Finding historical discourse on natural environment: Australian newspapers 1900-1990

Peeter Tinits (University of Tartu)

DHNB 2023

Research background

Industrial Modernity - 250 years of cultural accumulation in industrial societies and their socio-technical systems

Kanger & Schot 2019: One characteristic of Industrial Modernity is a lack attention paid to natural environment and its welfare: resources are infinite, nature is endlessly adaptable, problems like pollution can be fixed later.

During Industrial Modernity people have worried less about the welfare of nature compared to societal issues, politics, trade, economy, jobs. Can we see a change in this?

Studying history with historical newspapers

It’s a sociological question: e.g. which is more important - economic growth or wellbeing of natural environment? About masses, not single texts.

We have surveys (like World Value Survey) on this since 1981, but how to go further?

=> Digitized newspapers

- Newspapers carry popular discourse.
- Digitized newspapers can be studied en masse.

YORKSHIRE
Daily
THE GUARDIAN
DAILY EXPRESS
THE TIMES
Inflation rate rises to 18.4% and trade gap widens
Double blow to Government
Polish Premier forced to resign after attacks at party congress

Searching for discourse

How much was natural environment discussed in mainstream newspapers?

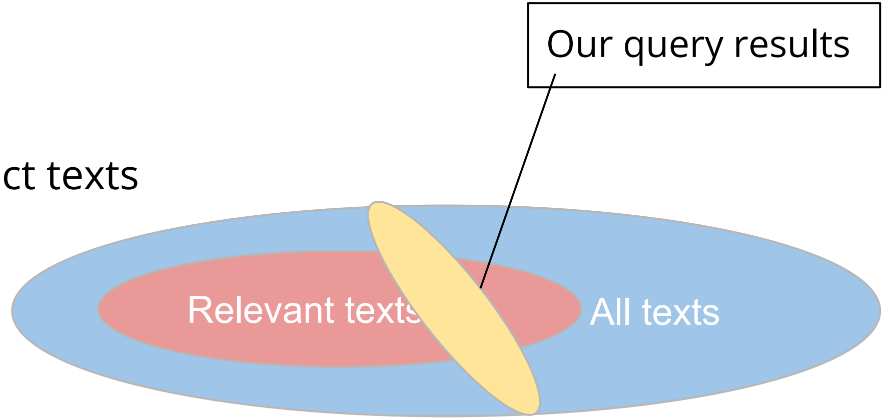
Problems to tackle

- Language change (words get new meanings)
- Discourse change (common phrases rise and fall)
- OCR errors (digitization errors)
- Scope of each is difficult to know beforehand.

Query construction

Finding discourse

- Precision: We want to find mostly/only correct texts
- Recall: We want to find as many as possible



A good keyword has the following properties:

- Accurately identifies the idea, object, person, or place under analysis.
- Does not return a significant number of irrelevant hits.
- Is not susceptible to linguistic fads (unless language change is the object of study)
- It not susceptible to OCR errors.

(Nicholson 2012)

From keywords to complex queries

A good set of keywords is difficult to construct, and requires iterative exploration of the datasets by domain experts.

Keywords search can be expanded by various text mining tools - e.g. topic models, word embeddings, pre-made natural language models.

For example

- Oberbichler & Pflanzelter (2021) classified articles as relevant or not and then used topic models to find similar articles.
- De Wildt, van de Poel, Chappin (2022) used keywords as anchor words and built topics around them.

Many different ways to improve queries.

Case: Environment in Australia 1900-1990

We want to find discourse on natural environment in Australia 1900-1990

- Large text corpus
- Long time-frame
    - Language and discourse change
- Diverse topics
    - E.g sustainable development, environmental policies, protection of animals, agricultural advice, natural disasters

Difficult to construct specialist keywords. Unknown how much is there.

