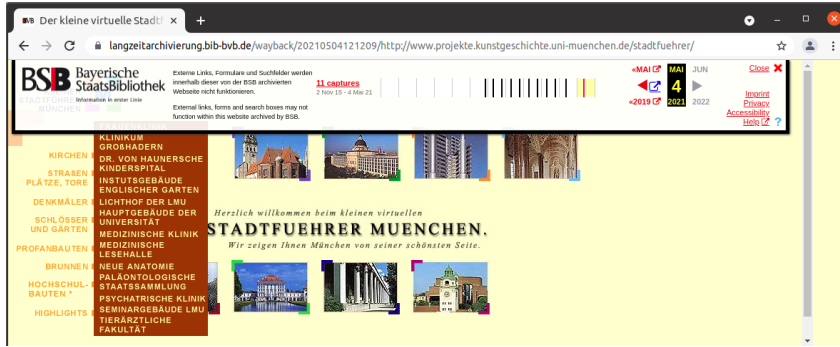
Dateiformat WARC

- Standardisiertes Containerformat für die Webarchivierung, in dem die einzelnen Dokumente der Original-Website gespeichert werden
- Metadaten zum Crawlprozess
- Große Verbreitung in der Webarchivierung
- Grundlage für die strukturierte Darstellung in der WaybackMachine

Technische Infrastruktur für die Webarchivierung

- **Software für das Harvesting einer Website**
 - Heritrix, Webrecorder
- **Management des Workflows**
 - Web Curator Tool, Netarchive Suite
- **Viewer**
 - OpenWayback, PyWb
- **Serviceangebote/Dienstleister**
 - Archive-It, MirrorWeb

Zugang zu Webarchiven: Viewer OpenWayback



Website

<http://www.projekte.kunstgeschichte.uni-muenchen.de/stadtfuehrer>

gecrawlt am 04.05.2021

- Im Katalog der BSB können Forschende bereits jetzt gezielt nach archivierten Websites suchen (Facette „Medienart“) und sich einzelne Zeitschnitte im Browser über den Viewer OpenWayback anzeigen lassen.
- Dank der Genehmigung der Rechteinhaber*innen sind die archivierten Websites öffentlich zugänglich.

Zugang zu Webarchiven: Viewer OpenWayback

✓ URL-basierter Zugang

✗ Suche nach bestimmten Inhalten

✓ intellektuelle Analyse einzelner Websites in der Browseransicht

✗ Analyse größerer Datenmengen mit (teil-)automatisierten Verfahren

Zugang zu Webarchiven: Volltextsuche

Exploration einer Sammlung von archivierten Websites

Suche im Inhalt

Filterung anhand von Metadaten wie Domain, Datentyp, ...

Zugang zu Webarchiven: Volltextsuche

The screenshot shows a search interface with a search bar containing the word 'inklusion' and a red 'Search' button. Below the search bar, there are three filter categories: DOMAIN, ORT, and PERSON. Each category has a list of items with checkboxes and counts. The search results are displayed in two cards. The first card is titled 'Mehr Bewusstsein für eine inklusive Gesellschaft | Katarina Barley' and shows a URL and four dates: 7.6.2019, 14.6.2019, 21.6.2019, and 28.6.2019. The second card is titled 'Inklusives Wahlrecht: SPD setzt sich durch | Katarina Barley' and shows a URL and the same four dates.

inklusion Search

DOMAIN

- skakeller.de 70,810
- patrick-brayer.de 21,751
- peter-liese.de 11,426
- martin-schirdewan.eu 9,220
- fw-europa.com 6,800
- [+ More](#)

ORT

- europa 100,550
- europawahl 87,348
- europäischen 81,272
- brüssel 77,266
- rue 75,355
- [+ More](#)

PERSON

- robert 70,072
- schuman 69,697
- ska 69,463
- ska,keller 69,269
- benjamin-bremer 69,024
- [+ More](#)

Mehr Bewusstsein für eine inklusive Gesellschaft | Katarina Barley

"URL": <https://katarina-barley.de/mehr-bewusstsein-fuer-eine-inklusive-gesellschaft/embed/>

"Gecrawlt am":

- 7.6.2019
- 14.6.2019
- 21.6.2019
- 28.6.2019

Inklusives Wahlrecht: SPD setzt sich durch | Katarina Barley

"URL": <https://katarina-barley.de/inklusives-wahlrecht-spd-setzt-sich-durch/embed/>

"Gecrawlt am":

- 7.6.2019
- 14.6.2019
- 21.6.2019
- 28.6.2019

Dargestellt ist die prototypische Volltextsuche basierend auf Daten des Event Crawls zur Europawahl 2019.

Suchergebnisse lassen sich zusätzlich anhand der Facetten „Domain“, „Ort“ und „Person“ filtern. Eigennamen von Personen und Orten wurden mittels Named Entity Recognition (NER) im Rahmen der Indexierung automatisch ausgewertet.

Die Suche dient der Annäherung an Fragen wie:

- Wie positioniert sich eine Partei auf ihrer Website zu einem bestimmten Thema?
- Welche Personen kommen in Zusammenhang mit einem Thema vor?

