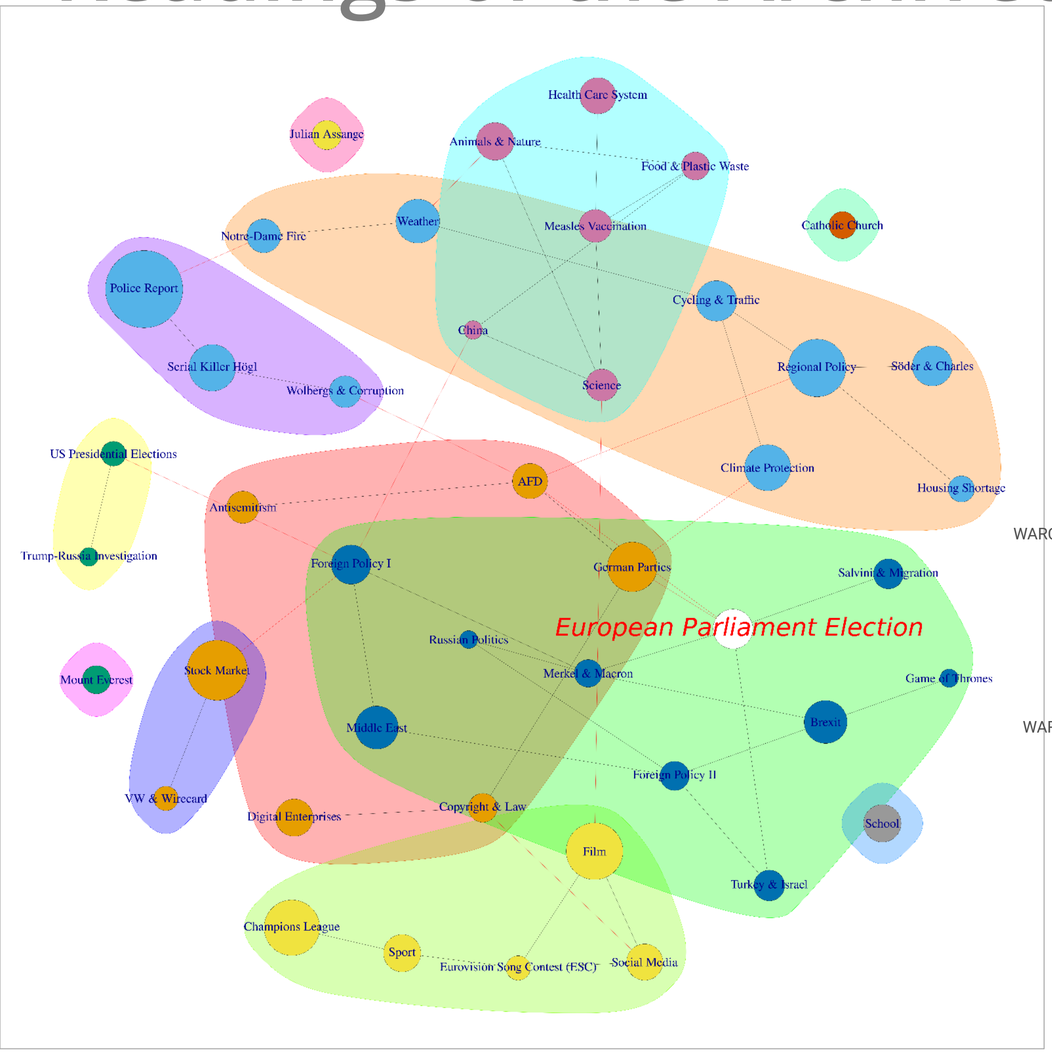
Blind Spots & Silent Pages - Distant Readings of the Archived Web



WARC/1.1

WARC-Filename: digitales-archiv.wacz#/archive/data.warc

WARC-Date: 2021-03-09T09:52:11.810Z

WARC-Type: warcinfo

WARC-Record-ID: <urn:uuid:119a4404-ab67-4abd-9276-69517b1fb882>

Content-Type: application/warc-fields

Content-Length: 125

software: Webrecorder ArchiveWeb.page 0.5.9 (via warcio.js 1.4.1)

format: WARC File Format 1.1

isPartOf: Digitales Archiv

WARC/1.1

WARC-Page-ID: ao8jp3bgxb7uwbuidqnb7p

WARC-Payload-Digest: sha-256:5ba2ecec5c043bda57ead00135635524a69749360e22985a

WARC-Target-URI: https://www.digitales-archiv-und-kanon.de/

WARC-Date: 2021-03-09T09:47:53.016Z

WARC-Type: response

WARC-Record-ID: <urn:uuid:efecc0e0-b0d4-4173-a5d8-f30385ed56b2>

Content-Type: application/http; msgtype=response

WARC-Block-Digest: sha-256:51c246eb86a8daaec6db9ded10efcca5daec1f781b5e45ee2cb25709640ae784