Data

Sydney Morning Herald (1900-1940) and Canberra Times (1930-1995)

Source: Trove archives (Australian National Library)

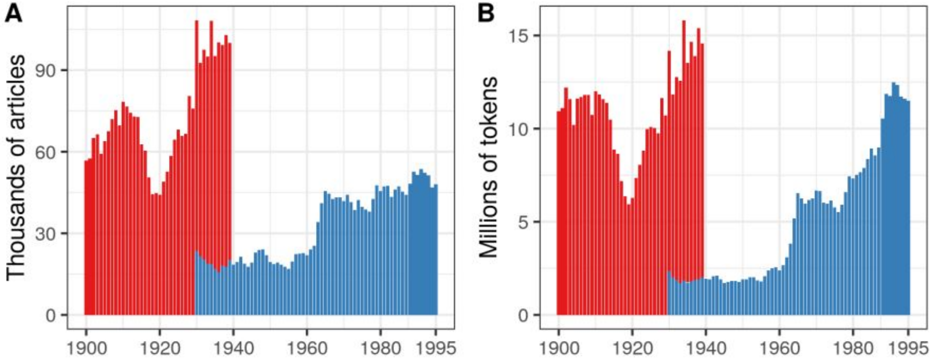
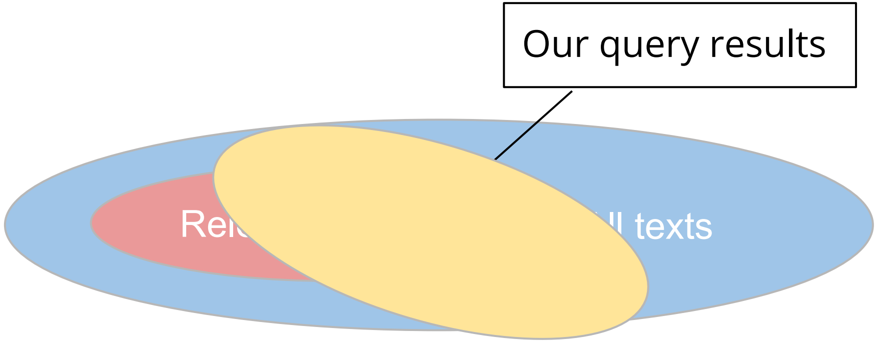


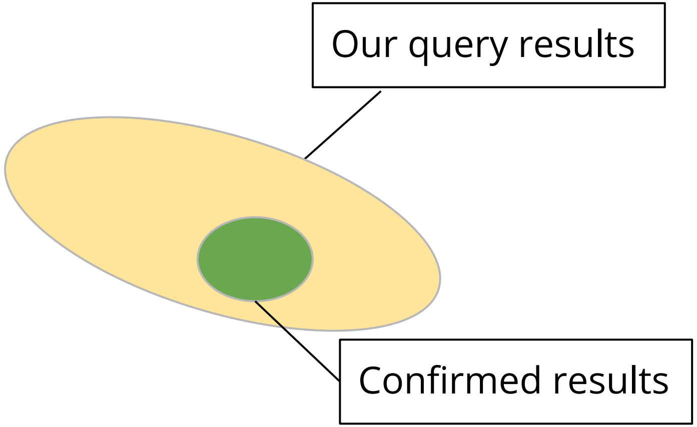
Figure 1: The sources used in the example. A: Thousands of articles per year, B: Millions of tokens per year. Sidney Morning Herald (1900-1940) in red, Canberra Times (1930-1995) in blue.

Our approach

Start with broad keywords to get a large collection of results,

then refine the set via topic models, and exclude many irrelevant ones.





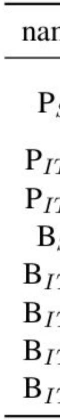
Our approach

Broad terms

- Natur[ea]*, environment* (nature, natural, environmentalism)
- E.g. ‘human nature’, ‘school environment’

Use topic models to improve precision

- Keyword-in-context (25 words + MATCH + 25 words) as document
- Topics annotated for relevance
- Keep only matches with above threshold relevance



