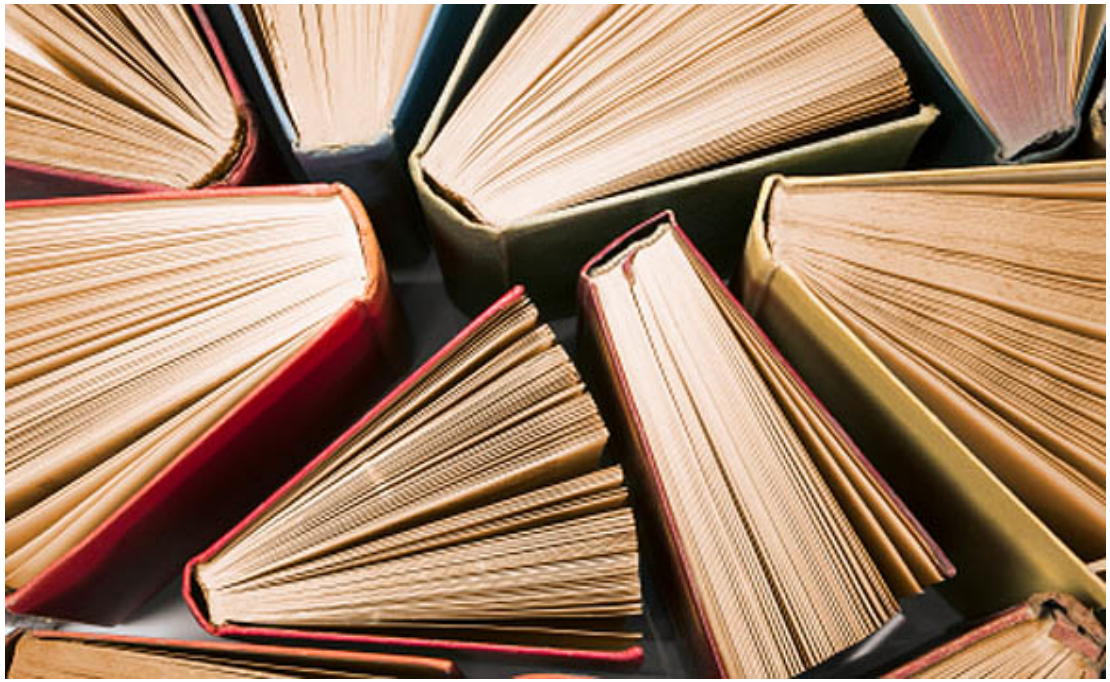
FONDAZIONE
BRUNO KESSLER



Developing a platform for the analysis of historical sources… and more

Sara Tonelli
Digital Humanities Joint Research Project
Fondazione Bruno Kessler

FONDAZIONE BRUNO KESSLER

The Digital Humanities joint research project

ALESSANDRO MARCHETTI
E-mail:
amarchetti@fbk.eu
PHONE:
0461314974
VIEW PROFILE

STEFANO MENINI
E-mail:
menini@celct.it
PHONE:
0461314942
VIEW PROFILE

GIOVANNI MORETTI
E-mail:
moretti@fbk.eu
PHONE:
0461314807
VIEW PROFILE

RACHELE SPRUGNOLI
E-mail:
sprugnoli@fbk.eu
PHONE:
0461314879
VIEW PROFILE

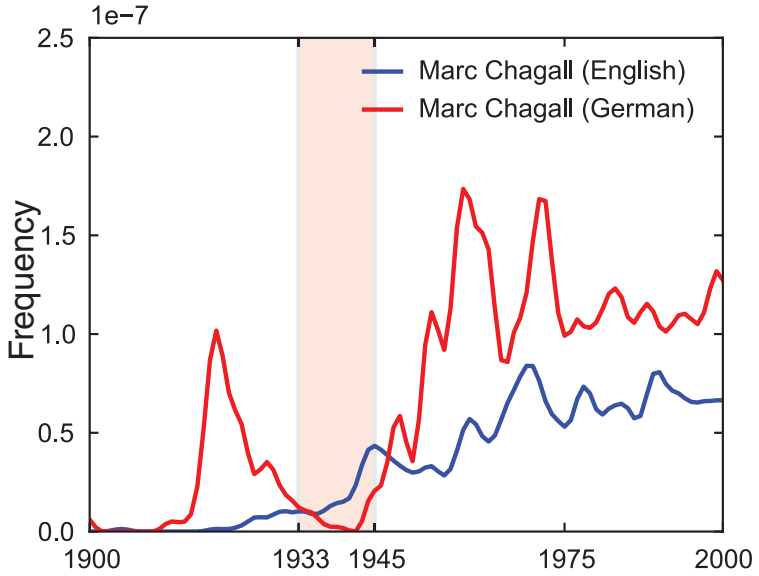
SARA TONELLI
E-mail:
satonelli@fbk.eu
PHONE:
0461314542
VIEW PROFILE

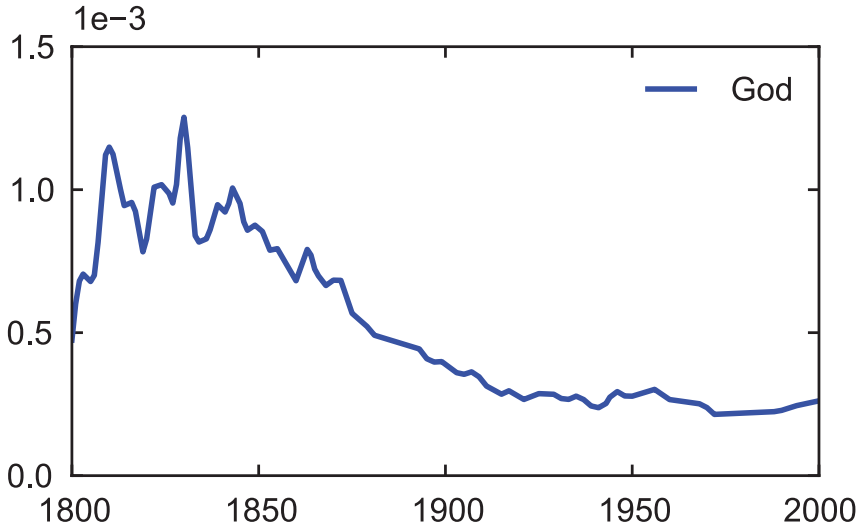
EMANUELE PIANTA

A bit of Culturomics

Quantitative Analysis of Culture Using Millions of Digitized Books, Jean-Baptiste Michel et al., Science 331, 176-182 (2010)

- Corpus of digitized texts containing about 4% of GoogleBooks.
- *Culturomics*: application of data collection and analysis techniques to the study of human culture





"God is not dead but needs a new publicist"

FONDAZIONE BRUNO KESSLER

“Close” and “distant” reading

Graphs, Maps, Trees: Abstract Models for a Literary History, Franco Moretti, 2007, Verso, London.