Zugang zu Webarchiven: Daten

Für computergestützte Analysen

Ausgangsformat WARC: Containerformat, außerhalb der Webarchivierung wenig bekannt

Filterung und Extraktion von Daten Voraussetzung für weitere Analysen

→ abgeleitete Datensets (Derivate)

Derivate mit dem Archives Unleashed Toolkit

- Für die Extraktion von Daten aus WARC-Dateien steht bereits entsprechende quelloffene Software zur Verfügung, wie z. Bsp. das Archives Unleashed Toolkit.
- Anhand der archivierten Website <https://www.gruene.de> soll das Vorgehen bei der Datenextraktion veranschaulicht werden (siehe nächste Folie):

Mit dem Archives Unleashed Toolkit wurden aus den archivierten HTML-Seiten Textinhalte extrahiert und in einer CSV-Datei gespeichert. Das Textderivat enthält die Textinhalte der einzelnen Webseiten ohne HTML-Markup sowie ergänzende Metadaten wie die ursprüngliche URL und das Crawldatum.

- Analog können mit dem Archives Unleashed Toolkit auch andere Arten von Daten wie ausgehende Links oder Bilder aus den WARC-Dateien extrahiert werden.

Derivate mit dem Archives Unleashed Toolkit



Website <https://www.gruene.de>
gecrawlt am 22.06.2019

| Crawl-datum | URL | Textinhalt |
|-------------|---|--|
| 20190622 | https://www.gruene.de/europawahl/ | ... Am 26. Mai haben wir es in der Hand: Wir können den Zusammenhalt ... |
| 20190622 | https://www.gruene.de/partei/ | ... Seit unserer Gründung kämpfen wir für die Natur und eine Welt, in der alle ... |
| 20190622 | https://www.gruene.de/beschluesse-und-programme/ | ... Grüner wird's nicht! Hier gibt es wichtige Dokumente von BÜNDNIS 90/DIE ... |
| ... | | |

Forschung mit Webarchiven

Exploration der Derivate mit gängigen Analysewerkzeugen

Bsp. Textanalyse: Voyant Tools

Bsp. Netzwerkanalyse: Gephi

Datenbasis sind Text- und Linkderivate, die mit dem Archives Unleashed Toolkit aus der archivierten Website <https://www.gruene.de> erzeugt wurden.

Analyse eines Textderivats mit Voyant Tools

Voyant Tools ermöglicht verschiedene Perspektiven auf die Texte einer Website: den Blick aus der „Vogelperspektive“ auf den Korpus in seiner Gesamtheit und die Detailansicht einzelner Dokumente.

Zur Veranschaulichung dienen die beiden Fenster „Summary“ und „Reader“ (siehe nächste Folie):

- „Summary“ zeigt Merkmale des gesamten Korpus wie die häufigsten Begriffe oder die Begriffe, die verglichen mit dem Rest des Korpus charakteristisch für ein Dokument sind. Anhand letzterer lässt sich bereits eine Hypothese über den Inhalt der Dokumente aufstellen.
- Mit einem Klick kann im gegenüberliegenden „Reader“ das gewünschte Dokument angezeigt werden, um die Hypothesen zu überprüfen.

Analyse eines Textderivats mit Voyant Tools

The screenshot shows the Voyant Tools interface with a blue header. The navigation bar includes 'Terms', 'Links', 'Summary', 'Reader', and 'TermsBerry'. The main content area is divided into two panels. The left panel displays statistical data and a list of distinctive words. The right panel shows a preview of the text being analyzed, with a search bar at the bottom.

Average Words Per Sentence:

- Highest: (124.0); (72.7); (61.0); (34.3); (28.9)
- Lowest: (3.0); (5.7); (6.2); (6.7); (7.3)

Most frequent words in the corpus: **menschen** (540); **grünen** (465); **mehr** (415); **europa** (298); **grüne** (278)

Distinctive words (compared to the rest of the corpus):

1. : **garantiesicherung** (18), **hartz** (16), **iv** (10), **sanktionen** (7), **jobcenter** (7).
2. : **neue** (31), **fragen** (19), **mensch** (15), **antworten** (8), **digitalisierung** (8).
3. : **neue** (32), **fragen** (20), **mensch** (15), **antworten** (8), **digitalisierung** (8).
4. : **polizei** (23), **sicherheitsbehörden** (10), **innenpolitik** (8), **kriminallität** (7), **innen** (17).
5. : **investitionen** (25), **schuldenbremse** (16), **mrp** (15), **infrastruktur** (19), **euro** (21).
6. : **diversität** (9), **gerechtigkeit** (17), **entwicklung** (10), **gewalt** (8), **ebene** (7).
7. : **entwicklung** (16), **strukturpolitik** (6), **globale** (9), **nachhaltigen** (8), **agenda** (6).

items:

https://www.gruene.de/artikel/anreiz-statt-sanktionen-bedarfsgerecht-und-bedingungslos/

Wie wir die Beschlusslagen der Partei umsetzen und ein Garantiesystem aufbauen. Ein Debattenbeitrag von Robert Habeck zum Grundsatzprogramm. Unser Sozialstaat beruht auf dem Vertrauen, dass er uns Sicherheit garantiert. Dieses Garantieverprechen ist

