

# Extracting literature references in German Speaking Geography – the GEOcite project

Bastian Birkeneder<sup>1</sup>, Philipp Aufenvenne<sup>1</sup>, Christian Haase<sup>1</sup>, Philipp Mayr<sup>2</sup> and Malte Steinbrink<sup>1</sup>

<sup>1</sup>Chair of Human Geography, University of Passau, Germany

<sup>2</sup>GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany

## Abstract

The paper outlines the motivation and build-up of the DFG-funded GEOcite project at University of Passau. The project works on a domain-specific approach to automatically extract, segment, match and visualize literature references in the German speaking geography domain with the objective to provide a novel basis for a scientometric monitoring instrument for the community. In this paper, we describe the GEOcite corpus, its construction and elaborate on a preliminary evaluation of different approaches to extract and segment references from the digitized part of the corpus. We further evaluate the EXCITE segmentation model [1] on different datasets of German research papers. The results of our evaluation show small improvements with domain-specific and increased training data.

## Keywords

Reference extraction, Geography papers, Network analytics, Scientometric monitoring,

## 1. Introduction

The GEOcite project presented in this paper is a central part of the overarching research project "The Pillars of Unity and Disciplinary Bridges: Geographical Research between Rhetoric and Practice", which has been funded by the German Research Foundation (DFG) since 2013. The main focus of the project is the question of the unity of geography. This question is nearly as old as the discipline itself. While human geography sees itself as social and cultural science, physical geography is assigned to the natural sciences. Not only in German speaking Geography the relationship between physical and human geography has always been a matter of concern deeply interwoven with the discipline's identity [2, 3]. While the idea of bringing together the natural and the social sciences is claimed as the discipline's unique selling point, there is a growing awareness of centrifugal tendencies within geography threatening its integrity and cohesion as one academic discipline [4]. So far, these discussions about the disciplines unity lack empirical support. Therefore, the project aims to provide an empirical basis by using bibliometric and network analytic methods. Based on an analysis of publication and

---


*ULITE workshop at JCDL 2022*

✉ Bastian.Birkeneder@uni-passau.de (B. Birkeneder); Philipp.Aufenvenne@uni-passau.de (P. Aufenvenne); Christian.Haase@uni-passau.de (C. Haase); philipp.mayr@gesis.org (P. Mayr); Malte.Steinbrink@uni-passau.de (M. Steinbrink)

🆔 0000-0002-2460-4920 (B. Birkeneder); 0000-0001-7957-5752 (P. Aufenvenne); 0000-0002-6656-1658 (P. Mayr); 0000-0001-7503-2750 (M. Steinbrink)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

citation patterns, the disciplinary structure of German speaking geography is investigated. In the first phase of the project (2014-2018), the citation relationships of all geographers who held a professorship at a German, Austrian or Swiss university in 2012 were collected. The citation data was retrieved from journal papers published by these actors in the decade from 2003 to 2012. The data collection was carried out partly automated using Scopus. Reference data from geographical journals not listed in Scopus were manually extracted. The results of the first project phase show that the discipline has clearly split into different subdisciplinary clusters [5, 6, 2]. However, the subdisciplines are still more or less linked by citations. Though, the temporal dimension of the structuring process could not be taken into account. So it is not clear whether the current situation is the result of growing together or drifting apart. Therefore, in the second project phase (since 2019), the data basis was comprehensively expanded in order to enable longitudinal analyses focusing on disciplinary dynamics over time. Our aim is to include all journal publications of geography professors from German speaking countries from 1949 until today. For this purpose the scientometric monitoring tool *GEOcite* was developed. In the following, the structure and functionality of *GEOcite* will be explained.

This paper matches with a couple of focus topics at the ULITE workshop<sup>1</sup>: *GEOcite* completely builds on open source software and has the objective to produce "Open infrastructures and services for reference mining"; in addition, *GEOcite* is an application of an established software framework for reference extraction and matching *EXCITE* [7] in the Geography domain. Thirdly, *GEOcite* matches with the topic "Search, exploration and mining of the reference graph" in the way that the retrieved data will ultimately be used for network analysis aiming at a deeper understanding of historical changes and paradigmatic shifts within geography.