- Mode of analysis called *distant reading* (vs. *close reading*), where thousands of books are automatically scanned and analyzed in an attempt to understand long-term and large-scale trends
- Empirical data are the key to better knowledge of literature because they are “independent from interpretation”

FONDAZIONE BRUNO KESSLER

The ALCIDE project

Analysis of **L**anguage and **C**ontent **I**n a **D**igital **E**nvironment:

- A digital history project: "*capacity, accessibility, flexibility, diversity, manipulability, interactivity and hypertextuality*" (Cohen and Rosenzweig, 2005).
- First collaboration between ICT center and Italian-German Historical Institute at FBK
- **Goal**: develop an online platform to perform temporal, geographical, linguistic and semantic analysis of historical documents
- Tailor state-of-the-art tools for Natural Language Processing to the requirements of researchers in contemporary history

FONDAZIONE BRUNO KESSLER

ALCIDE Project: Requirements

Requirements:

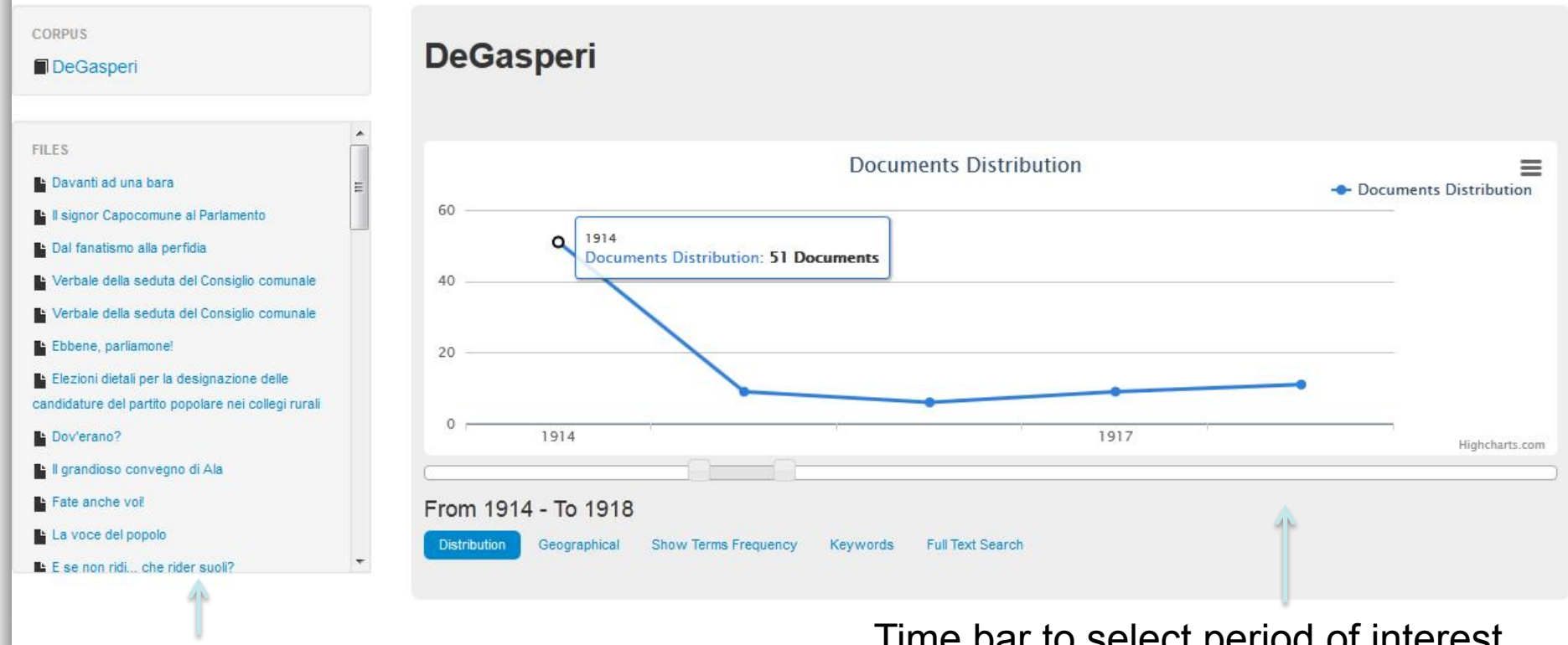
- Select the temporal span of the research on the fly
- Pass easily from *distant* to close *reading*
- The platform should not be bound to a specific historical period
- Reduce manual effort to the minimum
- Intuitive interface, quick system response, modularity
- Multilinguality
- Standard formats (XML for input files, HTML5 for visualization)

ALCIDE Project: : First use case

- Complete collection of Alcide De Gasperi's writings (1881 – 1954)
- Around 3.000 documents, 3.000.000 tokens (~70.000 lemmas)
- Includes articles, speeches transcriptions, official documents, etc.