nar

P

P

P

B

B

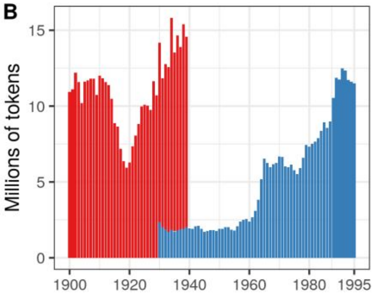
B

B

B

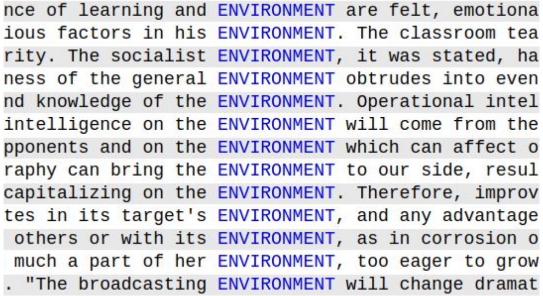
Our approach

Lemmatized corpus



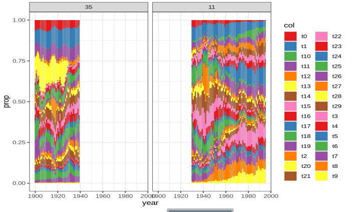
Keywords:
natur[ea]*,
environment*,
conservation*

KWICs +/- 25 words



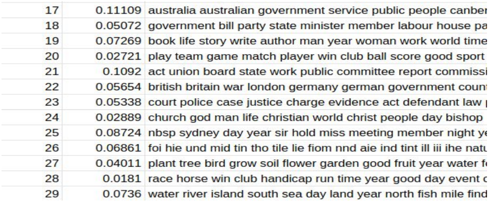
```
nce of learning and ENVIRONMENT are felt, emotiona
ious factors in his ENVIRONMENT. The classroom tea
rity. The socialist ENVIRONMENT, it was stated, ha
ness of the general ENVIRONMENT obtrudes into even
nd knowledge of the ENVIRONMENT. Operational intel
intelligence on the ENVIRONMENT will come from the
pponents and on the ENVIRONMENT which can affect o
raphy can bring the ENVIRONMENT to our side, resul
capitalizing on the ENVIRONMENT. Therefore, improv
tes in its target's ENVIRONMENT, and any advantage
 others or with its ENVIRONMENT, as in corrosion o
 much a part of her ENVIRONMENT, too eager to grow
. "The broadcasting ENVIRONMENT will change dramat
```

LDA topic models
10, 20, 30 topics

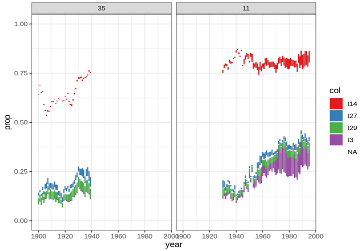


Annotate topics for relevance

Based on top 50 words, include at least a few relevant e.g. plant, river, sustainable, green



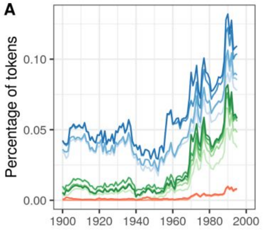
17	0.11109	australia australian government service public people canber
18	0.05072	government bill party state minister member labour house pa
19	0.07269	book life story write author man year woman work world time
20	0.02721	play team game match player win club ball score good sport
21	0.1092	act union board state work public committee report commissi
22	0.05654	british britain war london germany german government count
23	0.05338	court police case justice charge evidence act defendant law
24	0.02889	church god man life christian world christ people day bishop
25	0.08724	nbsp sydney day year sir hold miss meeting member night ye
26	0.06861	foi hie und mid tin tho tile lie fiom nnd aie ind tint ill iii ihe natu
27	0.04011	plant tree bird grow soil flower garden good fruit year water f
28	0.0181	race horse win club handicap run time year good day event
29	0.0736	water river island south sea day land year north fish mile find



Just relevant topics over time

Keep only matches with topic content above a threshold (we tried a few)