## 2. *GEOcite*: technical background

As a software solution, *GEOcite* aims to locate, collect and process extensive historical and recent citation data (from 1949 to the present). Digital and analogue archives are used for this. Extensive digitization work is being carried out, which forms the basis for an automated citation data extraction. For this task the *EXCITE* tools are used (see Figure 1). Our aim is to create a database for the analysis of current structures of disciplinary knowledge networks and their historical genesis and development. *GEOcite* enables us to create the conditions for bibliometric network analysis to better understand the disciplinary dynamics. The data obtained are made available to the scientific community and is thus permanently available for empirical research and historical discipline observation.

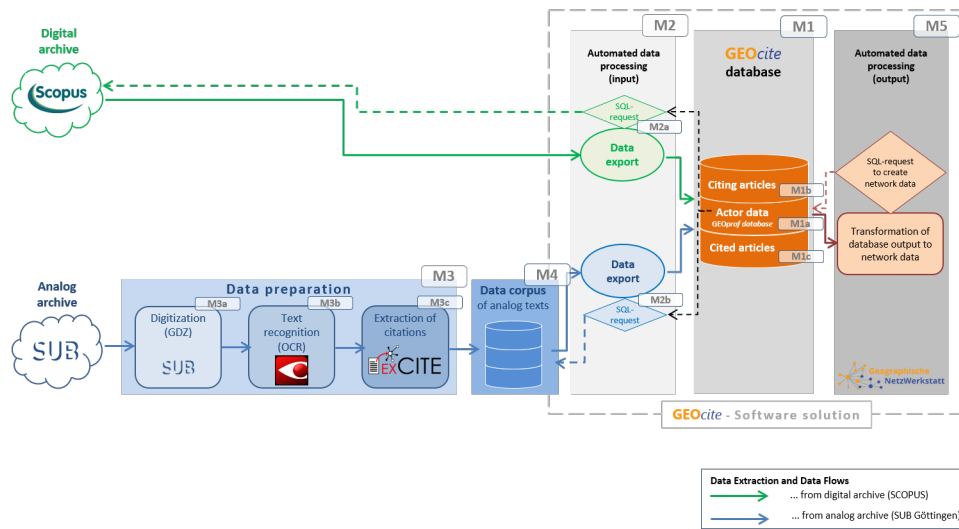
Figure 1 gives an overview of the structure and the data flows of the *GEOcite* tool. At the center is the *GEOcite* database [M1]. This links three datasets [M1a, b, c] necessary for the planned bibliometric network analyses.

In *GEOcite*, as in the first phase of the project, the actors considered are also the geographic professors in German speaking countries. The *GEOprof*-Database [actor data, M1a] contains a list of all geography professors since 1949 as well as additional biographical attribute data [8].

The second dataset [M1b] is a comprehensive compilation of the bibliographic information of journal papers published by those geography professors listed in the *GEOprof*-Database (citing

---

<sup>1</sup><https://exciteproject.github.io/ULITE-ws/>



**Figure 1:** GEOcite database and data flow

articles). To cover both historical and current publication activity, the database feeds from two different sources: In addition to texts listed in Scopus, the data is taken from analog journals that were digitized by the Göttingen Digitization Center (GDZ). Due to copyright legislation only the title pages and the bibliographies of a paper were scanned. The copies were provided as single files in TIFF format and needed to be processed in several steps: First, automated text recognition (OCR) was performed. Errors detected during text recognition were corrected manually. The image files were then converted into PDF format. After that the documents were merged to combine the title pages of an article and the associated bibliography into one file. The open source programs Cermin [9] and Grobid [10] were used to extract bibliographic data from the title pages, specifically authors and title of an article.

The third component depicted [M1c], represents a list of all works cited in the articles. In addition to the complete bibliographic information of the references, the dataset also contains the link to the actor data (GEOprof database) [M1a] and the source texts [M1b] in which they were cited. While this bibliographic information can be queried directly in Scopus, extracting the information of the cited works from the digitized corpus is more challenging. This is precisely the application field of the EXCITE project, which has been running since 2016 and is funded by the DFG [7]. EXCITE provides a tool for extracting literature references from PDF files. For this purpose, the reference strings in PDF documents are automatically detected, extracted and segmented. EXCITE was developed specifically for the extraction of citation data from social science texts and has been trained on mainly recent German-language papers. Relevant extracted bibliographic information is used to find matches between our actor database and processed scientific articles. The resulting links between citing actor (matched author of

an article) and cited actor (matched author in a reference) is then used to create our citation network.