ALCIDE: “Distant” reading

1. Documents distribution over time (metadata: **year** of publication)



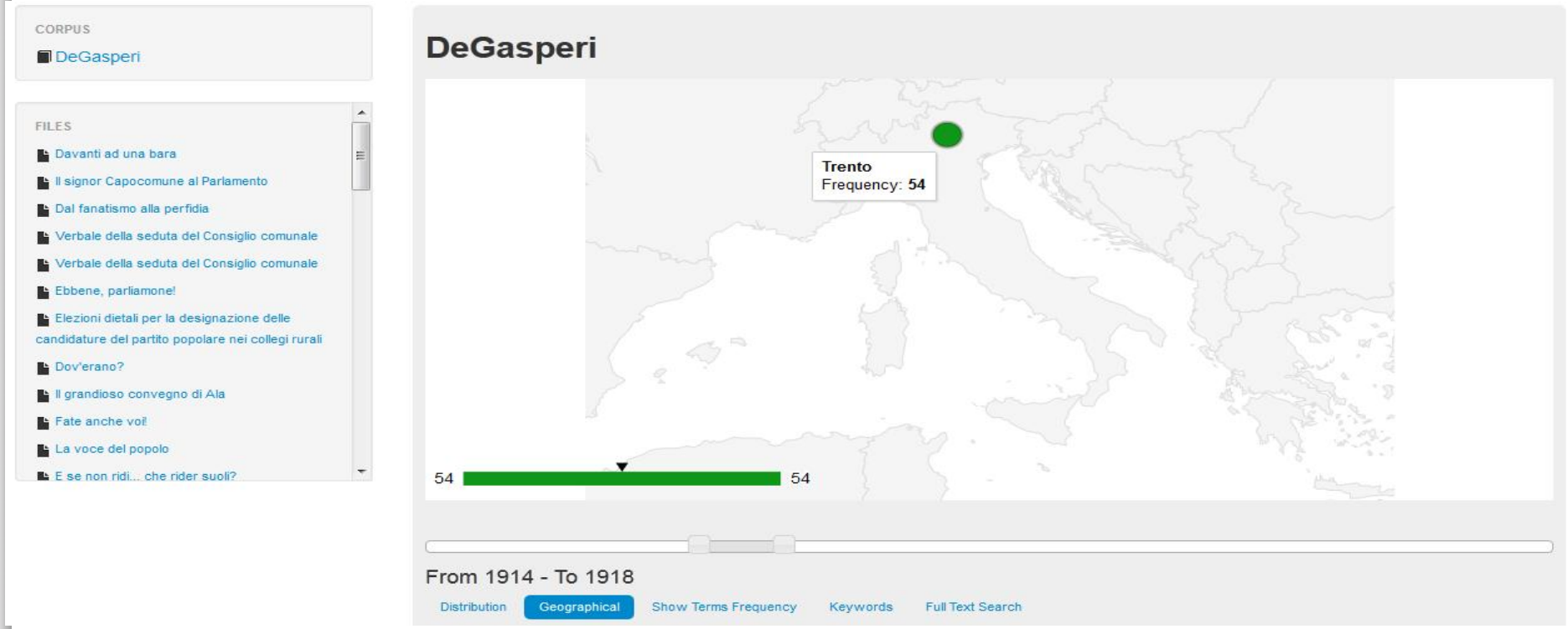
Documents list for the selected year

Time bar to select period of interest

Q: *In which year did De Gasperi publish most?*

ALCIDE: “Distant” reading

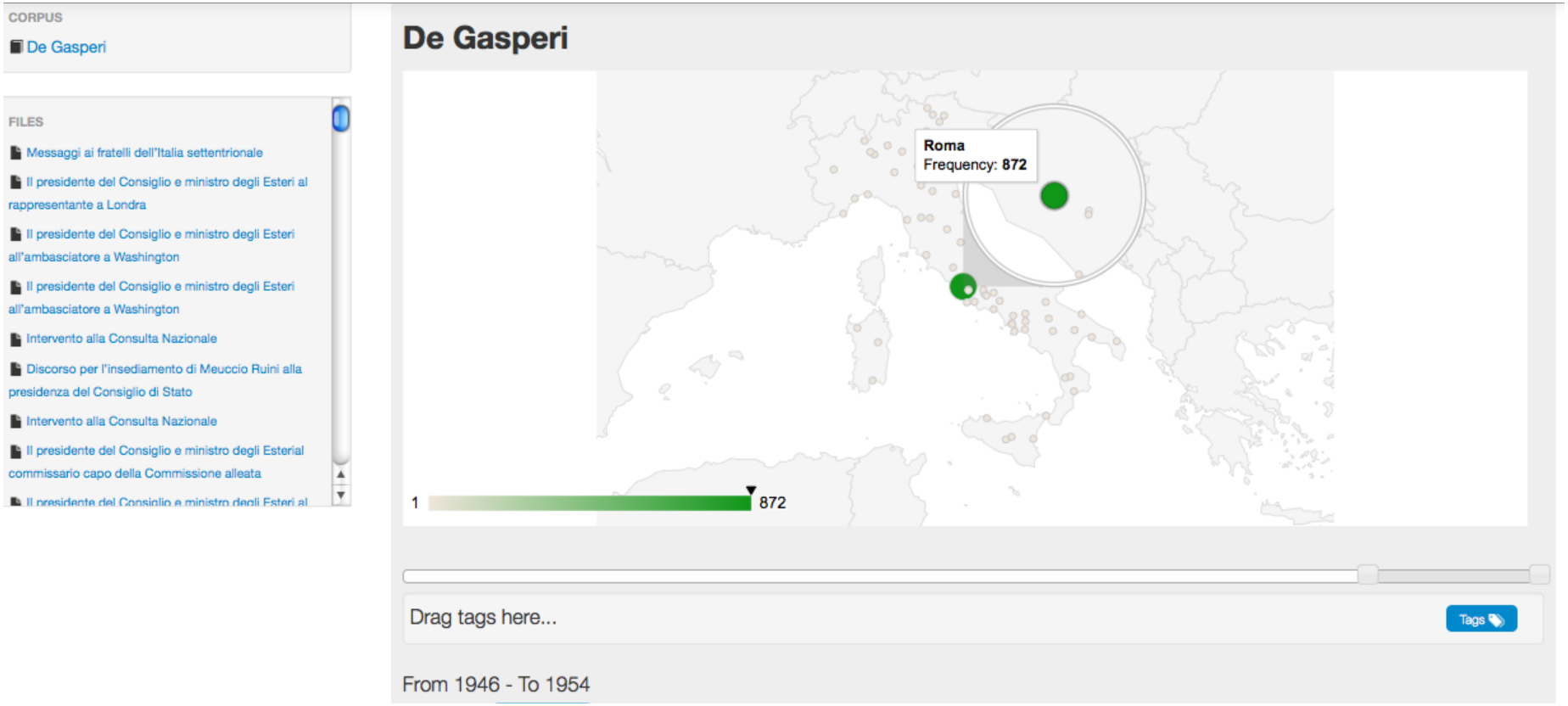
2. Documents distribution per **place** of publication



