Journal of Open Humanities Data

A Reproducible IT-Blog Corpus

SPECIAL COLLECTION: 'COMPUTATIONAL HUMANITIES RESEARCH DATA'

DATA PAPER

ubiquity press

ADRIEN BARBARESI

JENS POHLMANN

*Author affiliations can be found in the back matter of this article

ABSTRACT

The dataset comprises text and metadata extracted from several hundred IT-blogs and websites, along with a method to duplicate the data by updating its contents and downloading it to the user's local machine. The targets have been hand-picked with the intention to represent the discourse on blogs and websites dedicated to questions at the intersection of technology and society from Germany and the United States of America. The texts have been retrieved by web crawling techniques. The resulting corpus is accessible through a search platform and also reproducible with freely accessible descriptors and software.

CORRESPONDING AUTHOR:
Jens Pohlmann
Centre for Media, Communication & Information Research (ZeMKI), University of Bremen, Bremen, Germany
jpohlmann@uni-bremen.de

KEYWORDS:
web blogs; corpus linguistics; internet policy; discourse analysis; public discussion; freedom of expression

TO CITE THIS ARTICLE:
Barbaresi, A., & Pohlmann, J. (2021). A Reproducible IT-Blog Corpus. *Journal of Open Humanities Data*, 7: 17, pp. 1–4. DOI: *https://doi.org/10.5334/johd.35*

(1) OVERVIEW

REPOSITORY LOCATION

DWDS-platform (dwds.de): *https://www.dwds.de/d/korpora/it_blogs*

DOI: *10.5281/zenodo.4569734*

CONTEXT

This data was produced for an ongoing research project concerning public discourse about Internet Policy and some of it has been discussed by Barbaresi & Pohlmann (2020).

(2) METHOD

STEPS

We first compiled a list of German IT-blogs and websites by identifying the main websites in this field and then looking for similar sites and keywords. In this process, we were looking for blogs and websites that report about and discuss the latest developments in information technology and products, IT-law, and policy, as well as sites dedicated to commentary on the societal impact of technology. Aside from specific IT-news and policy portals, such as *www.netzpolitik.org*, and IT-product portals (e.g., *www.mobilegeeks.de*), we collected blogs by IT-lawyers and those of scholars, intellectuals, and journalists working in the fields of Communication and Media Studies, Tech, Law, Policy, and Philosophy. After finding these initial sites, we manually extended the list by handpicking recommendations from the engine *https://www.similarsites.com* that fit the corpus profile.

We fetched sitemaps from the sites of interest (the sitemaps protocol primarily allows webmasters to inform search engines about pages on their sites that are available for crawling) and identified content for the remaining sites by web crawling (Olston & Najork 2010).

Text extraction focuses on the central part of the texts (e.g., without navigation or footer information), comments (potential user-generated content listed at the bottom of an article), and metadata (at least title, date, and URL and possibly author, tags, categories, and summary). The documents are then stored as XML and processed by a platform for lexicographic corpus research (Geyken et al. 2017).

For more information see *https://trafilatura.readthedocs.io/*. The software is published under an open-source license: *https://doi.org/10.5281/zenodo.3460969*.

SAMPLING STRATEGY

All pages found on the websites have been processed in full provided it was technically possible, and the resulting documents contained a meaningful amount of text (e.g., no image galleries). Additionally, metadata, text, and comments have been extracted and indexed on the platform, which allows for further text-based filtering using faceted searches.

QUALITY CONTROL

The corpus has been checked for consistency and completeness (especially concerning the accuracy of the scraping process). Furthermore, relevant metadata on the website level have been checked as well (e.g., copyright licenses).

(3) DATASET DESCRIPTION

OBJECT NAME

IT-blog corpus, EN+DE.

FORMAT NAMES AND VERSIONS

TXT, Python package, CSV and XML data export.

CREATION DATES

2019-09-03, update pending.

DATASET CREATORS

Adrien Barbaresi (Berlin-Brandenburg Academy of Sciences), Jens Pohlmann (ZeMKI, University of Bremen; CESTA, Stanford University).

LANGUAGE

English and German.

LICENSE

Access restricted to free login; software under GPLv3+ license; list of sources under CC BY-SA license v4.0.

REPOSITORY NAME

DWDS-platform (dwds.de)

PUBLICATION DATE

2020-09-15.

SIZE

Hundreds of different sources, German version: 1.5 million documents, amounting to 900 million tokens, English version: 2 million documents and about 1.3 billion tokens.

(4) REUSE POTENTIAL

Quotes extracted from the corpus can be used in a variety of formats. The whole dataset cannot be copied and re-used as such. However, the corpus can be re-created from the sources list using the open-source corpus building software Trafilatura (*https://github.com/adbar/trafilatura*), thus making it free to copy while bypassing potential copyright concerns. The data is of interest to other corpus linguists, political scientists, sociologists, cultural studies researchers, but also for market studies or technology impact assessment.