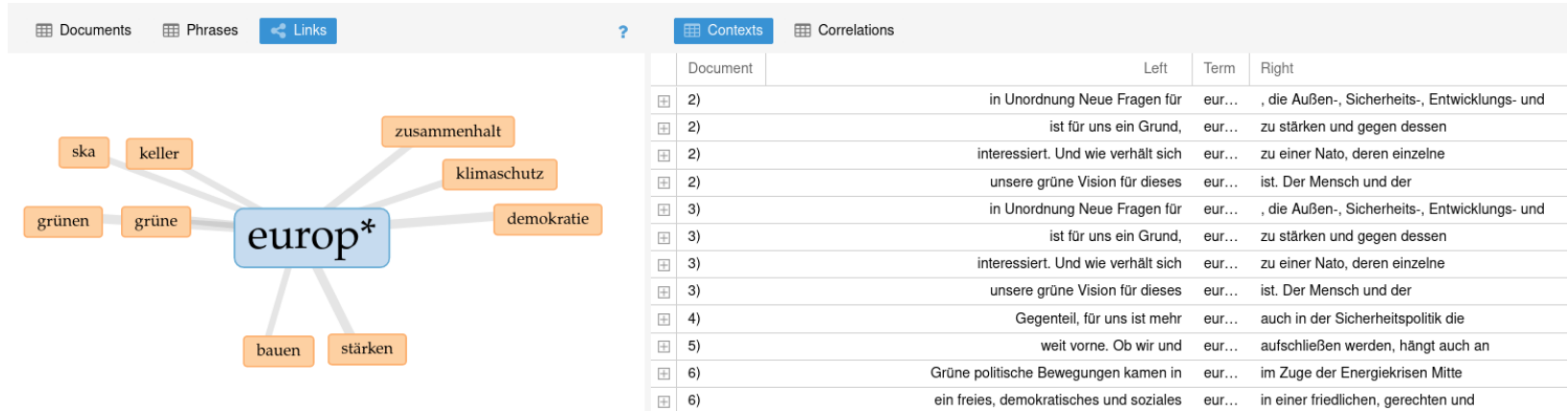
g... wir einen extremen und hoch...
...haben. Es ist für die M...
...neue...
...ist

Analyse eines Textderivats mit Voyant Tools

Auch bei der Auswertung von Begriffskontexten können sich die beiden Perspektiven abwechseln und ergänzen (siehe nächste Folie):

- Das „Links“-Fenster bietet den Blick auf den Gesamtkorpus als Kollokationsgraph. Die Begriffe, die am häufigsten in Zusammenhang mit dem zentralen Begriff „europ*“ vorkommen, sind hier als Netzwerk visualisiert. Sie können Hinweise darauf geben, wie ein Konzept auf der Website geframed wird.
- Konkrete Vorkommen des Begriffs sind rechts im Fenster „Contexts“ sichtbar und können dazu dienen, die Framing-Thesen mit Beispielen zu untermauern.

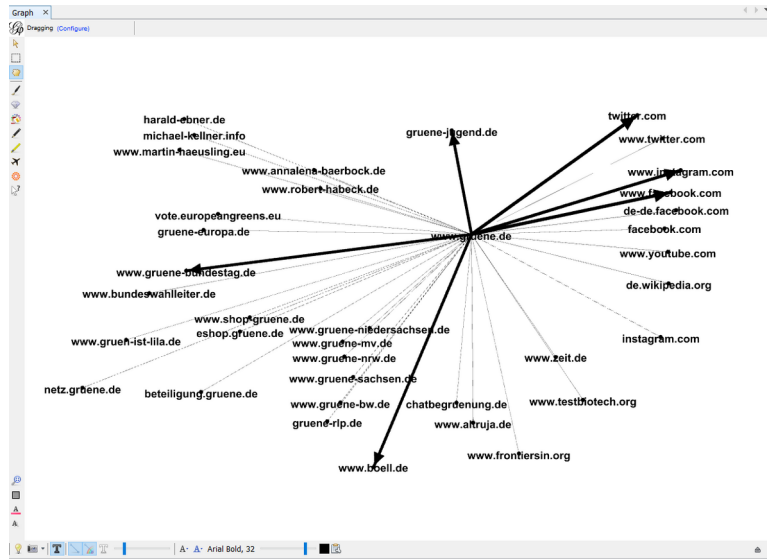
Analyse eines Textderivats mit Voyant Tools



The screenshot displays the Voyant Tools interface. On the left, a word cloud for the term 'europ*' is shown, with related terms in orange boxes: 'ska', 'keller', 'zusammenhalt', 'klimaschutz', 'demokratie', 'bauen', 'stärken', 'grünen', and 'grüne'. The central term 'europ*' is in a blue box. On the right, the 'Contexts' tab is active, showing a table of document contexts.

| Document | Left | Term | Right |
|----------|---|--------|---|
| 2) | in Unordnung Neue Fragen für | eur... | , die Außen-, Sicherheits-, Entwicklungs- und |
| 2) | ist für uns ein Grund, | eur... | zu stärken und gegen dessen |
| 2) | interessiert. Und wie verhält sich | eur... | zu einer Nato, deren einzelne |
| 2) | unsere grüne Vision für dieses | eur... | ist. Der Mensch und der |
| 3) | in Unordnung Neue Fragen für | eur... | , die Außen-, Sicherheits-, Entwicklungs- und |
| 3) | ist für uns ein Grund, | eur... | zu stärken und gegen dessen |
| 3) | interessiert. Und wie verhält sich | eur... | zu einer Nato, deren einzelne |
| 3) | unsere grüne Vision für dieses | eur... | ist. Der Mensch und der |
| 4) | Gegenteil, für uns ist mehr | eur... | auch in der Sicherheitspolitik die |
| 5) | weit vorne. Ob wir und | eur... | aufschließen werden, hängt auch an |
| 6) | Grüne politische Bewegungen kamen in | eur... | im Zuge der Energiekrisen Mitte |
| 6) | ein freies, demokratisches und soziales | eur... | in einer friedlichen, gerechten und |

Analyse eines Linkderivats mit Gephi



- Die ausgehenden Links der Website sind als gerichteter Graph visualisiert (Knoten = Domain, Kante = Link).
- Die Darstellung gibt z.B. Hinweise darauf, welche Sozialen Medien die Partei nutzt oder auf welche Nachrichtenseiten sie sich bezieht. Auch die Organisationsstruktur der Partei spiegelt sich teilweise wieder in den Links auf die Websites der Landesverbände.
- So können möglicherweise weitere relevante Websites identifiziert werden, die archiviert und analysiert werden sollen.