Q: *In which city were most documents published during World War I?*

ALCIDE: “Distant” reading

2. Documents distribution per **place** of publication

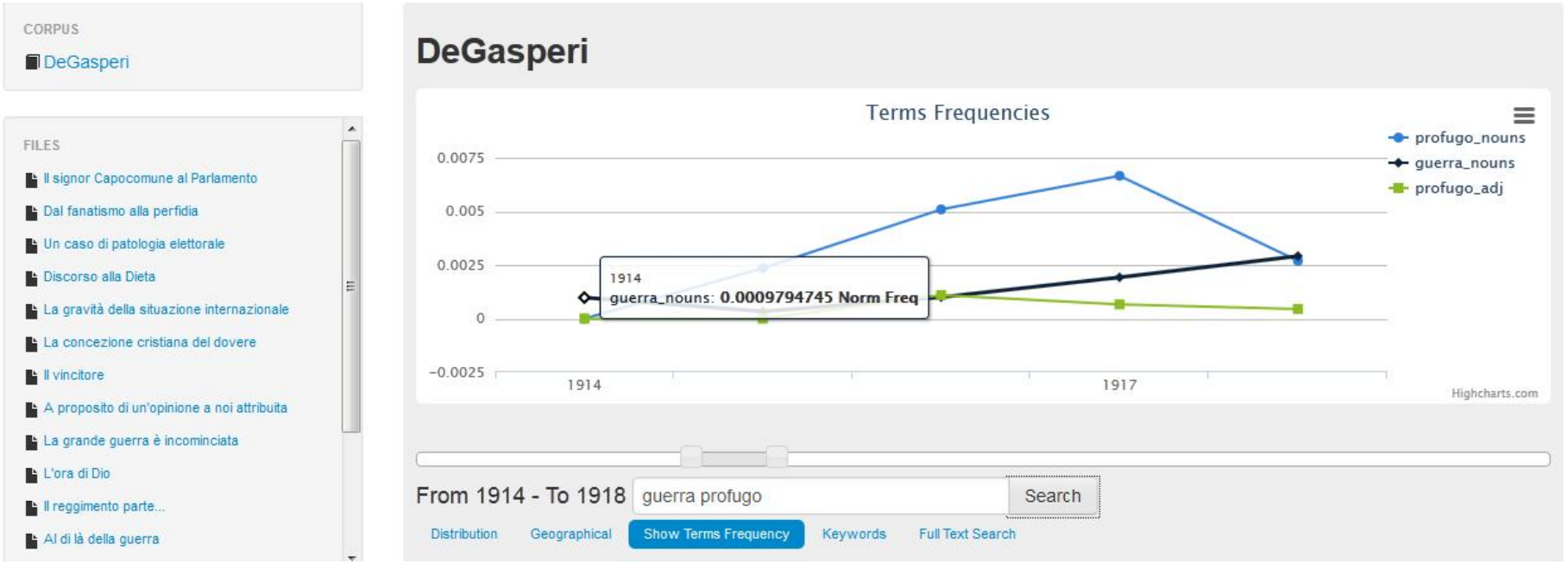


Q: *In which city were most documents published after World War II?*

ALCIDE: “Distant” reading

3. Lemmas & tokens distribution over time

Pre-processing with PoS tagger and lemmatizer from TextPro

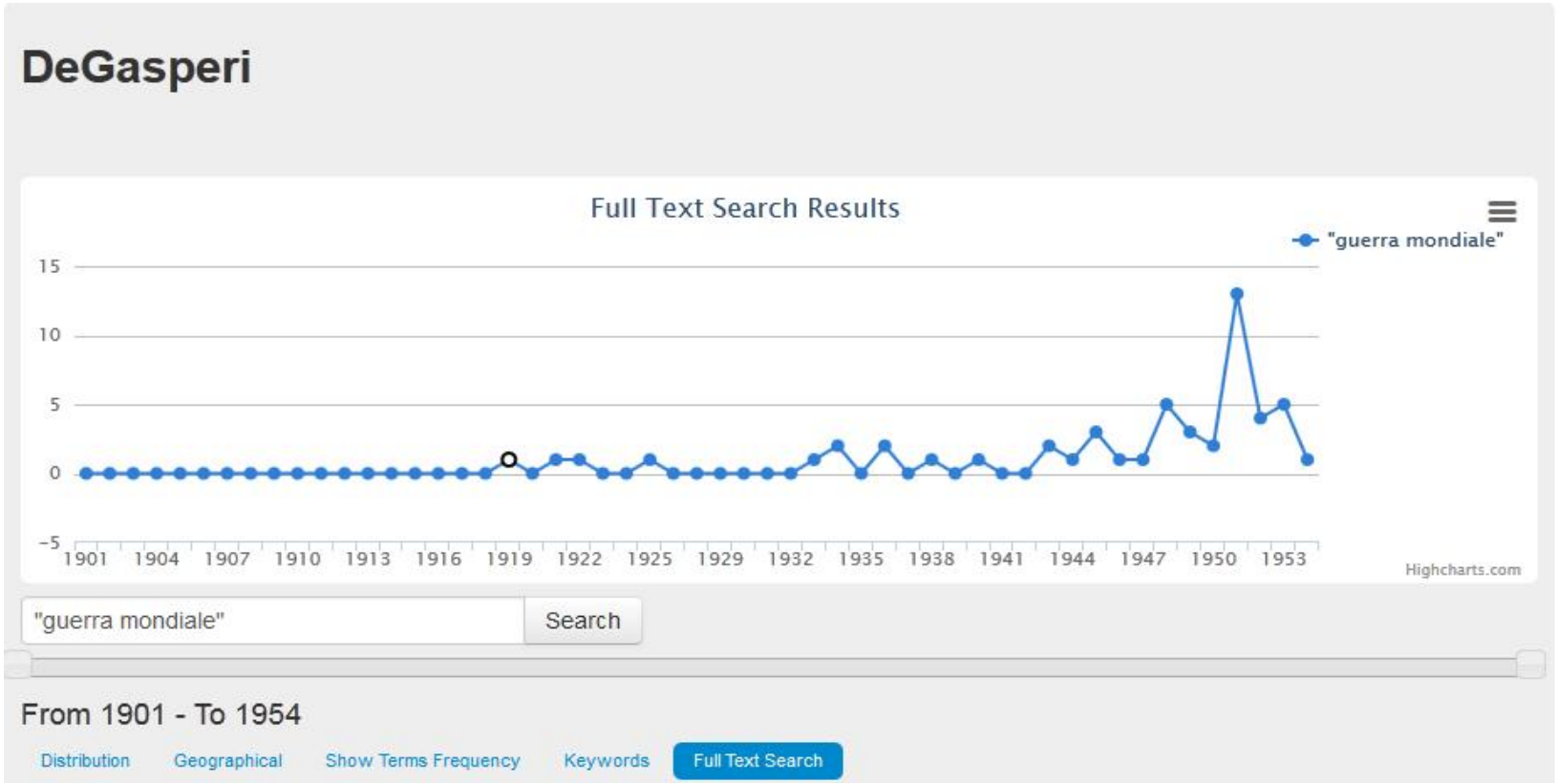