E.g. 491,057 ->
187,327 matches



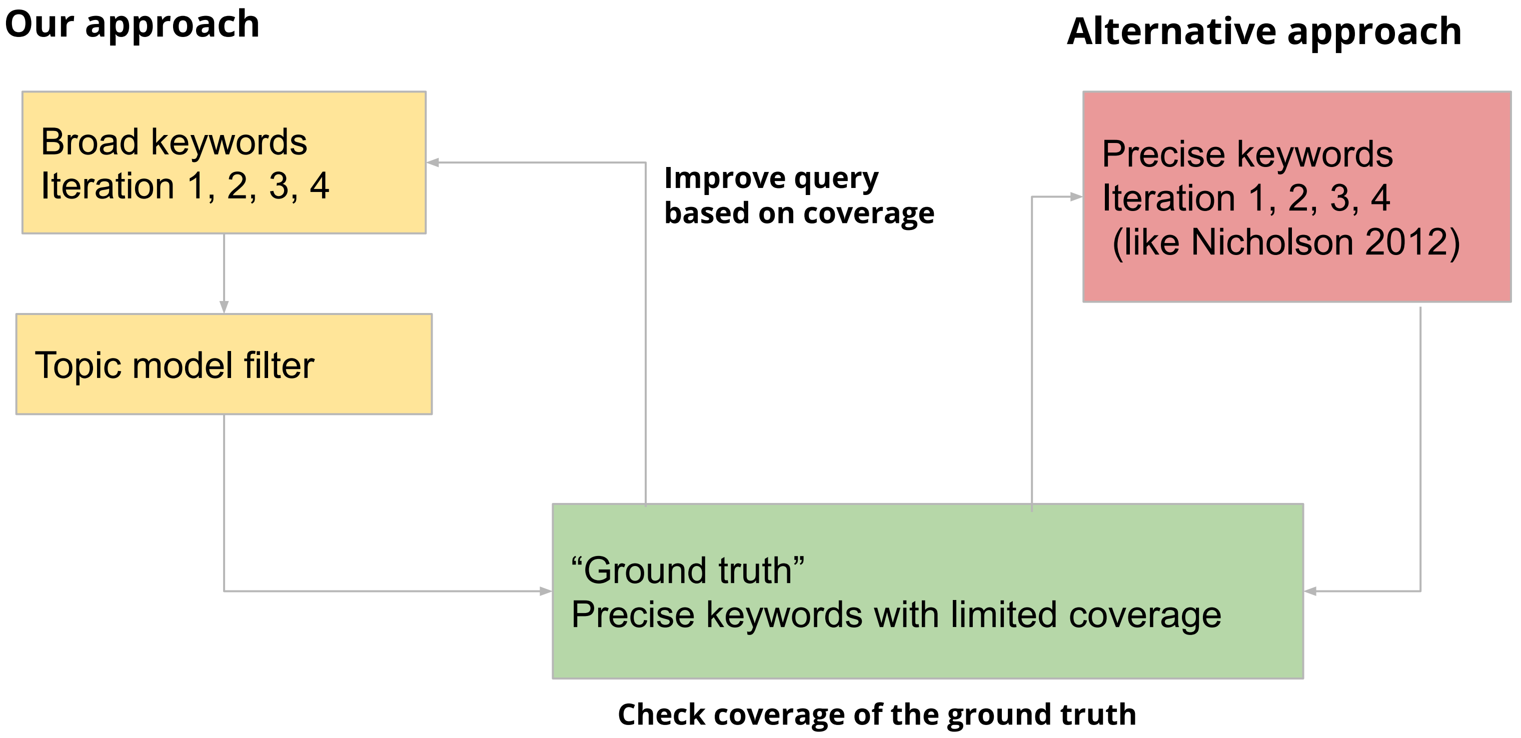
Prevalence over time

Goals

Our goals are here

- To measure changes in prevalence over time
- To check that the workflow works as we expect

Testing the approach



Our approach

Alternative approach

Broad keywords
Iteration 1, 2, 3, 4

Topic model filter

Improve query based on coverage

Precise keywords
Iteration 1, 2, 3, 4
(like Nicholson 2012)

“Ground truth”
Precise keywords with limited coverage

Check coverage of the ground truth

Testing the approach

For each query

- Recall = how many texts from “ground truth” found
- Precision = check 100 matches for relevance