Fallbeispiel Europawahlkampf 2019

Exemplarische Forschungsfrage (n)

- Welche Themen werden in deutschen Medienwebseiten in der “heißen Phase” des Europawahlkampfes von 2019 diskutiert?
- Und welche Themenkonjunkturen sind beobachtbar?

Materialfülle als erkenntnistheoretische Herausforderung

“Distant reading’, I have once called this type of approach; where distance is however not an obstacle, but a specific form of knowledge: fewer elements, hence a sharper sense of their overall interconnection. Shapes, relations, structures. Forms. Models.”

Moretti, Franco. *Graphs, Maps, Trees: Abstract Models for a Literary History*. London: Verso, 2005.

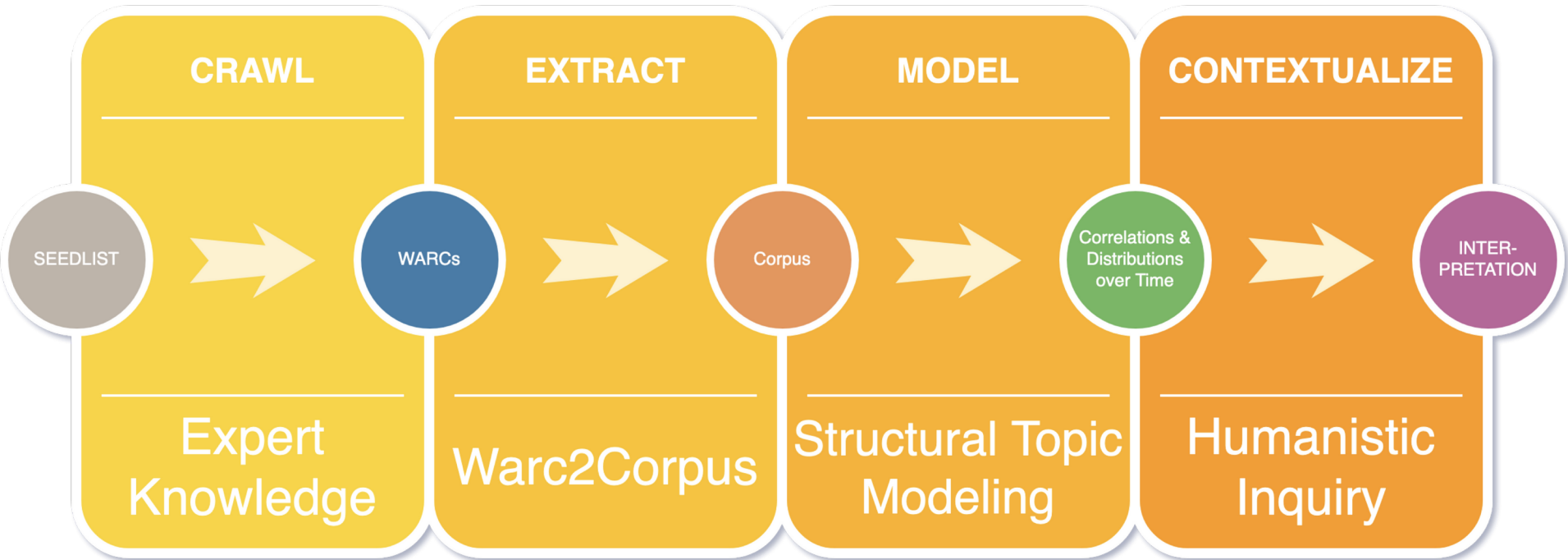
“Distant reading: where distance, let me repeat it, is a condition of knowledge: it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes-or genres and systems.”

Moretti, Franco. *Distant Reading*. New York / London: Verso, 2013.

Zeit in Webarchiven

- Zeit ist essentiell, um komplexe sozial- und kulturwissenschaftliche Phänomene wie Diskurse oder Begriffsgeschichte abzubilden.
- Im Prozess des “close reading” nehmen wir den zeitlichen Kontext einer Webseite explizit oder implizit zur Kenntnis.
- In Webarchiven können wir die Zeit normalerweise nicht explizit fassen, da es keinen standardisierten und weithin akzeptierten Weg gibt, um die Entstehung und Veränderungen einer Webseite zu erfassen.
- WARC Metadaten (z.B. Crawl-Datum) können allenfalls als Annäherung herangezogen werden, aber sind nicht verlässlich.

