# EPINetz: Exploration of Political Information Networks

John Ziegler[1] Alexander Brand[2] Julian Freyberg[1] Tim König[2] Wolf Schünemann[2]
Marina Walther[1] Michael Gertz[1]

**Abstract:** Different societal challenges, such as information overload, emerge due to the digital transformation of the media landscape. This also demands new competencies from citizens that often lack the means to contextualize arguments or actors and to understand their interrelationships in complex topics. The EPINetz project is an approach to bridge the outlined skills gap by developing an appropriate political information system. It provides access to political news collected from multiple data sources, including social media, and offers various network exploration capabilities. Different entities such as political actors or topics are extracted from collected data and shown within their respective contexts modelled as weighted and time-varying information networks. Thereby, interested citizens and especially schoolchildren can discover current political topics and understand relationships between relevant entities.

**Keywords:** Digital Literacy; Information Networks; Political Information Systems

## 1 Introduction

How to approach and convey complex topics of social and political importance? How to enable citizens ranging from schoolchildren to adults to develop and improve their media competence and digital literacy? In an increasingly complex media landscape, many citizens lack the appropriate skills.
The EPINetz project aims to develop a Web-based platform that allows users to explore political information networks. Rather than focusing only on a few select media outlets, the platform integrates data from various publicly accessible (German) sources, including Twitter and news outlets. The project has the following objectives: (1) Integrate politically relevant information from different sources and make it accessible through a unified information retrieval system. (2) Extract and visualize entities such as actors (e.g., politicians) and topics, as well as their relationships in a temporal-sensitive way. By applying NLP-based information extraction methods, collected documents are mapped to interactively explorable information networks where nodes represent entities and edges time-varying and weighted relationships. This differs significantly from traditional search engines where users are simply provided with a list of documents barely offering any contextualized view.

[1] Heidelberg University, Institute of Computer Science, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany; ziegler@informatik.uni-heidelberg.de, freyberg@stud.uni-heidelberg.de, m.walther@stud.uni-heidelberg.de, gertz@informatik.uni-heidelberg.de
[2] Hildesheim University, Institute of Social Sciences, Universitätsplatz 1, 31141 Hildesheim, Germany; alexander.brand@uni-hildesheim.de, tim.koenig@uni-hildesheim.de, wolf.schuenemann@uni-hildesheim.de

The remainder of this paper is structured as follows: In Section 2, we give a summary of related projects, contrast them with the EPINetz approach, and outline relevant concepts of digital literacy. This is succeeded by a description of the EPINetz platform in Section 3, along with some sample user stories. Finally, we conclude the paper with a summary and outlook on the project's roadmap in Section 4.

## 2 Background

In this section, we first give an overview of related work, contrast the same with the approach taken in the EPINetz project and, secondly, outline our understanding of digital literacy.

### 2.1 Related Work

To focus on the core functionality of the platform, we narrow down the scope of related work to the topic of political information systems. In the past, different projects have aimed at building similar tools. These approaches differ from EPINetz in several ways. First and foremost, the Media Cloud [MC21] project offers a platform for general-purpose media analysis. Despite its diverse and very sophisticated capabilities, such as topic mapping and source management, it does not offer any of the network-based exploration functionality provided by the EPINetz project. This also applies to the Vox Civitas tool [Di11] which was designed to analyze social media content around broadcast events. Similarly, the European Media Monitor [EMM21] provides basic statistics, trending topics and activity detection to the end-user, but again lacks sufficient exploration functionality. Furthermore, information from social media is not taken into account. This is also true for the TopExNet [SAG19] tool, even though it offers network-based exploration capabilities for entity relationships extracted from news articles. Probably most similar to our work is the LeadLine system [Do12]. It offers a visual analytics system for events extracted from news articles and social media. In contrast, EPINetz focuses on the less general domain of political information and more specifically targets the German media landscape.

### 2.2 Digital Literacy

EPINetz aims to improve both general and domain-specific digital literacy. As to our general conception of digital literacy, there is no canonical terminology to build on, but rather a great variety of concepts being used in the field [DD09]. The concept of literacy is generally oriented towards comprehensive understandings of citizen education and participation in the digital age. First, our concept emphasizes the informational skills that individual users need to develop and apply a critical understanding of their digital information ecosystems and lifeworlds. Thus, we can build on Jones-Kavalier's and Flannigan's definition of digital

literacy as “a person’s ability to perform tasks effectively in a digital environment [...] Literacy includes the ability to read and interpret media (text, sound, images), to reproduce data and images through digital manipulation, and to evaluate and apply new knowledge gained from digital environments.” [JF06] Beyond this fundamental concept, recent developments in digitalization need to be reflected. Therefore, we include skills and evaluative capacities concerning datafication, algorithmic filtering or machine learning that have been termed data literacy elsewhere. [PS20] Beyond processing and presenting information in context, EPINetz allows a ’look behind the scenes’ of datafication by providing opportunities to observe and practice data-scientific methods. All in all, our basic conception is compatible with strategies and frameworks issued by governance actors such as the European Union [Vu16] and the German Conference of Education Ministers [KMK16] as it takes up key components defined in those frameworks, mostly covering the competence area "information and data literacy", but also citizen engagement and a critical understanding of media in a digital world. As to our domain-specific orientation, it is important to add that in contrast to many other studies and projects on politics and policy using computational social science methodology, we do not operate in an exclusively data-driven way. EPINetz instead accounts for the pre-structuration of policy fields that resonates in public debates even when mediated in digital environments.