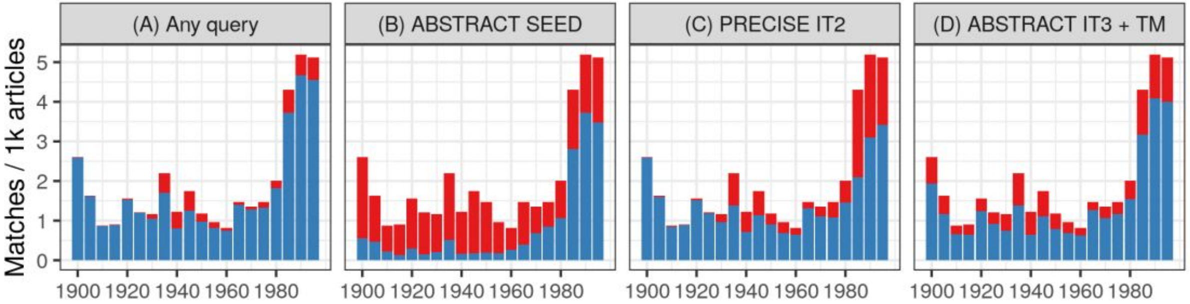


Figure 2: The recall of different approaches by five year periods as a proportion of articles. A: Articles found by any query. B: Initial seed for broad terms. C: Articles found by best precise query (Precise, iteration 2). D: Articles found by best query via topic models (Broad, iteration 3). The height shows the number of matches per 1,000 articles in each 5-year slice. Blue matches were successfully found, red matches were not.

Our keywords

“Ground truth” keywords

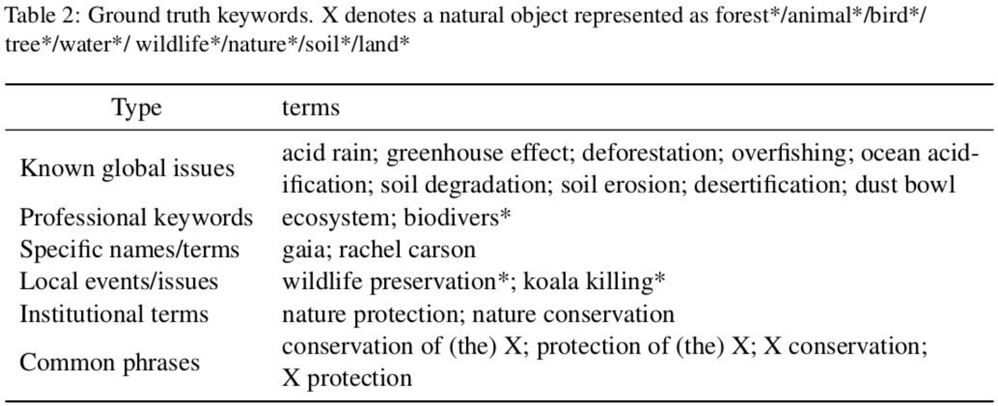
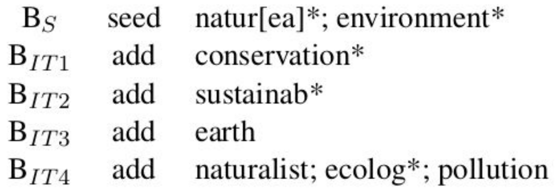


Table 2: Ground truth keywords. X denotes a natural object represented as forest*/animal*/bird*/tree*/water*/ wildlife*/nature*/soil*/land*

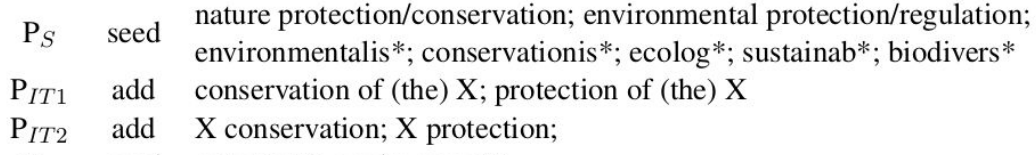
Type	terms
Known global issues	acid rain; greenhouse effect; deforestation; overfishing; ocean acidification; soil degradation; soil erosion; desertification; dust bowl
Professional keywords	ecosystem; biodivers*
Specific names/terms	gaia; rachel carson
Local events/issues	wildlife preservation*; koala killing*
Institutional terms	nature protection; nature conservation
Common phrases	conservation of (the) X; protection of (the) X; X conservation; X protection

Broad keywords



B_S	seed	natur[ea]*; environment*
B_{IT1}	add	conservation*
B_{IT2}	add	sustainab*
B_{IT3}	add	earth
B_{IT4}	add	naturalist; ecolog*; pollution