Q: *When did the issue of refugees become relevant for De Gasperi?*

ALCIDE: “Distant” reading

3. Lemmas & tokens distribution over time

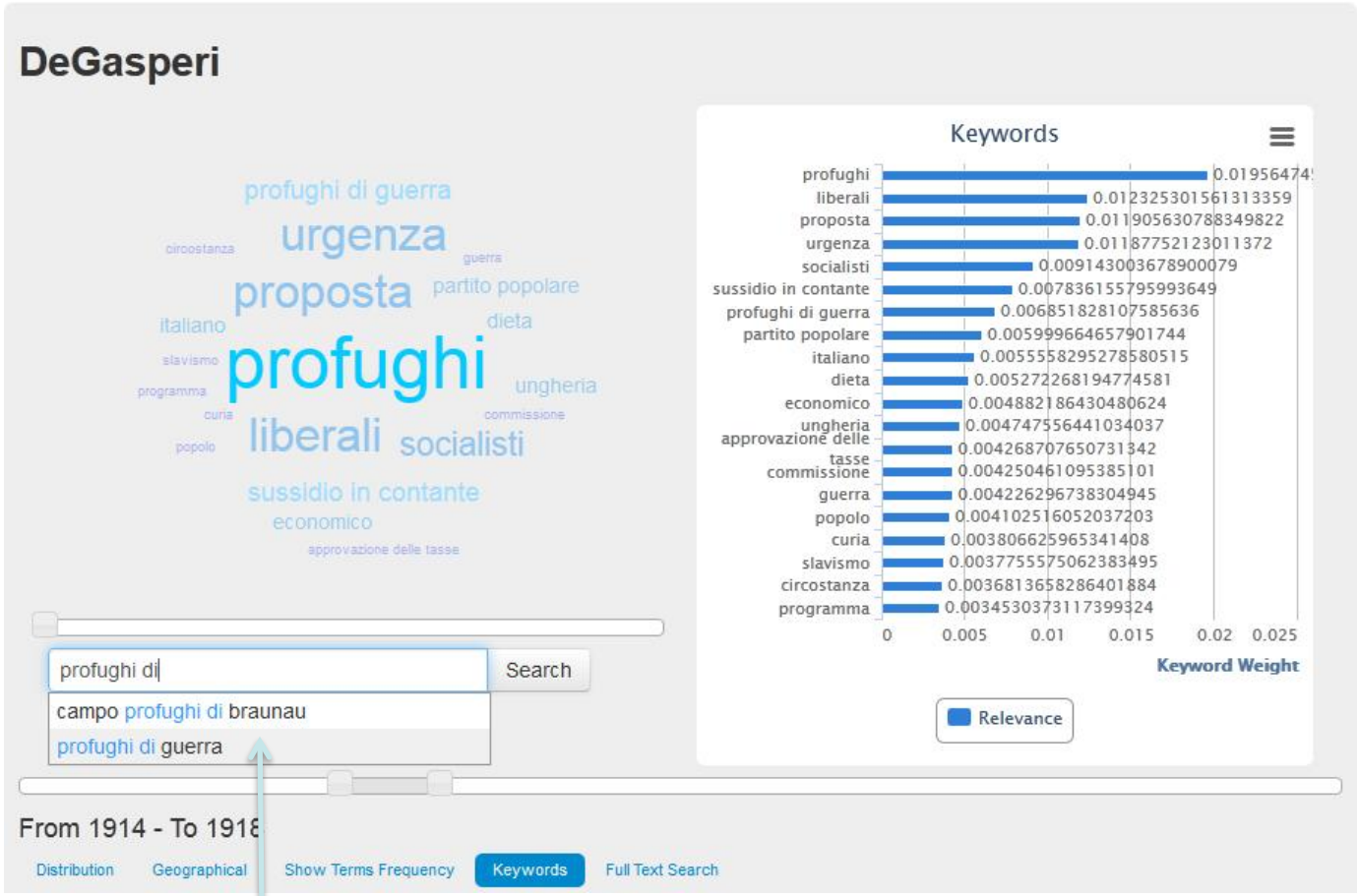
Pre-processing with PoS tagger and lemmatizer from TextPro



Q: *When was “world war” first mentioned by De Gasperi?*

ALCIDE: “Distant” reading

4. Search & visualization of key-concepts



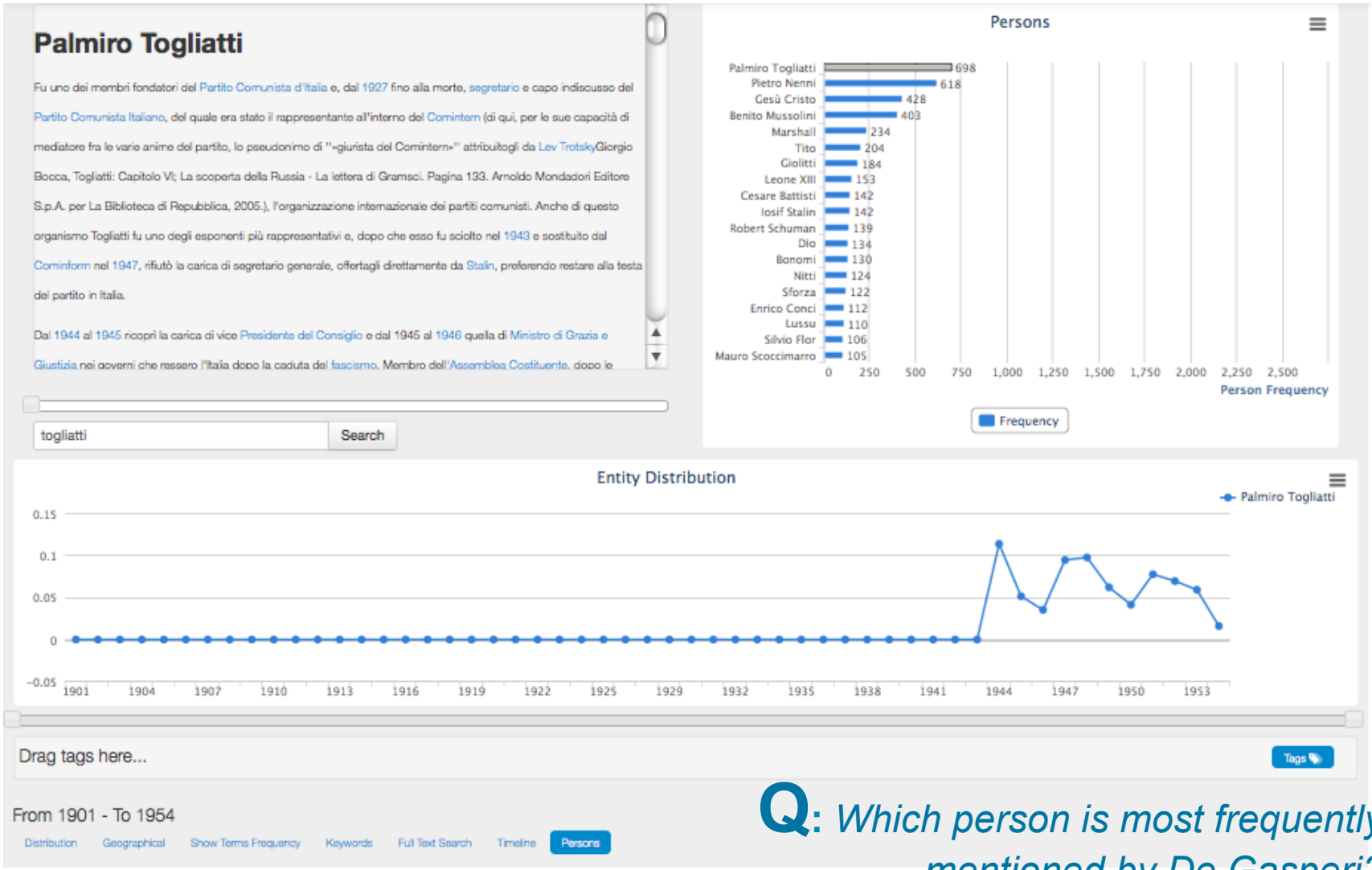
Key-concepts extracted at document level and then merged