Forschungsprozess



Der Forschungsprozess lässt sich wie folgt beschreiben:

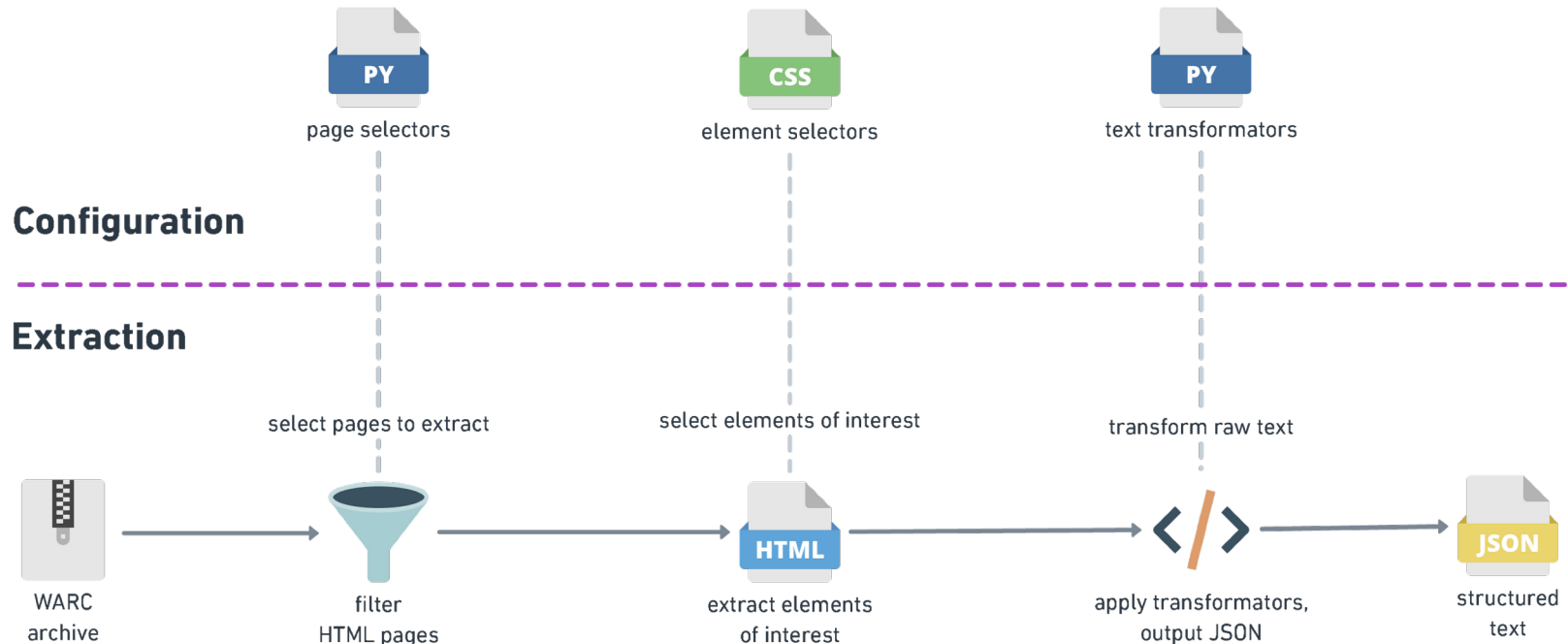
- Auf der Basis von Expertenwissen wird eine Seedlist erstellt
- Aus den aus dem Crawl hervorgehenden WARC Dateien müssen anschließend relevante Inhalte extrahiert werden. Dazu setzen wir die selbst entwickelte Software Warc2Corpus ein, die Textinformationen und Metadaten wie das Erstellungsdatum granular extrahiert.
- Auf diesem Korpus berechnen wir ein Computer-linguistisches Modell, ein Topic Model. Damit können wir einmal unser Material weiter eingrenzen (Zeitraum, Themenrelevanz).
- Diese Ergebnisse können wir dann interpretieren und durch qualitative Untersuchung empirisch absichern.

Funktionsweise von Warc2Corpus

Filter HTML pages to extract by path and/or by inspecting content, e.g. `"/news/\d+"` containing an `<article>` element.

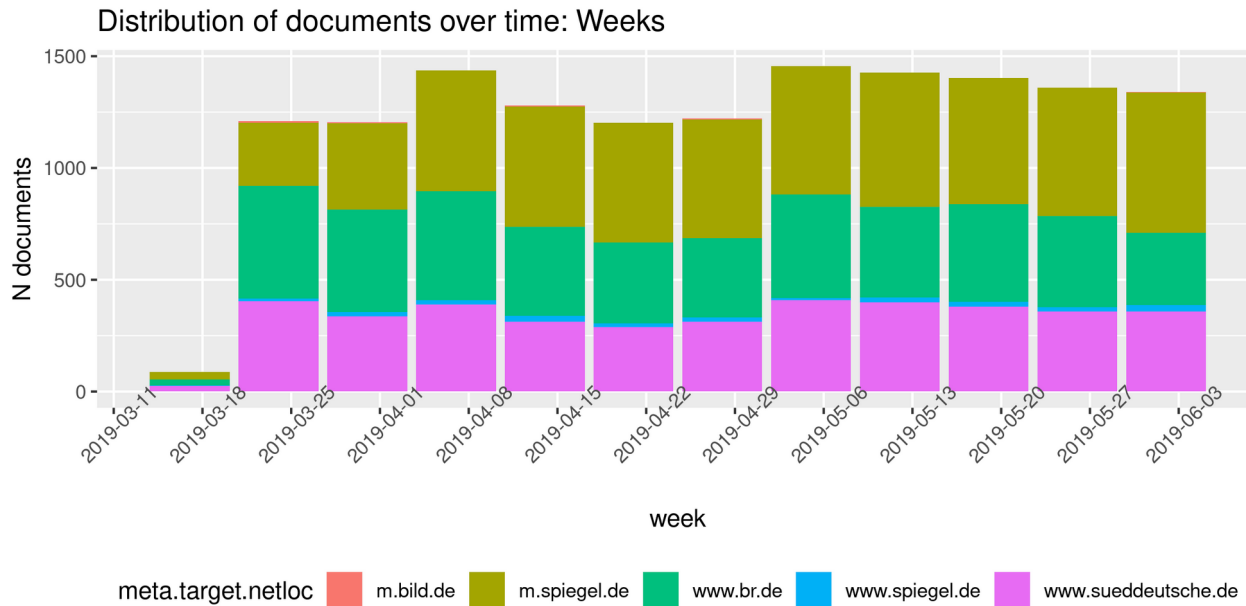
Apply CSS selectors on HTML text to extract only structured information of interest, e.g. title, body, date-of-publication.