Content-Length: 6948

HTTP/1.1 200 OK

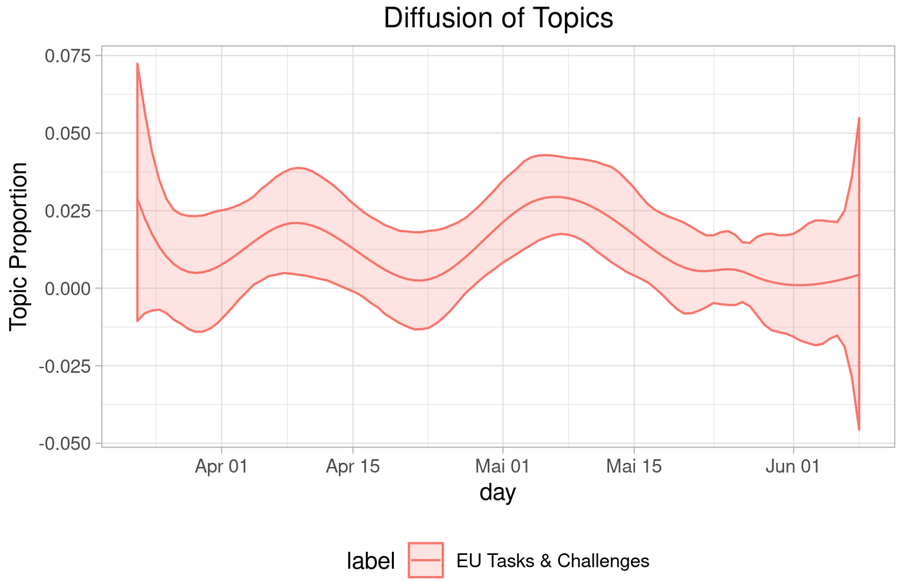
accept-ranges: bytes

access-control-allow-origin: *

age: 21

cache-control: max-age=600

>



Simon Donig

Markus Eckl

Sebastian Gassner

Florence Reiter

Malte Rehbein

RESAW 2021 | 18/06/2021

„Webarchive – Methoden der Digital Humanities in Anwendung für den Aufbau und die Nutzung von Webarchiven“

Bayerische Staatsbibliothek

Lehrstuhl für Digital Humanities
(Prof. Dr. Malte Rehbein)

Jean-Monnet-Lehrstuhl für Europäische Politik
(Prof. Dr. Daniel Göler)

Gefördert durch
DFG Deutsche Forschungsgemeinschaft

Project duration: 2018 – 2021
Project number 395175156

This Paper

Here we would like to

- *discuss*, what can, epistemologically, be seen seen as blind spots and silences with regard to webarchiving
- *share* how we proceed to make them articulate,
  - firstly by generating granular metadata-enriched derivatives and
  - secondly by analysing them with scalable reading methods;
- *reflect* our approach with regard to which forms of blind spots and silences exist at which level.

Case study: concatenated corpus of websites of print-media outlets from a crawl on European Parliament election 2019.

Specific interest: dimension of **scale and the silence of time in webarchives**.

Definitions

- *“Silences”* or *“silent pages”*: web content that exists but is temporarily silent until we make it articulate, because it is opaque to the method.
- *“Blind spot”*: limitation induced by the method as to what can be perceived at a given point after all. Fundamental epistemological barriers for the inclusion of content.
- Blind spots and silences can exist at different levels of inquiry, both when a webarchive is created and when the archived web is analysed.

Potential Blind Spots & Silences (non exhaustive):

(Research) Design

- archiving is selecting: research design, legal confines…
- delineating and initialising the crawl: parameterisation etc.
- experience vs. content focused crawl

Randomness

- random-walk pattern used by the crawler when traversing a target
- embedded content: advertising and content delivery networks

Access

- walled gardens, paywalls, intranet, dark web and other barriers
- server side vs client side crawling (content-database)

Crawl Technique Limitations

- barrierfree design, interactives pages
- social media elements
- negotiating unforeseen events in the crawl

Silences: Focus & Limits of Knowledge Generation Method

- multimodality & visuality (visuals, video)
- scale (in terms what can be perceived)
- Temporality (when read at scale)

UNIVERSITÄT PASSAU

“‘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.