Precise keywords



P_S	seed	nature protection/conservation; environmental protection/regulation; environmentalis*; conservationis*; ecolog*; sustainab*; biodivers*
P_{IT1}	add	conservation of (the) X; protection of (the) X
P_{IT2}	add	X conservation; X protection;

Our query results

Stricter threshold increases precision

- Trade-off between precision and recall still there.
- Used 33% content in 50% of models as threshold

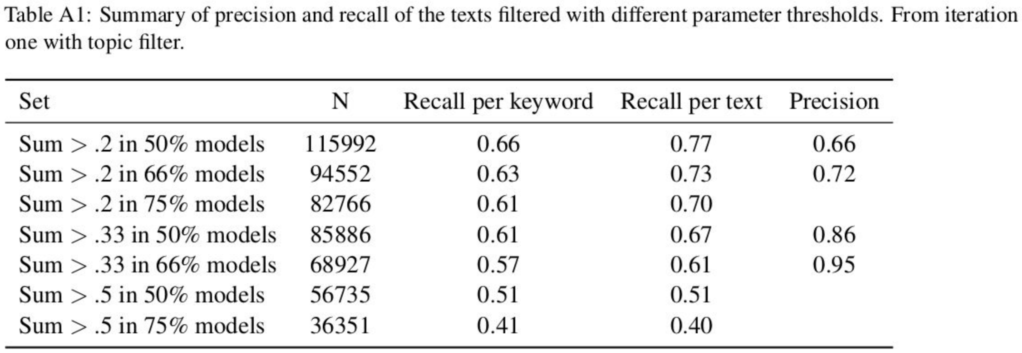


Table A1: Summary of precision and recall of the texts filtered with different parameter thresholds. From iteration one with topic filter.

Set	N	Recall per keyword	Recall per text	Precision
Sum > .2 in 50% models	115992	0.66	0.77	0.66
Sum > .2 in 66% models	94552	0.63	0.73	0.72
Sum > .2 in 75% models	82766	0.61	0.70	
Sum > .33 in 50% models	85886	0.61	0.67	0.86
Sum > .33 in 66% models	68927	0.57	0.61	0.95
Sum > .5 in 50% models	56735	0.51	0.51	
Sum > .5 in 75% models	36351	0.41	0.40	

Our query results

Applying topic model filters

- 10+ more matches
- Great improvement in precision from topic models
- Marginal loss in recall from topic models

Precision

- 0.21 -> 0.83
- 0.30 -> 0.67
- 0.31 -> 0.77

Recall

- 0.39 -> 0.32
- 0.86 -> 0.78
- 0.86 -> 0.76

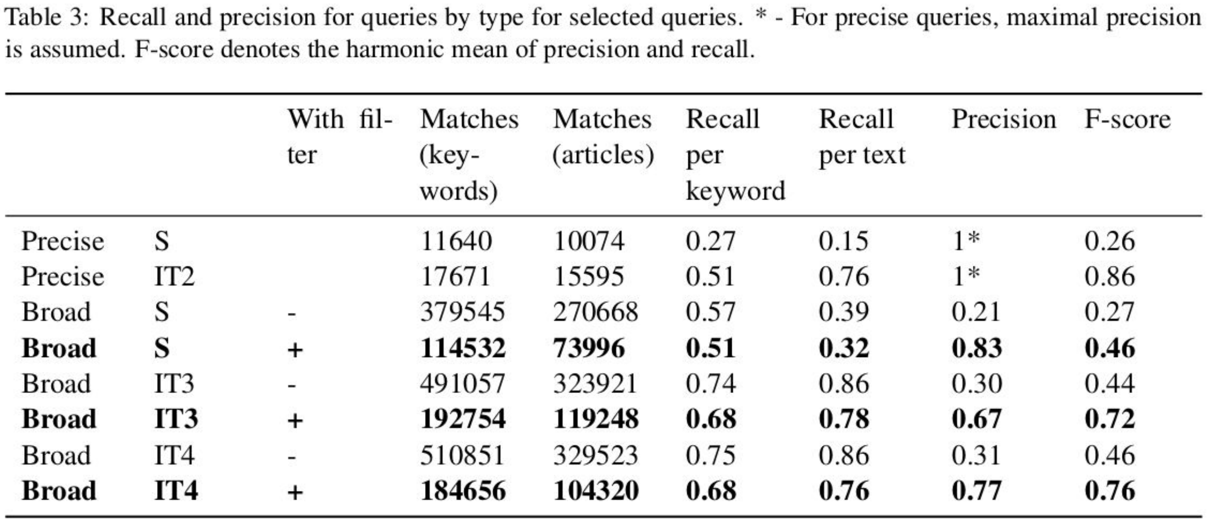


Table 3: Recall and precision for queries by type for selected queries. * - For precise queries, maximal precision is assumed. F-score denotes the harmonic mean of precision and recall.