Apply lambdas on the elements extracted to transform data, e.g. convert the string `"14. Februar 2018"` into a Python date object.



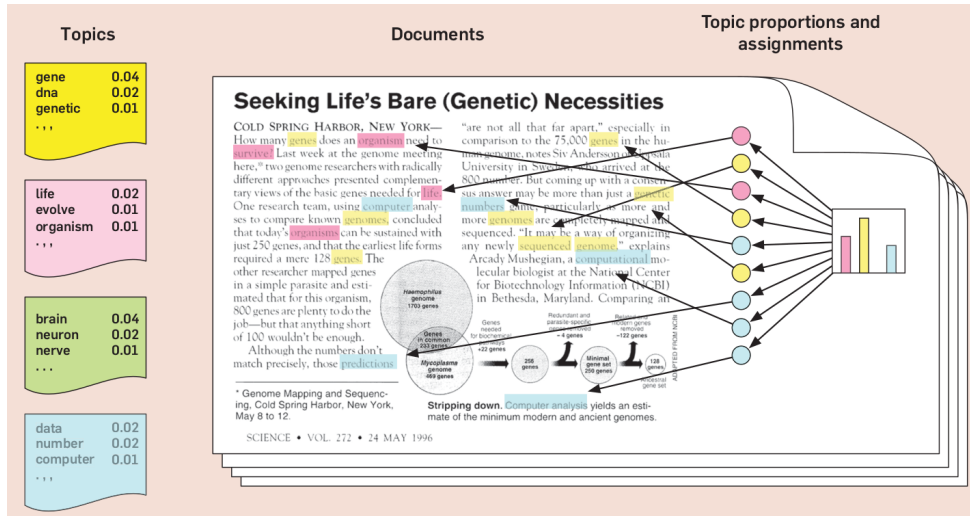
Funktionsweise von Warc2Corpus

- Warc2Corpus erfüllt vier wichtige Funktionen: Die Identifikation relevanter Inhalte; die Extraktion dieser Inhalte, die Ablage der extrahierten Inhalte in strukturierter Form und die Standardisierung von bestimmten Datenobjekten (z.B. des Datums).
- In einem ersten Schritt werden über eine Konfigurationsdatei die wesentlichen zu extrahierenden Inhalte benannt und auf Grundlage der URL einzelne HTML-pages extrahiert.
- Im zweiten Schritt werden für jede Website die wesentlichen Elemente lokalisiert (Wo ist der Titel, Text, Veröffentlichungsdatum?) und ebenfalls in eine Konfigurationsdatei gespeichert. Die Anpassung dieser Konfigurationen für den Extraktor ist derzeit noch zeitaufwändig und soll zukünftig möglichst stark automatisiert werden.
- Im dritten Schritt können dann mit den Extraktoren im großen Stil Informationen extrahiert werden, die strukturiert in eine JSON-Datei gespeichert werden. Bestimmte Datenobjekte werden dabei möglichst standardisiert.