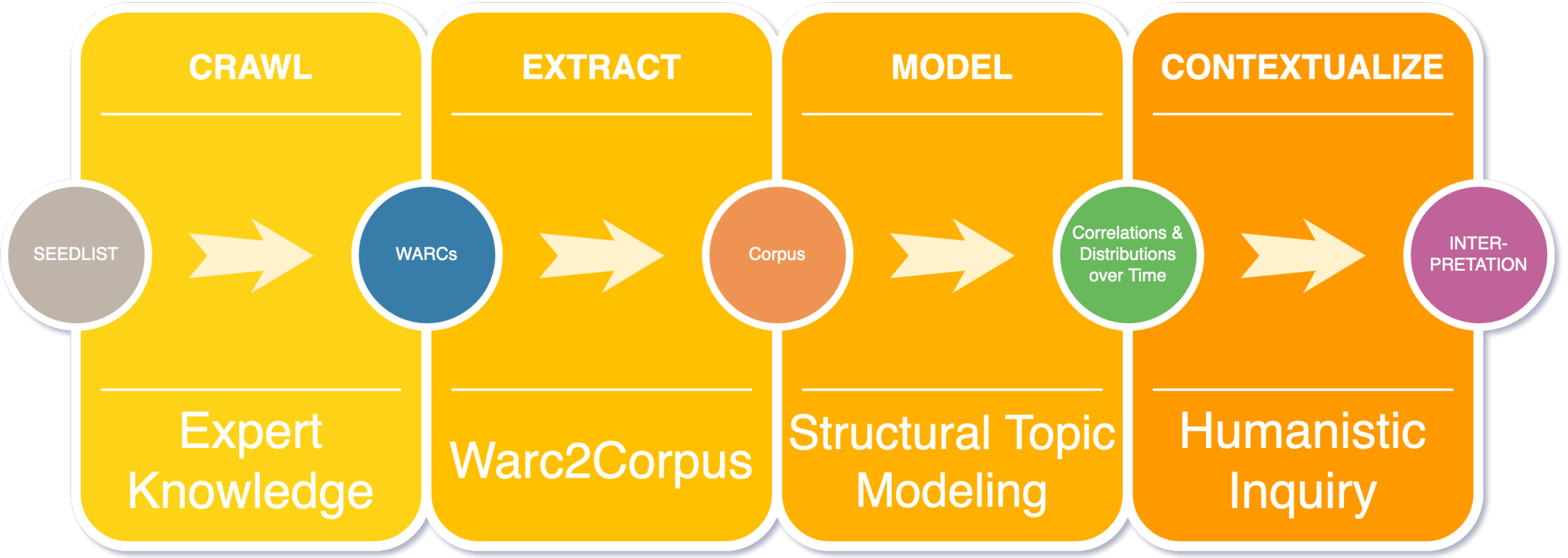
“Distant reading: where distance, let me repeat it, is a condition of knowledge: it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes-or genres and systems.”

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

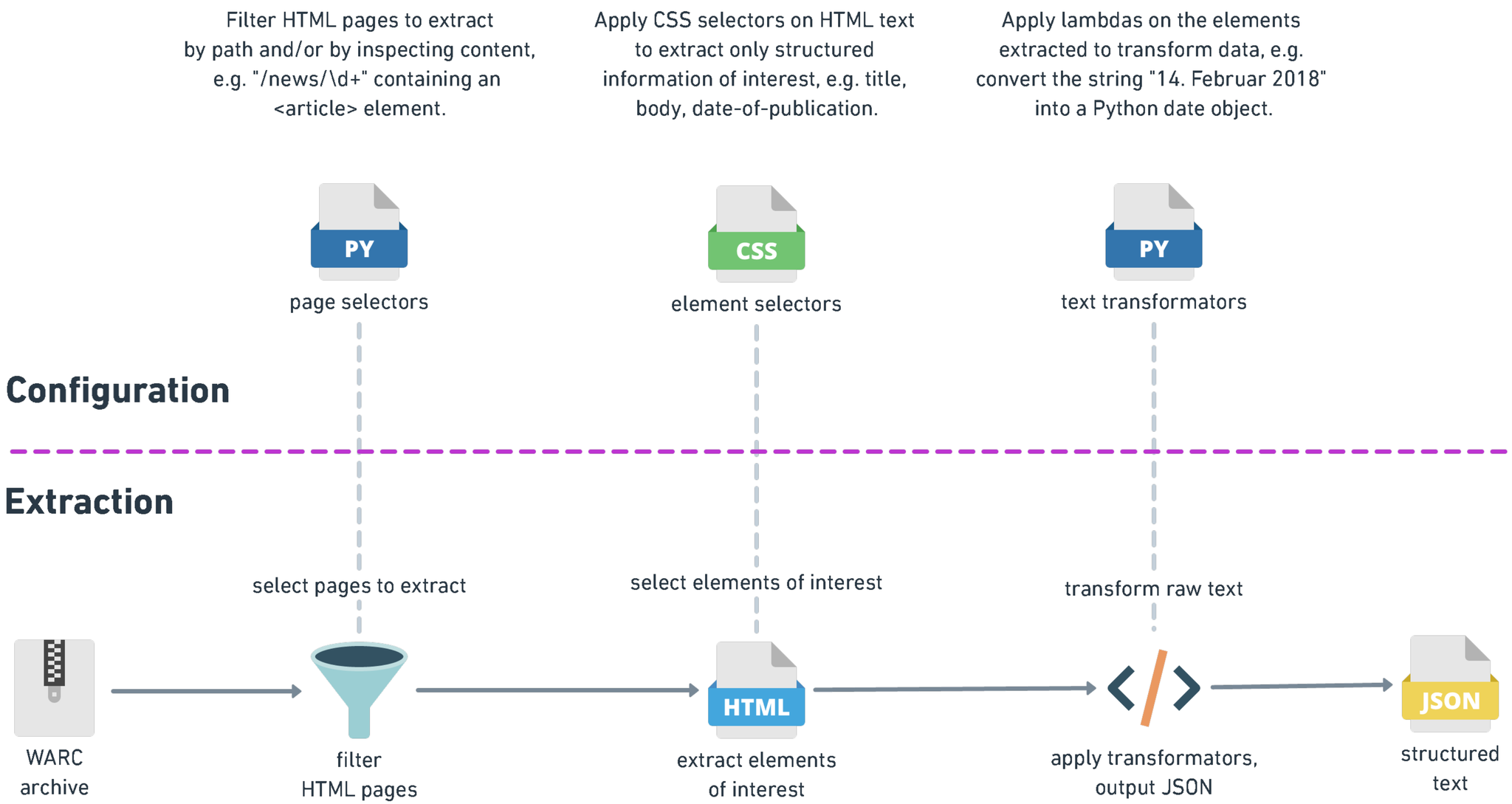
- **Time is essential** for the exploration of historical phenomena, like *history of concepts* or *discourse*.
- When relying on **close reading we explicitly or implicitly take into account the temporal context** of a given webpage.
- **Time is usually silent in archives**: no explicite, standardised and widely applied way of recording creation and alteration of webpages.
- WARC metadata (crawldate) can be a proxy but are not really helpful.