In our work, the data is used to analyse the discourse about Internet Policy questions in Germany and the United States. IT-blogs represent an expert discourse regarding questions at the intersection of technology and society and may have a considerable impact on the discussion of these matters in traditional media (e.g., newspapers), and thereby on the conversation in broader swaths of the population (Barbaresi & Pohlmann 2020). In a pilot study, we compare the discussion about a particular German anti-hate speech law, the Netzwerkdurchsetzungsgesetz (NetzDG), on German IT-blogs with the discourse that is simultaneously taking place in the most important German newspapers. Based on this setup, we can draw conclusions about the impact of IT-blogs and websites on more traditional print media.

Separate from analyses that are predicated on themes and specified search terms, users can also apply more corpus-driven text data mining techniques and examine, for example, the contents of complete blogs or groups of blogs and websites in order to determine topics that prevail in these posts over time. Furthermore, they can inspect reference networks through exploring linkages between specific blogs/websites and thereby study communication practices within the IT-blog sphere.

The corpus needs to be re-created from the sources list to avoid potential copyright concerns, as some of the blog posts and website articles contained within the corpus are copyright protected. However, this does not hinder the application of text data mining techniques to the data when it comes to producing results that are either highly aggregated and do not allow for the reconstruction of the original texts, or regarding results that only provide snippets of the texts in question.

Note, however, that this stipulation may generate difficulties when it comes to sharing the underlying dataset of a study with peer-reviewers and the research community at large. The dataset can be freely reproduced by using Trafilatura to create a specific subcorpus. However, its composition may change after the initial corpus has been produced if web pages or whole websites are deleted. Recovery can be automatically attempted from the Internet Archive (*https://archive.org*). However, reviewers or anyone who wants to rebuild the initial corpus may end up with a slightly different text base, especially if there is a substantial amount of time between the creation of the respective corpora. Consequently, the verifiability of individual research results may be limited, particularly if the research in question mainly draws on distant reading practices and statistical analysis of large amounts of text and metadata. For projects that follow a "blended reading" (Stulpe & Lemke 2016) approach and integrate elements of close reading, these limitations may be less troubling. A comprehensive discussion of this accessibility issue exceeds the scope of the short data paper format.

ACKNOWLEDGEMENTS

We would like to thank Lukas and Yannick Kozmus (BBAW), Victoria Kratel (University of Bremen), and Uma Phatak (Stanford University) for their work on quality assurance and the collection of relevant metadata.

FUNDING STATEMENT

The DWDS project is part of the *Akademienprogramm* of the German Academies of Sciences and of the Center for Digital Lexicography of German (ZDL).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Adrien Barbaresi: Conceptualization, Data curation, Methodology, Software, Supervision, Writing – original draft, Writing – review & editing.

Jens Pohlmann: Conceptualization, Data curation, Supervision, Writing – original draft, Writing – review & editing.

AUTHOR AFFILIATIONS

Adrien Barbaresi *https://orcid.org/0000-0002-8079-8694*
Center for Digital Lexicography of German, BBAW, Berlin, Germany

Jens Pohlmann *https://orcid.org/0000-0001-6614-5358*
Centre for Media, Communication & Information Research (ZeMKI), University of Bremen, Bremen, Germany; Center for Spatial and Textual Analysis (CESTA), Stanford University, Stanford, USA

REFERENCES

Barbaresi, A., & **Pohlmann, J.** (2020). Mapping the German Tech Blog Sphere and Its Influence on Digital Policy. In S. Breidenbach, P. Klimczak & C. Petersen (Eds.), *Soziale Medien* (pp. 139–157). Wiesbaden: Springer Fachmedien. DOI: *https://doi.org/10.1007/978-3-658-30702-8_7*

Geyken, A., Barbaresi, A., Didakowski, J., Jurish, B., Wiegand, F., & **Lemnitzer, L.** (2017). Die Korpusplattform des "Digitalen Wörterbuchs der deutschen Sprache" (DWDS). *Zeitschrift für germanistische Linguistik*, *45*(2), 327–344. DOI: *https://doi.org/10.1515/zgl-2017-0017*

Olston, C., & **Najork, M.** (2010). Web crawling. *Foundations and Trends in Information Retrieval*, 4(3), 175–246. DOI: *https://doi.org/10.1561/1500000017*

Stulpe, A., & **Lemke, M.** (2016). Blended Reading. In M. Lemke & G. Wiedemann (Eds.), *Text Mining in den Sozialwissenschaften* (pp. 17–61). Wiesbaden: Springer Fachmedien. DOI: *https://doi.org/10.1007/978-3-658-07224-7_2*

TO CITE THIS ARTICLE:
Barbaresi, A., & Pohlmann, J. (2021). A Reproducible IT-Blog Corpus. *Journal of Open Humanities Data*, 7: 17, pp. 1–4. DOI: *https://doi.org/10.5334/johd.35*

Published: 22 July 2021

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.