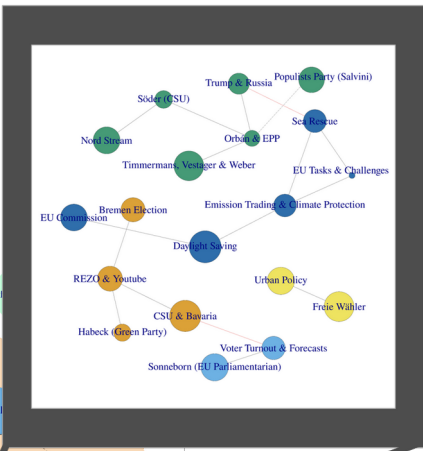
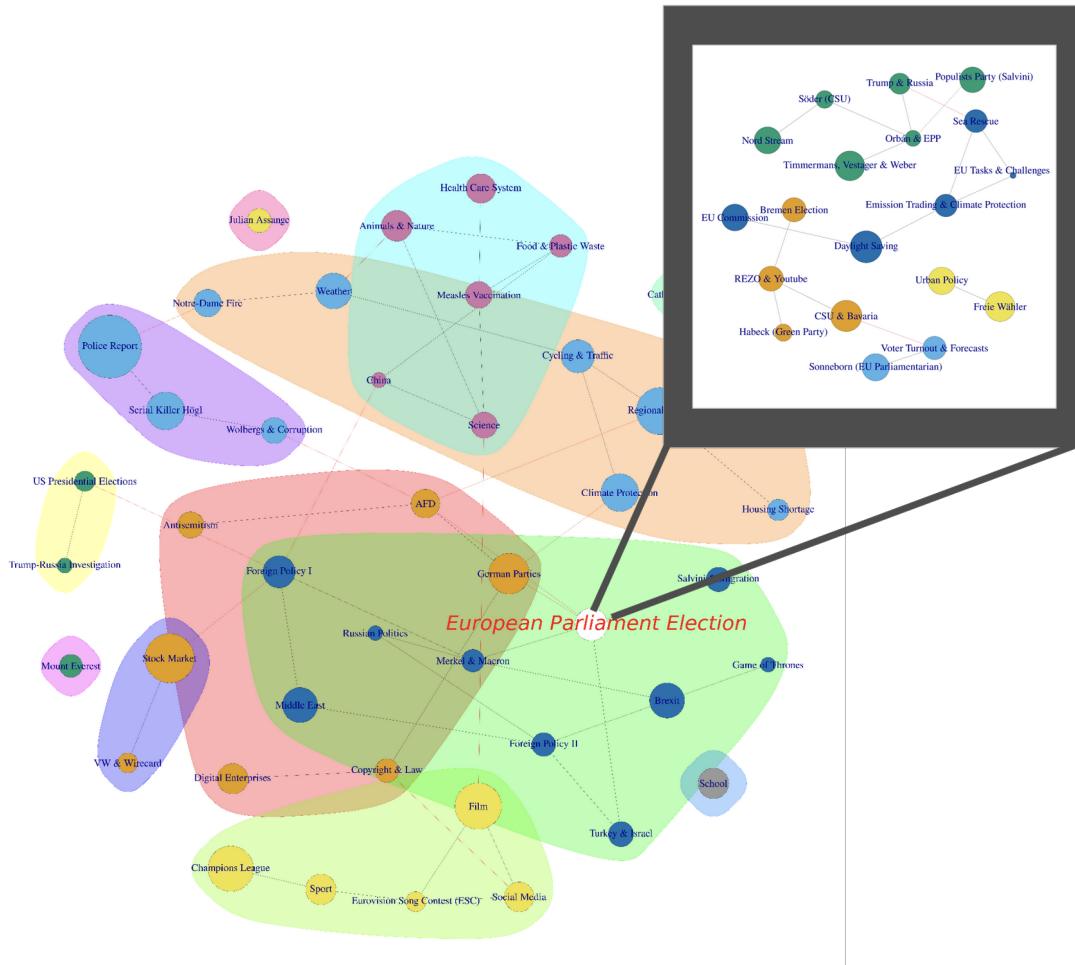
- Das Ergebnis dieses Prozesses erlaubt es z.B. ausgeben zu lassen wie sich die Menge der extrahierten Seiten, für die es ein Datum gibt, über den Untersuchungszeitraum verteilt. Für die durchgeführte Analyse sollten nur Dokumente verwendet werden, die im Zeitraum des zentralen Wahlkampfes der Europawahl veröffentlicht wurden.
- Unser Beobachtungszeitraum erstreckt sich daher von Ende März bis Juni 2019.
- Die Grafik zeigt die wöchentliche Anzahl der extrahierten Dokumente differenziert nach Medium, in dem sie erschienen sind. Insgesamt wurden für diesen Zeitraum 14.627 Dokumente extrahiert.

Latent Dirichlet Allocation (LDA)



Blei 2012

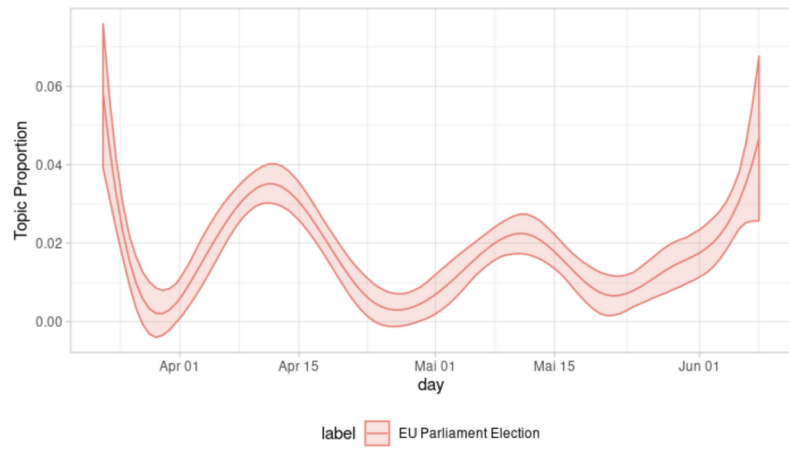
- LDA ist eine Methode des Topic Modelings.
- Eine Grundannahme der LDA ist, dass ein Dokument nur eine bestimmte Anzahl an Themen (Topics) enthält. Die Themen wiederum werden bestimmt durch die im Dokument enthaltenen Wörter.
- Es besteht somit eine Wahrscheinlichkeitsverteilung, dass ein Dokument aus bestimmten Themen besteht. Eine zweite Wahrscheinlichkeitsverteilung verweist auf die Wahrscheinlichkeit, dass ein Thema aus bestimmten Wörtern besteht.
- In dieser Studie verwenden wir das R Package Structural Topic Modeling (STM), um Drittvariable (z.B. Zeit) in das Modell zu integrieren. Zudem bietet STM die Möglichkeit, die Topics in Korrelation zueinander zu setzen. Hierfür wird ein Korrelationsgraph erstellt (Knoten = Labels der Topics; Kanten = Wahrscheinlichkeit, dass zwei Topics gemeinsam in einem Dokument erscheinen; Farbe = Clustering)