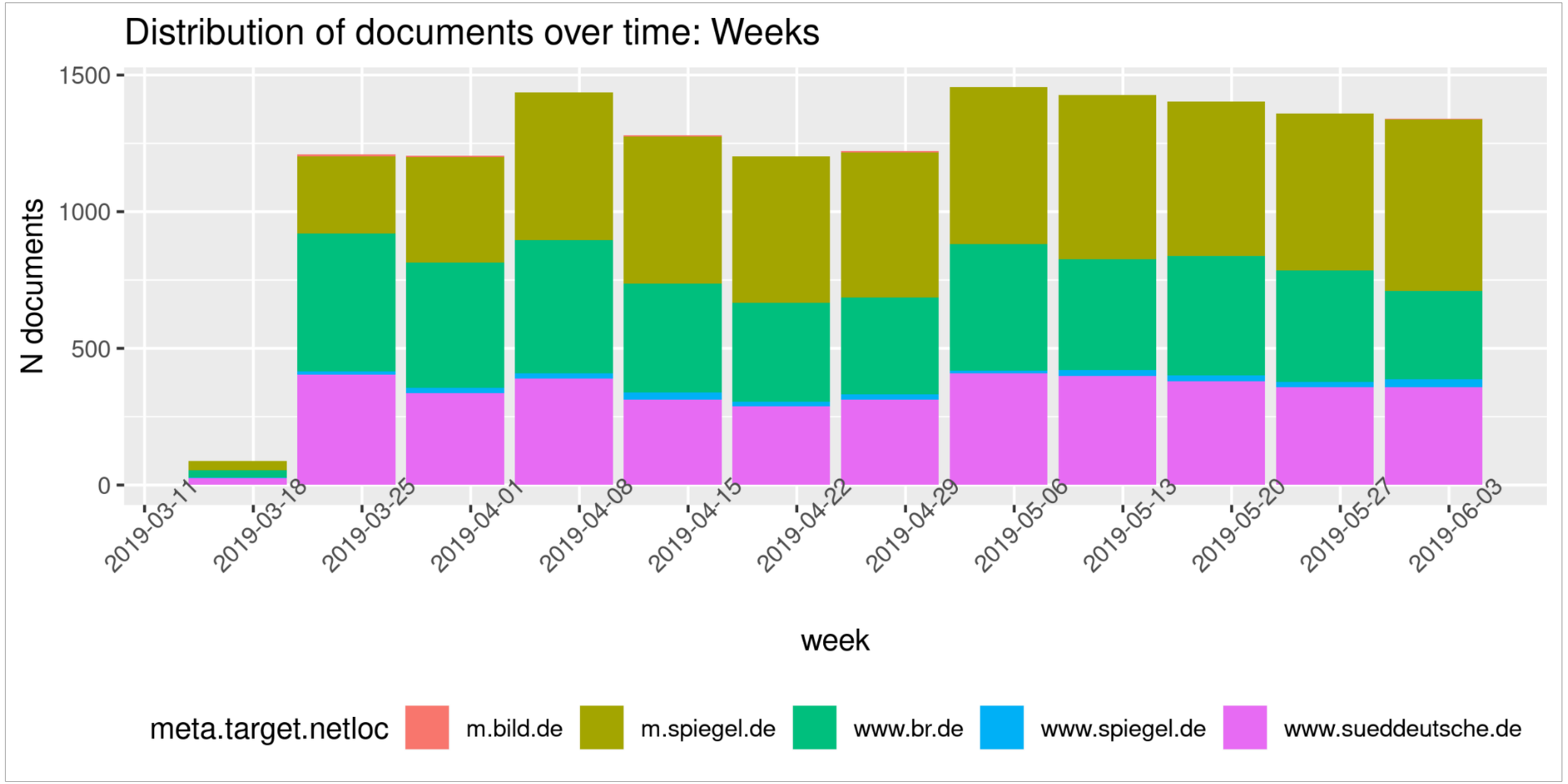
Method & Illustrative Case Study

Workflow


CRAWL
EXTRACT
MODEL
CONTEXTUALIZE
SEEDLIST
WARCs
Corpus
Correlations & Distributions over Time
INTER-PRETATION
Expert Knowledge
Warc2Corpus
Structural Topic Modeling
Humanistic Inquiry

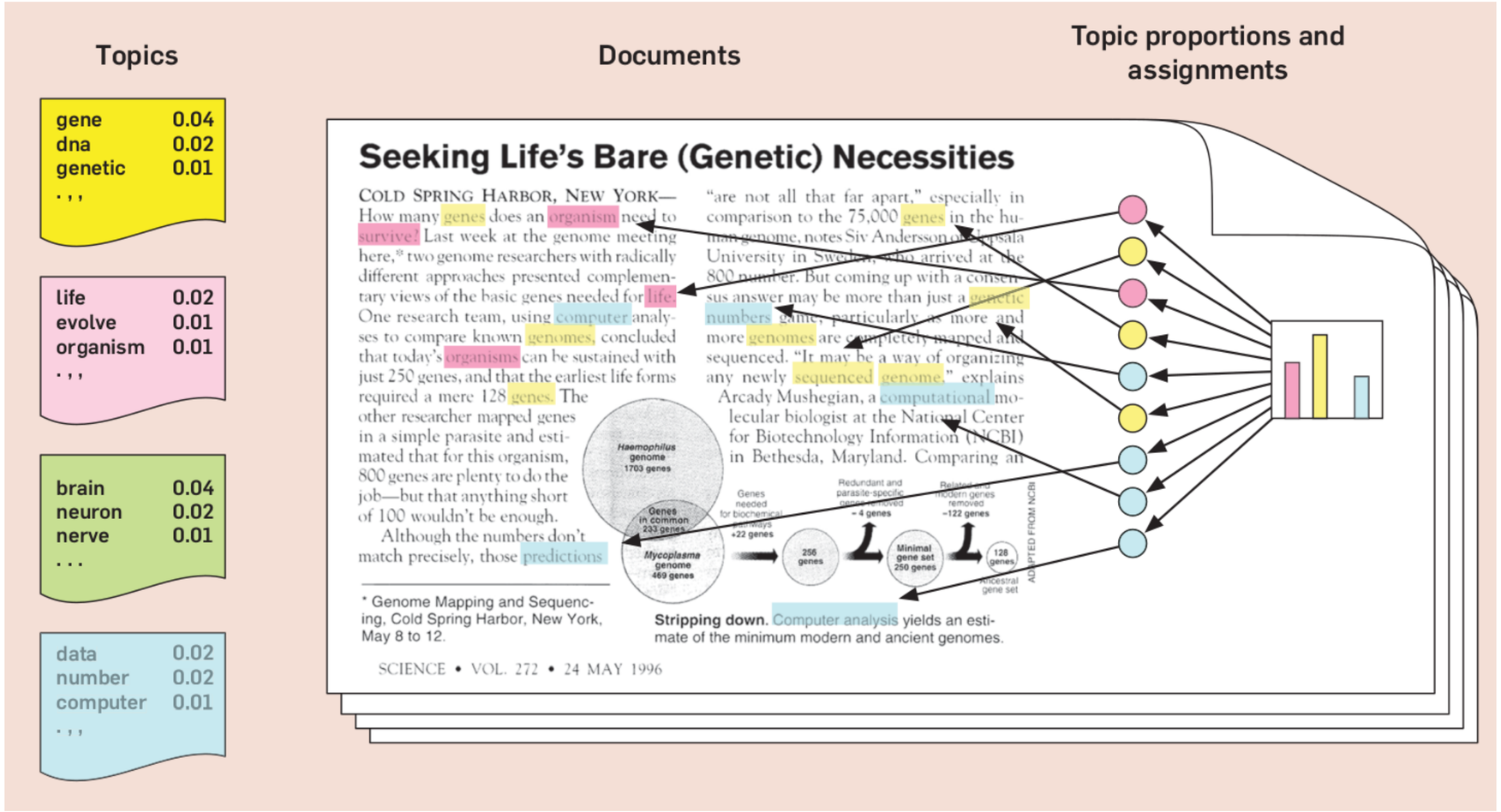
Extraction


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.
PY
page selectors
CSS
element selectors
PY
text transformators
Configuration
Extraction
select pages to extract
select elements of interest
transform raw text
WARC archive
filter HTML pages
HTML
extract elements of interest
apply transformators, output JSON
JSON
structured text