Key-concept suggestion

Q: *Which were the main topics dealt with by De Gasperi between 1914 and 1918?*

ALCIDE: “Distant” reading

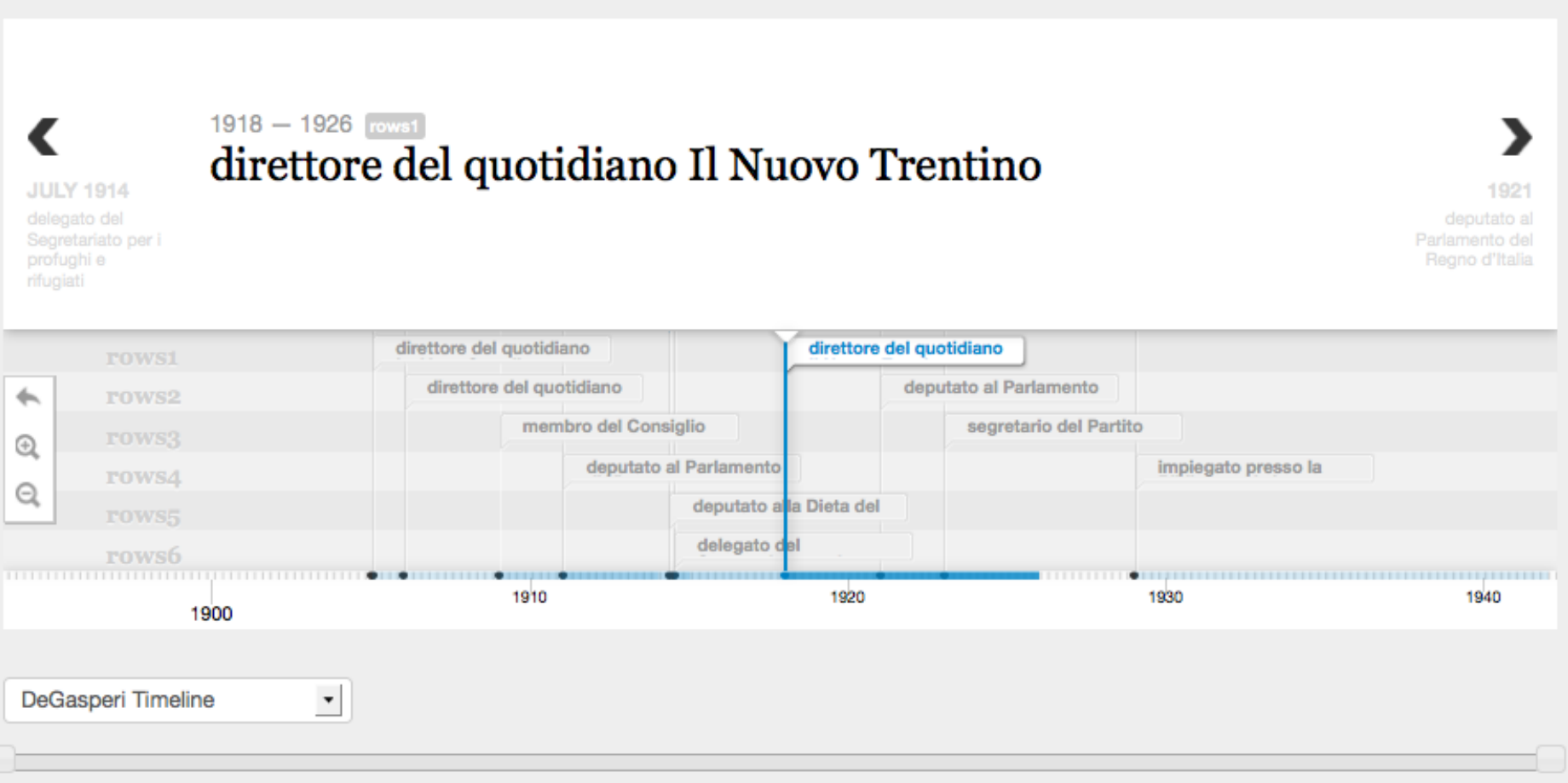
5. Search and visualization of persons’ names