		With filter	Matches (keywords)	Matches (articles)	Recall per keyword	Recall per text	Precision	F-score
Precise	S		11640	10074	0.27	0.15	1*	0.26
Precise	IT2		17671	15595	0.51	0.76	1*	0.86
Broad	S	-	379545	270668	0.57	0.39	0.21	0.27
Broad	**S**	**+**	**114532**	**73996**	**0.51**	**0.32**	**0.83**	**0.46**
Broad	IT3	-	491057	323921	0.74	0.86	0.30	0.44
Broad	**IT3**	**+**	**192754**	**119248**	**0.68**	**0.78**	**0.67**	**0.72**
Broad	IT4	-	510851	329523	0.75	0.86	0.31	0.46
Broad	**IT4**	**+**	**184656**	**104320**	**0.68**	**0.76**	**0.77**	**0.76**

Measuring prevalence

Impact of improved query on results of prevalence

- Broad no filter
- Broad with filter
- Precise

Iterations

- 1, 2, 3, 4
- Darker = newer

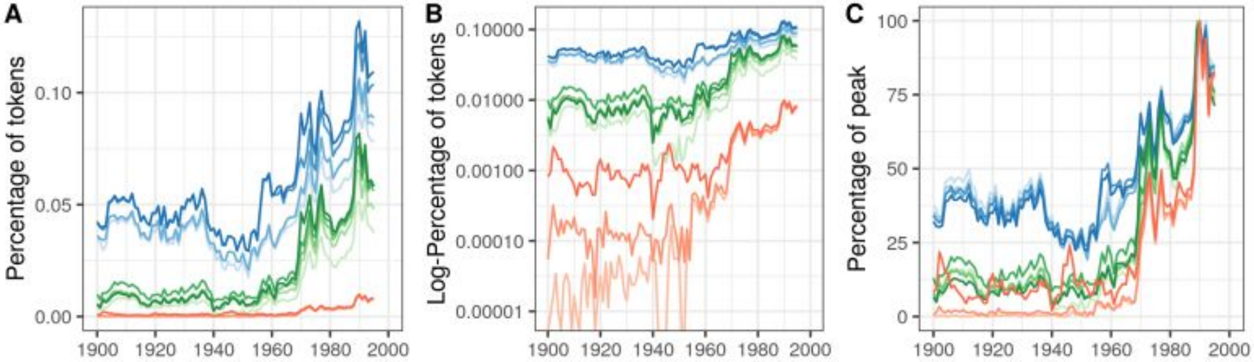


Figure 3: Results of the queries colored by type and iteration. Blue - broad query, no filter. Green - broad query - with filter. Red - precise query. The lightest line is the seed set of keywords, colors get darker with each iteration. A: Frequencies on a linear scale; B: Frequencies on a logarithmic scale, C: Frequencies as a proportion of their maximal value.