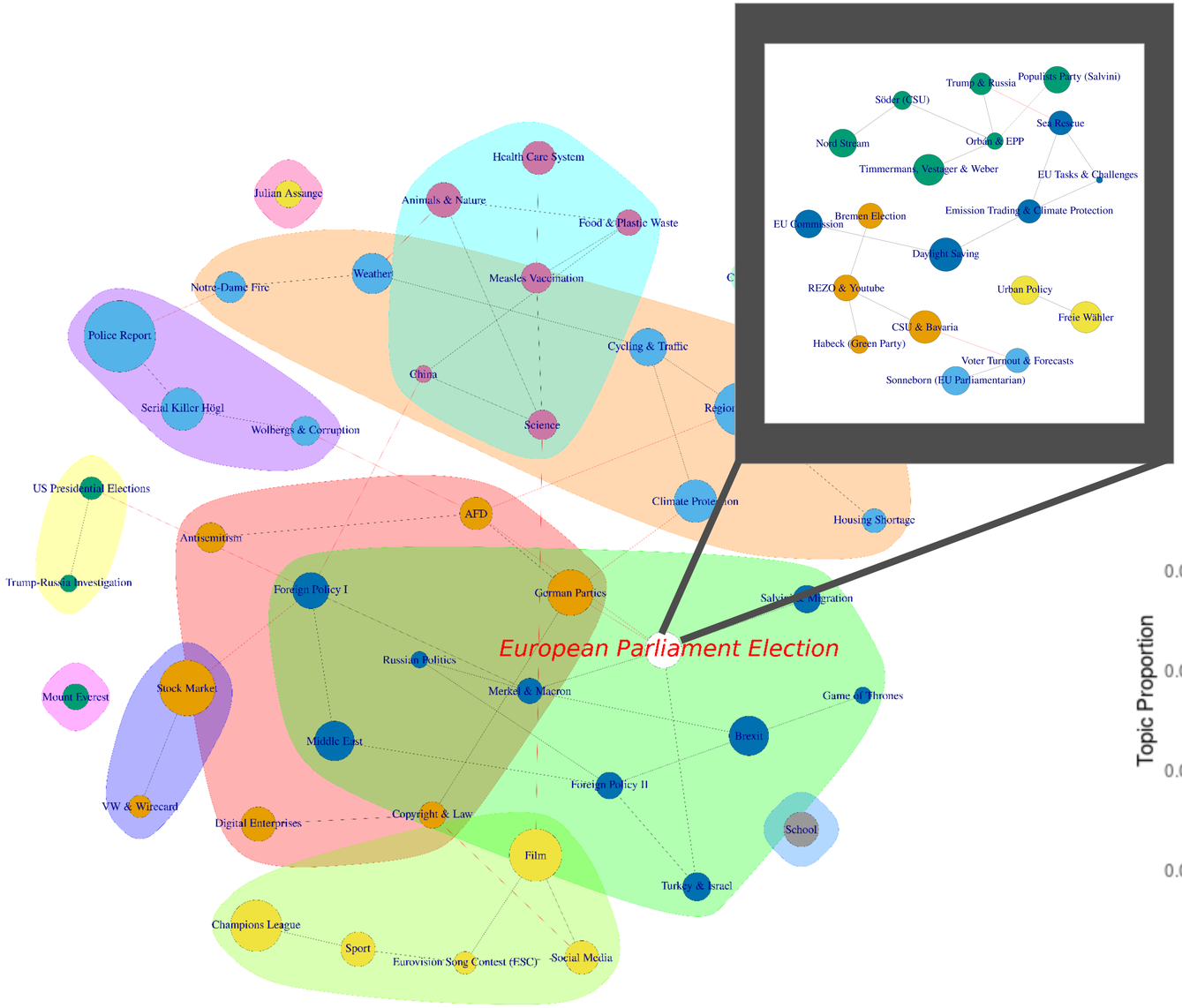
N = 14.627 Documents

Latent Dirichlet Allocation(LDA) & Structural Topic Modeling (STM)



Blei, David M., Andrew Y. Ng, and Michael I. Jordan. „Latent Dirichlet Allocation". Journal of Machine Learning Research 3 (2003): 993–1022.