Q: *Which person is most frequently mentioned by De Gasperi?*

ALCIDE: “Distant” reading

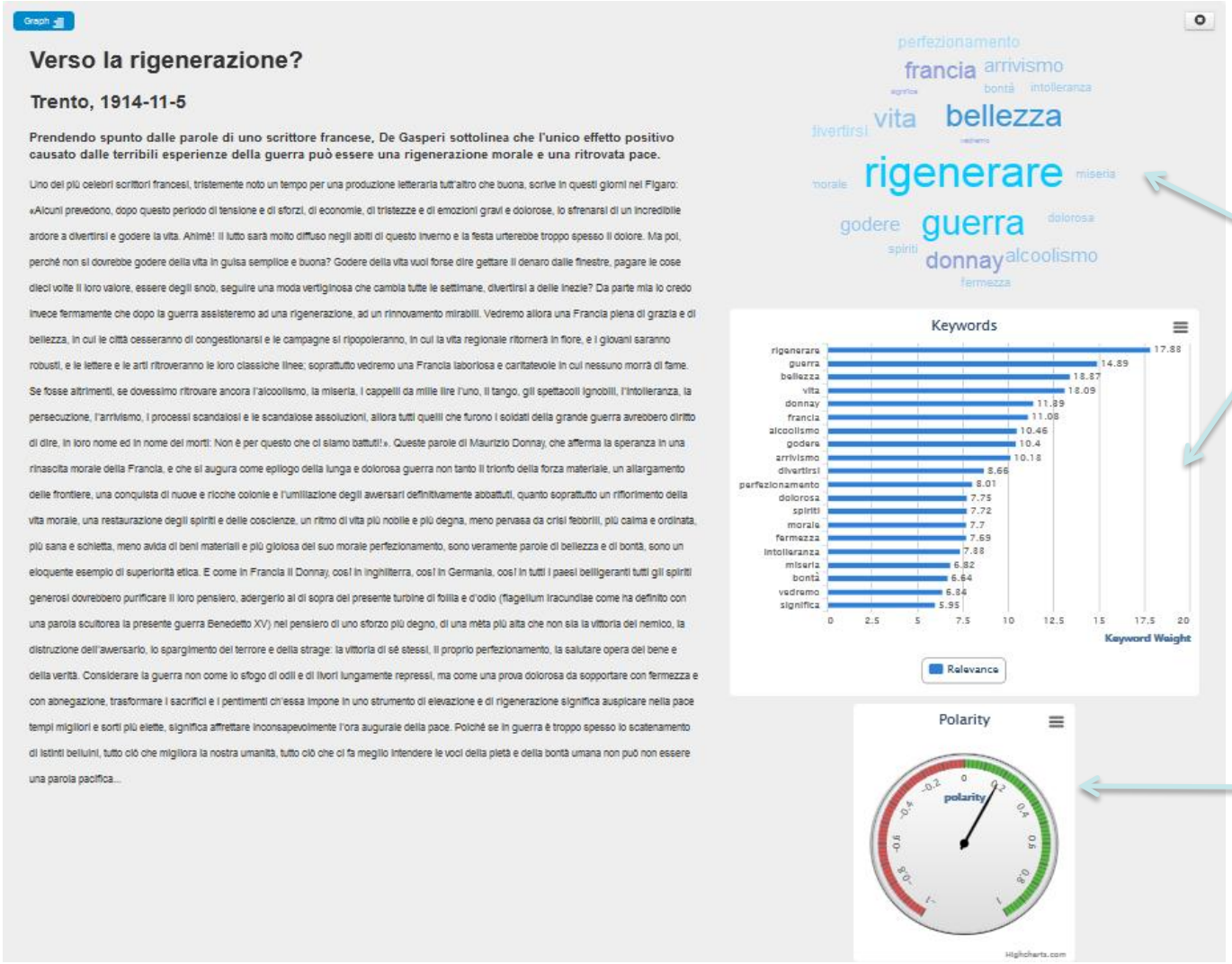
6. Timeline personalization, visualization and comparison