Measuring prevalence

Best keyword sets

- Broad no filter
- Broad with filter
- Precise

Top - trendlines

Bottom - first derivations

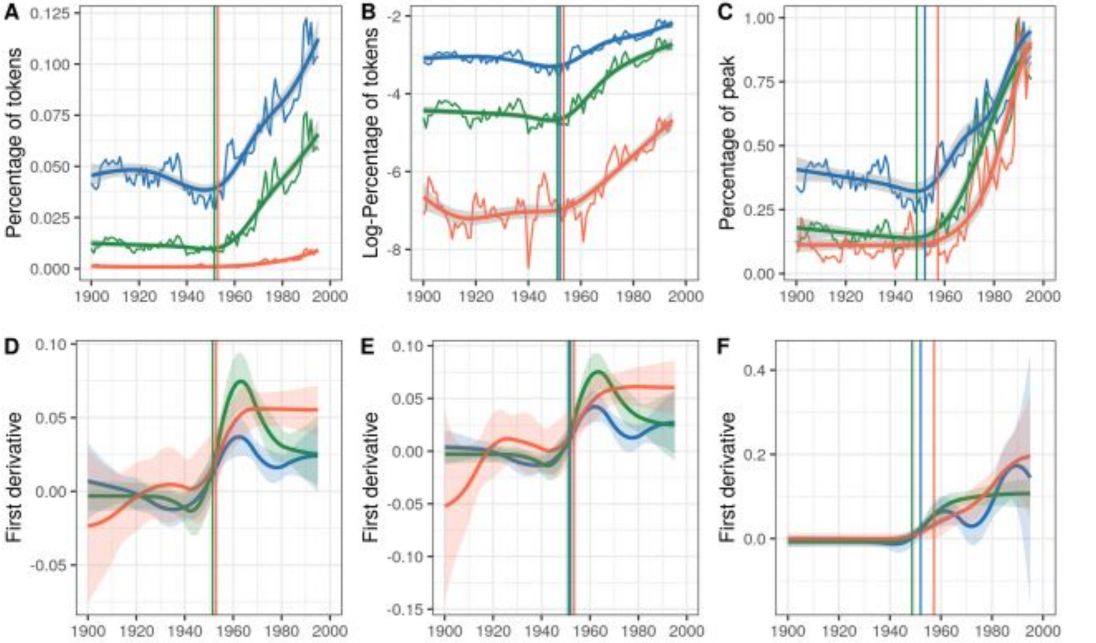


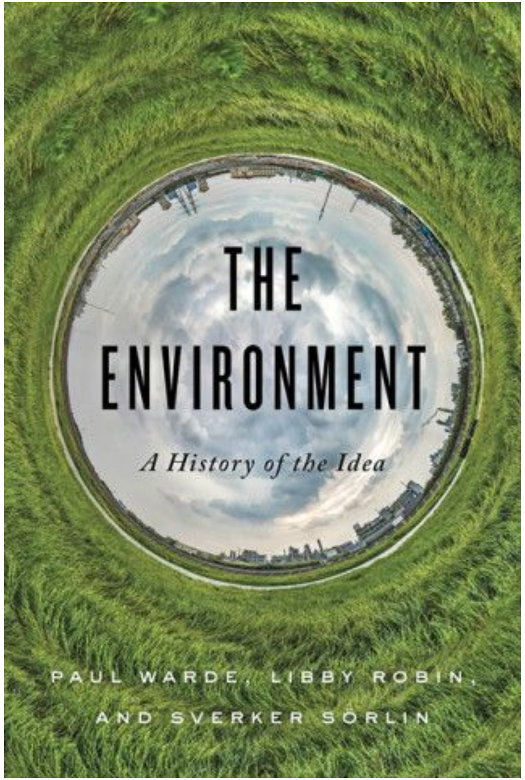
Figure 4: Search results of the best model for historical interpretation.A: Frequencies on a linear scale; B: Frequencies on a logarithmic scale, C: Frequencies as a proportion of their maximal value. D-F: The first derivatives on the graphs above them. The vertical lines in each graph show the beginning of the period of growth - the point when the first derivative was significantly different from 0.

Discussion

The quick growth of 1960s agrees with literature on environmentalism

Using topic models to filter the results greatly improves precision of the results

Using broad keywords with such filters can be a viable approach to reduce the requirements of domain expertise and corpus experimentation

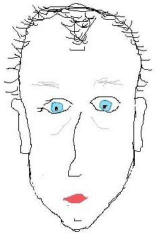


Open topics for DH

Keyword frequency analysis is quite common in DH, with some open issues:

- Difficult to establish ground truth, describe precision and recall. Should we do this more in DH or trust the methods?
- Measuring prevalence: linear, log-linear, relative etc. Which growth to describe?
- Iterations of keyword sets. How (much) should we document them in DH?

Thank you



UNIVERSITY OF TARTU

“Reshaping Estonian energy, mobility and telecommunications systems on the verge of the Second Deep Transition” (2019-2023)