## 3 EPINetz Platform

In the following, we describe the main components of the EPINetz platform. In Section 3.1, we give a description of the different data sources and outline the data processing in Section 3.2. Finally, in Section 3.3, we detail the construction of the different types of information networks for text analysis and exploration purposes. Potential user stories are given in Section 3.4. Figure 1 serves as an overview of how the different components interact with each other.

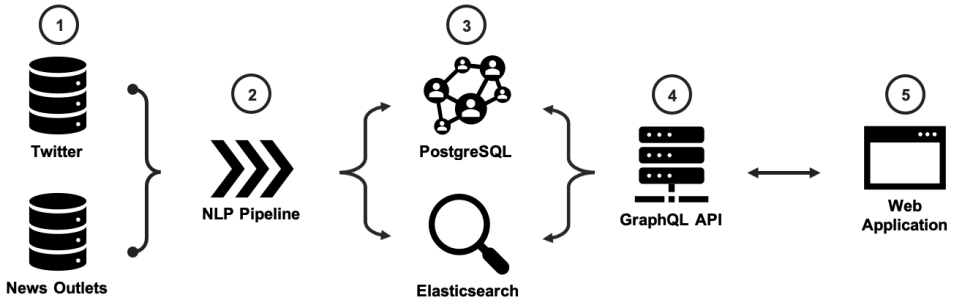



Fig. 1: Simplified overview of the architecture of the EPINetz platform.

1. Data is collected from multiple sources, such as Twitter and news outlets. 2. Textual information of the data is processed by the NLP pipeline. 3. Data is stored in PostgreSQL for network generation and Elasticsearch for information retrieval. 4. A GraphQL API is

used as a unifying data access layer. 5. Information can be explored by the end-user via the Web application.

### 3.1 Data Sources

A primary objective of the EPINetz platform is to provide users access to political topics and actors from different media sources. For this, we resort to social media, more precisely Twitter, as well as close to 100 major German news outlets, e.g., Spiegel, FAZ, Bild. News article collection is restricted to the general topics of politics, economy, finance, and society. Furthermore, the collected Twitter data is based on a curated list of Twitter accounts of close to 2,000 German politicians and political organizations. Twitter data is collected through the new Twitter API[3], providing access to both historic information and live data. Currently, our data pool contains around 12 million tweets and 8 million news articles.
We try to avoid filter bubbles by providing data on parliamentarians and news outlets across the political spectrum. Through network exploration and information retrieval capabilities the user can have a look at multiple and potentially diverse aspects of a political debate. It is important to note that already biased media coverage might be propagated to the platform as we take the raw data without any filtering or sampling.

### 3.2 Data Processing

Both news articles and tweets are processed via dedicated NLP pipelines, primarily using spaCy[4], for named-entity extraction (e.g., persons, organizations, locations). For tweets, hashtags and user mentions are also extracted. By now, we refer to topics as densely connected sets of keywords and actors as occurring in our information networks (see Section 3.3). As illustrated in Figure 1, the processed data is managed via two systems, a PostgreSQL database used as the basis for the construction of our information networks and Elasticsearch for different information retrieval tasks.
To distinguish between different policy fields, e.g., public health and migration, we develop "policy parsing" as a novel method for the identification of such fields. This involves a mixture of supervised and unsupervised learning techniques, namely a pre-informed, enriched keyword-based grouping in the tagged dossiers of the Bundeszentrale für politische Bildung (Federal Agency for Civic Education) and clustering on node-embedded representations on the aforementioned information networks.

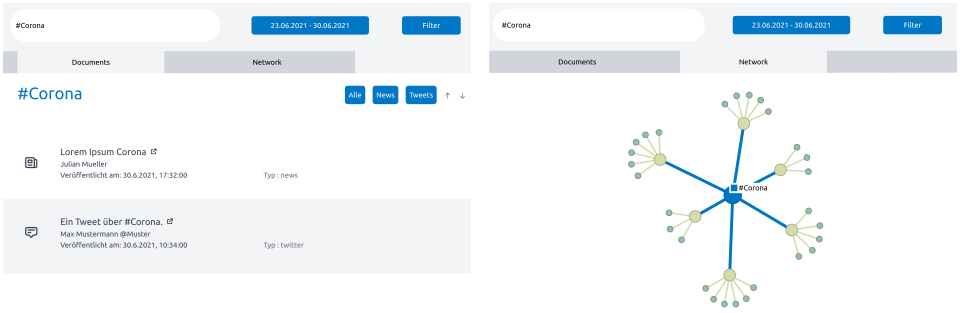
[3] Twitter API v2: Early Access, `https://developer.twitter.com/en/docs/twitter-api/early-access`, retrieved 28-06-2021
[4] spaCy · Industrial-strength Natural Language Processing in Python, `https://spacy.io/`, retrieved 28-06-2021

### 3.3 Information Networks and Timelines

With fine-grained information about text data accessible in PostgreSQL, different types of information networks and timelines can be constructed. Some of these networks are built as soon as new documents come into the system while other networks are constructed on the fly in response to user requests issued via the platform interface. Such networks show, for example, the Twitter users mentioned by a user, the hashtags used by a user, or co-occurrences of hashtags, named entities, and keywords. To reflect the temporal dimension, all relationships are assigned time-encoding attributes, enabling the analysis and contrasting of entities and relationships in a time-sensitive manner.