European Parliament Election in a Topic Cluster

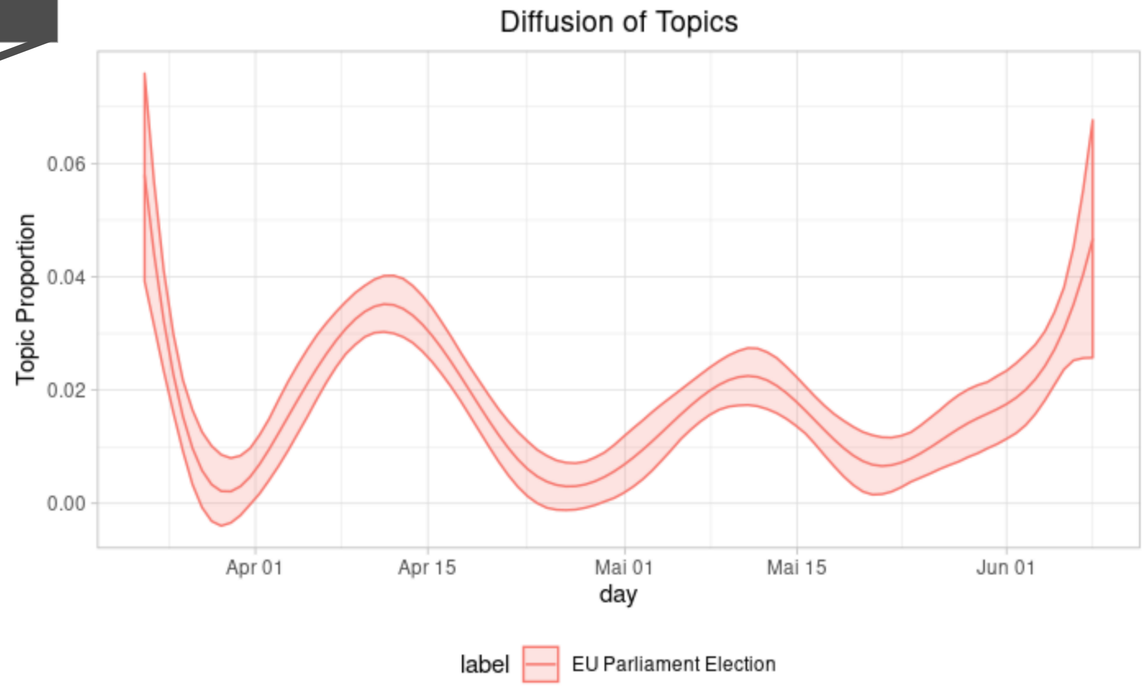


Topic:

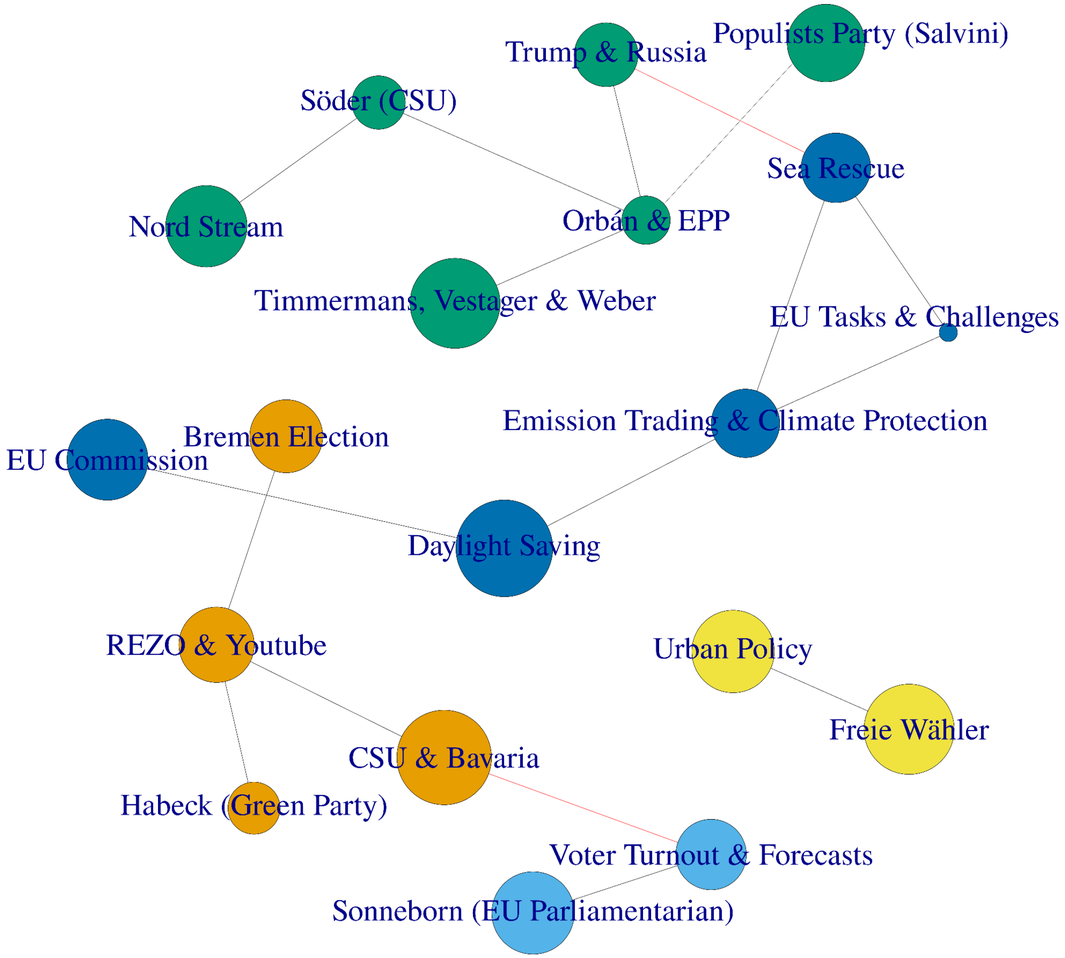
European Parliament Election

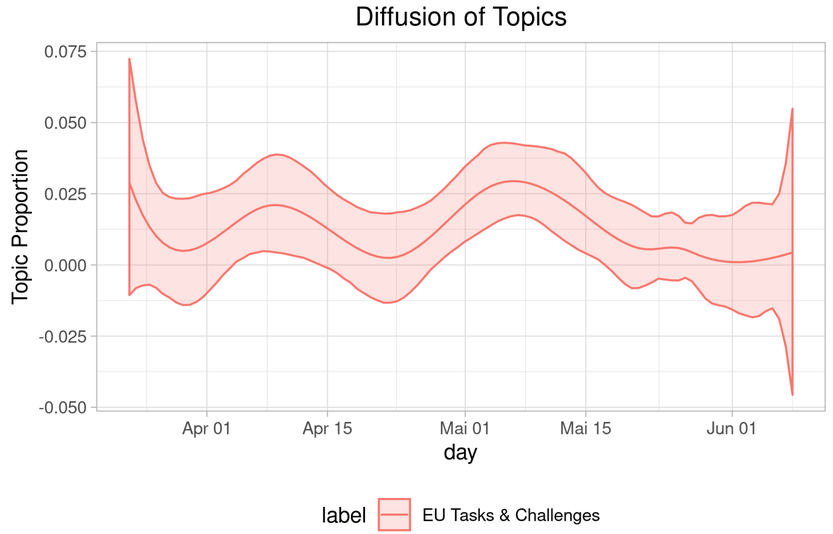
Wordlist:

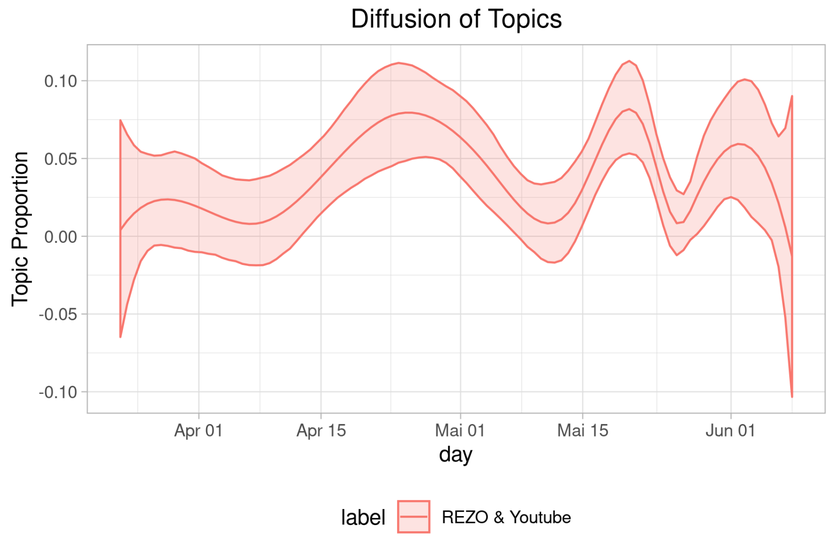
europawahl, manfred_weber, evp, afd, partei, europaparlament, csu, timmermans, spitzenkandidaten



Broad Topics and Discursive Trends: the Rezo Case







N = 1.318 Documents

- **Extraction** is blind to all archived content that cannot be associated with a particular timestamp - trade-off between granularity and inclusiveness.
- **Topic modeling** is a form of complexity reduction thus virtually blind to low frequency content. Furthermore the method influenced by hyper-parameterizations such as the predetermined number of topics.
- **Mitigation?** Research design that combines close and distant reading practices.
- Blind to **multimodal compounds**.

- Blind Spots and Silences intercede at different levels of knowledge extraction from webarchives.
- Our approach to addressing the challenge of scale and the silence of temporality in the archived web:
  - Warc2Corpus, a method for granular extraction of well formed structured data
  - Distant Reading methods such as Topic Modeling/STM, both for exploration and analysis
- Future work:
  - Improving usability of w2c
  - Integration with existing data extraction instruments such as AUT
  - exploration how far these methods could be used to enrich WARC metadata