Q: *What speeches did De Gasperi give during the Russian revolution?*

ALCIDE: “Close” reading

7. Visualization at document level

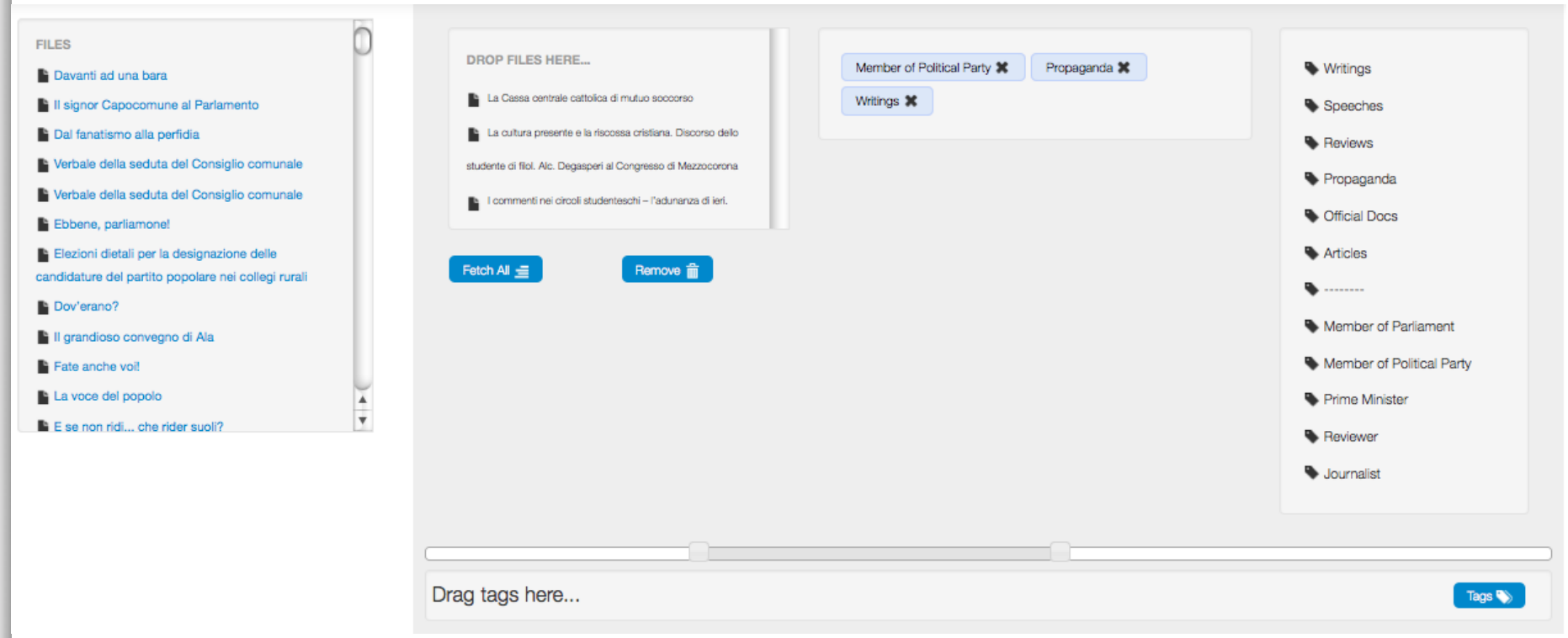


Key-concepts

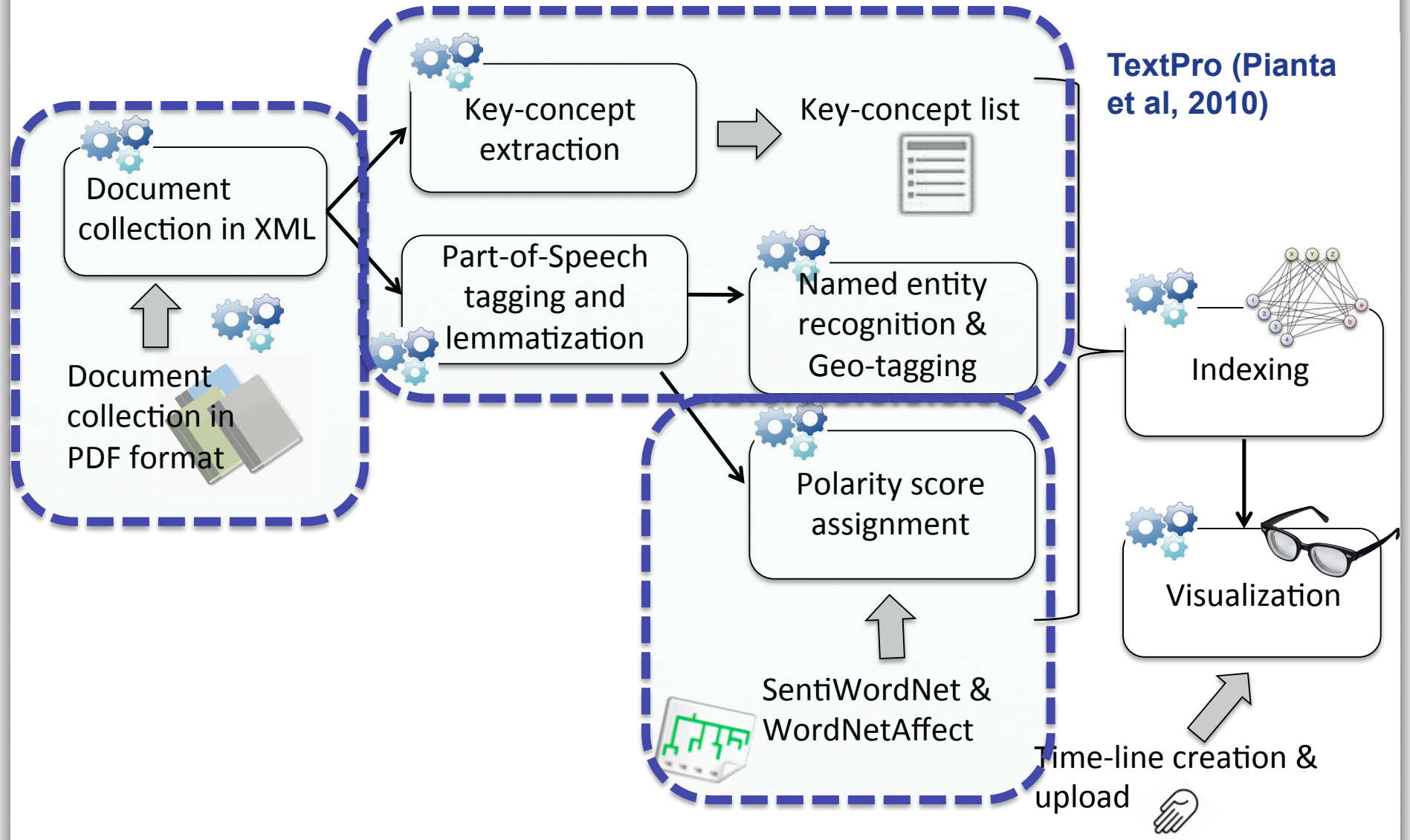
Polarity (positive or negative) at document level

ALCIDE: Tag manager for documents annotation

8. Manual tagging of single or sets of documents



From paper to bits



FONDAZIONE BRUNO KESSLER

Outcome of system evaluation

Focus group with expert users and evaluation based on System Usability Scale questionnaire:

- Some functionalities (*Timeline*, entities linked to *Wikipedia*) should be part of a learners' view
- Revise frequency normalization criteria
- Relevance of key-concepts is too "subjective"
- An automatic analysis of semantic change would be useful, for instance looking at co-occurrences: which association measures?
- Polarity at *concept* level would be more useful than at *document* level

FONDAZIONE BRUNO KESSLER

Polarity annotation: a feasible task?

- How difficult is it to annotate polarity at concept level **on historical data**?
- Manual annotation of "*trade union*" e "*trade unionism*", 525 sentences
- Annotation performed using the *CrowdFlower* platform
- 5 judgements for each occurrence, 4 possible labels: positive, negative, neutral, don't know.
- Gold standard of 60 sentences annotated in-house by two interns
- Compare crowdsourced annotation with gold standard and with automatic polarity labelling via SentiWordNet/ WordNetAffect

FONDAZIONE BRUNO KESSLER

Polarity annotation: a feasible task?

	Crowdsourcing	SentiWNW/ WNAffect
Total	68.30%	43%
Negative	55.50%	22%
Positive	**80%**	31%
Neutral	46.60%	**86%**
Don't know	0	n.a.

Accuracy w.r.t. gold data

Low agreement between annotators, both in-house and via crowdsourcing

FONDAZIONE BRUNO KESSLER

A.L.C.I.D.E. project: Next steps

- Extend the system with the functionalities highlighted during the experts' evaluation
- Tag the corpus to distinguish between transcriptions of speeches, official documents, propaganda writings, etc.: how do vocabulary, discourse structure, style, readability change?
- Add other corpora, also in English
- Use the platform to discover something relevant to history scholars!

FONDAZIONE BRUNO KESSLER

Conclusions

- Many lessons learnt in the first project months:
  - A platform like A.L.C.I.D.E. can be effectively used only if “digital” and “humanities” collaborate from the very beginning, must be a two-way communication
  - NLP technology must be re-thought: make processes more intuitive, include users in the loop, importance of visualization/interfaces
  - We are not just technology providers!

- Many challenges ahead:
  - New project involving MART and MUSEION for exploration platform of verbo-visual art: recommend artworks based on similarity? Crowdsourcing text transcription? How to encode the graphical component?
  - Horizon 2020

THANK YOU!