### 3.4 User Stories

A fictitious user could use the EPINetz platform to keep up with ongoing political debates. By providing an aggregated view based on data from different sources, the end-user will less likely suffer from information overload. Statistical information and timelines of the collected data might serve as potential entry points to currently important topics.
Once the user decided on a subset of entities, network exploration capabilities allow to grasp relationships in an intuitive manner (see Figure 2b). Usage scenarios could then be as simple as exploring topics and arguments over time, or more complex, e.g., contrasting actors and arguments on different media outlets.




(a) Document list (b) Network exploration

Fig. 2: Information can be retrieved as a list of documents or explored as a network.

A prototypical example of an information network based on the hashtag “#Corona” is shown in Figure 2b. Adjacent nodes might be co-occurring hashtags or Twitter users employing them frequently. Figure 2a shows a corresponding fictional list of documents.

## 4 Conclusion and Ongoing Work

With the steady increase of news outlets and channels in the media landscape, it will become more and more difficult for the public and citizens to deal with the amount and complexity of information, especially aspects related to political topics and debates. In the EPINetz project, we aim to address this problem by providing users with a Web-based platform that allows them to search, explore, and contrast information surrounding political actors, topics, and debates at different scales, with data uniformly integrated from diverse sources, including social media. In this paper, we have outlined how network-like structures provide a suitable means to explore actors and topics in context and over time while informing users from where and how the information has been obtained. Underlying EPINetz is a large collection of German news outlets as well as comprehensive Twitter datasets related to German politicians and organisations.

We are currently realizing the first front-end features for the platform, tailored to use cases suitable for a wide range of users, including schoolchildren, to explore the data. In particular, together with partner institutions from the education sector, we design more specific use cases and user stories as well as evaluation and quality criteria, the latter targeted towards schools with a focus on digital literacy in the landscape of political topics and debates.

## Acknowledgements

We thank the Klaus Tschira Stiftung gemeinnützige GmbH for funding EPINetz. The project's website can be found at `www.epinetz.de`.

## References

[DD09] van Deursen, A. J.; van Dijk, J. A.: Using the Internet: Skill related problems in users' online behavior. Interacting with Computers 21/5-6, pp. 393–402, June 2009.

[Di11] Diakopoulos, N.; Naaman, M.; Yazdani, T.; Kivran-Swaine, F.: Social media visual analytics for events. In: Social Media Modeling and Computing. Springer, pp. 189–209, 2011.

[Do12] Dou, W.; Wang, X.; Skau, D.; Ribarsky, W.; Zhou, M. X.: LeadLine: Interactive visual analysis of text data through event identification and exploration. In: 2012 IEEE Conference on Visual Analytics Science and Technology (VAST). Pp. 93–102, 2012.

[EMM21] European Media Monitor, retrieved 23-06-2021, 2021, URL: `https://emm.newsbrief.eu/overview.html`.

[JF06] Jones-Kavalier, B.; Flannigan, S. L.: Connecting the Digital Dots: Literacy of the 21st Century, https://er.educause.edu/articles/2006/1/connecting-the-digital-dots-literacy-of-the-21st-century, retrieved 30-06-2021, 2006.

[KMK16] KMK, Kultusministerkonferenz: Bildung in der digitalen Welt: Strategie der Kultusministerkonferenz, https://www.kmk.org/fileadmin/pdf/PresseUndAktuelles/2018/Digitalstrategie_2017_mit_Weiterbildung.pdf, retrieved 28-07-2021, 2016.

[MC21] Media Cloud, retrieved 23-06-2021, 2021, URL: https://mediacloud.org/.

[PS20] Pangrazio, L.; Sefton-Green, J.: The social utility of 'data literacy'. Learning, Media and Technology 45/2, pp. 208–220, 2020.

[SAG19] Spitz, A.; Almasian, S.; Gertz, M.: TopExNet: Entity-Centric Network Topic Exploration in News Streams. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. WSDM '19, Association for Computing Machinery, Melbourne VIC, Australia, pp. 798–801, 2019.

[Vu16] Vuorikari, R.; Punie, Y.; Carretero, S.; Van den Brande, L.: DigComp 2.0: The Digital Competence Framework for Citizens: Update Phase 1: the Conceptual Reference Model, https://ec.europa.eu/jrc/en/printpdf/150698, retrieved 28-07-2021, 2016.