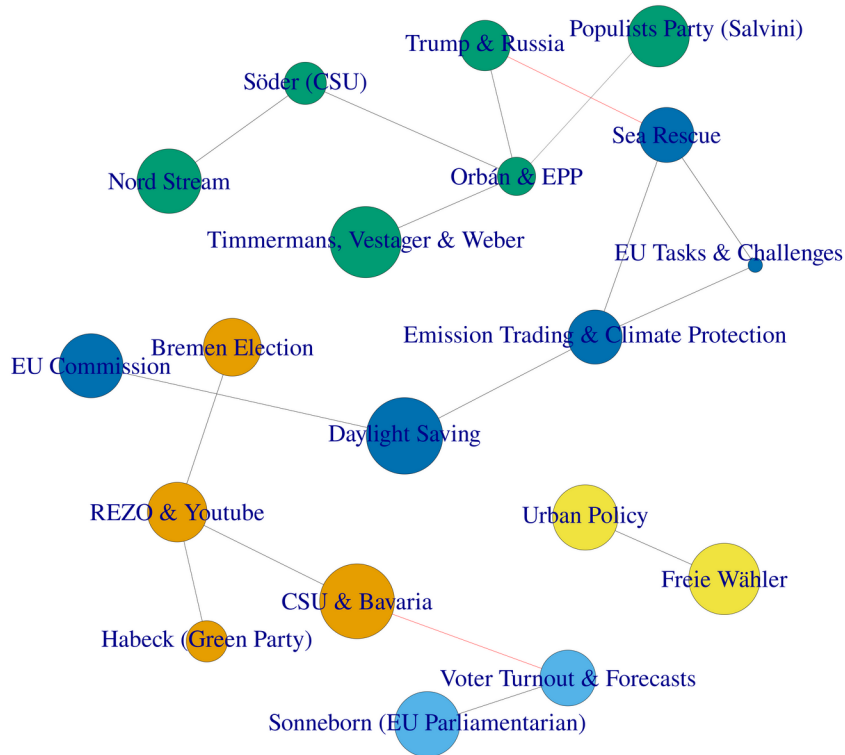
Topic:
European Parliament Election

Wordlist:
europawahl, manfred_weber, evp, afd, partei, europaparlament, csu, timmermans, spitzenkandidaten

Diffusion of Topics



Diskursfelder



- Der Cluster mit den grünen Knoten besteht beispielsweise aus Dokumenten, die Themen zur Außenpolitik oder zu institutionellen Dimensionen der Politik enthalten.
- Charakteristische Themen sind z.B. Trump & Russland, Nord Stream Pipeline, populistische Bewegungen oder die Rolle von Viktor Orbáns Fidesz in der Europäischen Volkspartei.
- Das zweite Cluster enthält Dokumente, die Themen mit Bezug zu ökologischen, sozialen oder komplexen ethischen Fragen wie Asyl und Seenotrettung, Umweltaspekte sowie Bürgerbeteiligung wie im Fall der Sommerzeit darstellen.
- Auch bemerkenswerte Ereignisse, die die nationale oder staatliche Ebene mit der europäischen Ebene verknüpfen, können ein eigenes Thema darstellen.
- Der heftige Angriff eines Youtubers mit dem Künstlernamen REZO auf die konservative Partei, der den Wahlkampf in Deutschland über mehrere Wochen dominierte, steht im Zentrum eines eigenen Clusters, das auch die Bundesländer Bremen und Bayern verbindet, die beide 2019 Wahlen hatten.

Zusammenfassung

1. Aufbau von Webarchivsammlungen in Gedächtnisorganisationen
2. Nutzungsszenarien von Webarchiven: close, blended, distant reading

Sind Webarchive als Forschungsdaten nutzbar?

- + standardisierte Verfahren/Archivformate in der Webarchivierung: z.B. WARC
- + es gibt open source Tools, die out of the box genutzt ggf. weiterentwickelt werden können
- + etablierte DH-Methoden anwendbar nach Bildung geeigneter Derivate: Text, Link, Bild, ...

- Bestände insgesamt noch heterogen und lückenhaft
- schwer auffindbar, Zugang häufig stark eingeschränkt
- Nachnutzung von spezifisch zusammengestellten Forschungskorpora nicht erlaubt

Referenzen

Nestor Thema 15 – Das Dateiformat WARC für die Webarchivierung:

<http://nbn-resolving.de/urn:nbn:de:0008-2021042614>

Software für die Webarchivierung im Überblick, zusammengestellt vom

IIPC: <https://github.com/iipc/awesome-web-archiving#tools--software>

Archives Unleashed Toolkit: <https://archivesunleashed.org/aut/>

Voyant Tools: <http://voyant-tools.org/>

Gephi: <https://gephi.org/>

warc2corpus: <https://github.com/sebastian/warc2corpus> (im Aufbau)

Blei, David M. 2012. „Probabilistic topic models : surveying a suite of algorithms that offer a solution to managing large document archives.“
Communications of the ACM 4 (55): 77–84.

<https://doi.org/10.1145/2133806.2133826>.

stm – an R package for the Structural Topic Model:

<https://www.structuraltopicmodel.com/>