GEOcite reuses the following EXCITE tools<sup>2</sup> [7]:

1. EXannotator<sup>3</sup> to build a dataset for Exparser model training.
2. Exparser [1] to process and extract the references from the PDF corpus.

## 2.1. GEOcite Data

In the following Table 1, we outline the GEOcite corpus consisting of active male and female professors of Geography in Germany and other German speaking countries. In addition, we list the amount of considered relevant papers in Scopus and our digitized article corpus. We divide our data into bins of 20 years (1949–1968; 1969–1988; 1989–2008; 2009–2022).

	1949-68	1969-88	1989-2008	2009-2022	total
Active professors <sup>1</sup>	180	565	759	567	1,180
Papers in Scopus <sup>2</sup>	231	3,149	10,762	15,263	29,484 <sup>3</sup>
Papers in digitized corpus <sup>4</sup>	3,352	4,119	5,758	2,823	16,052

<sup>1</sup> Included are those professors who were actively holding a professorship in the respective time interval.

<sup>2</sup> The Scopus corpus includes only scientific papers written by relevant actors.

<sup>3</sup> 23 articles from Scopus do not include a publication date.

<sup>4</sup> The digitalized corpus also contains articles from authors, which are not part of our research group.

**Table 1**

Overview of the GEOcite corpus.

## 2.2. GEOcite tools

As the first output of the GEOcite project, a comprehensive list of the geographic professors in Germany, Austria and Switzerland since 1949 is available. The GEOprof dataset is available to the research community as a static download (CSV format) at the geoscience data publisher [8]. There you will also find further information about the methods used to collect the data on the professorship. In addition, an interactive map to explore the dataset is available on the project's website<sup>4</sup> (see Figure 2).

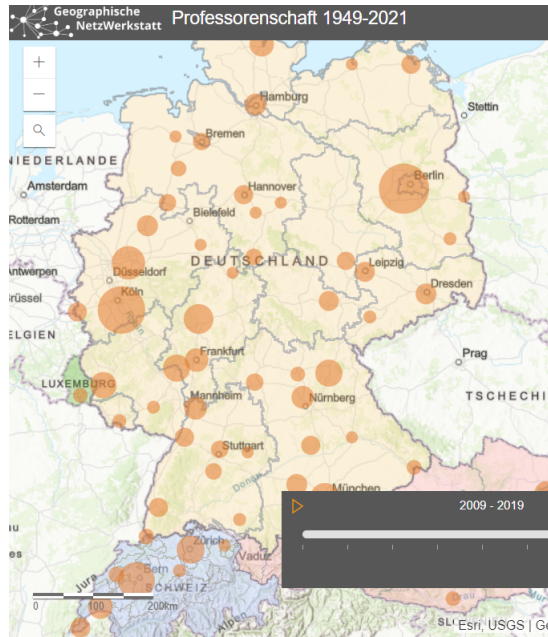
We further created a geography-specific dataset of annotated and segmented references, extracted from scientific articles. At the end of the project, a web-based platform for bibliometric network analysis of the collected geographic citation data is planned. All datasets and tools are or will be available for reuse<sup>5</sup>.

<sup>2</sup><https://github.com/exciteproject/>

<sup>3</sup><https://github.com/exciteproject/EXannotator>

<sup>4</sup><https://geographische-netzwerkstatt.uni-passau.de/de/geoprof/>

<sup>5</sup><https://github.com/GeoCite>



**Figure 2:** GEOprof demonstrator with active Professors in Geography from 1949–2021, see interactive map<sup>6</sup>

In the following section, we describe a preliminary evaluation and discussion of an analysis of reference segmentation in our Geocite corpus, as well as the next steps to be taken in the project in order to optimize the segmentation process.

### 3. Evaluation of the reference segmentation

#### 3.1. Set-up and Training

In order to create a comprehensive citation network between our actors, it is essential to extract important bibliographic data with preferably low error rates. While the default Exparser models provide sufficient results, the used toolchain allows its user to train a model on a custom dataset. To maximize the results of the segmentation of our extracted references, we trained three models on different datasets to examine the effects of more domain specific articles and more articles in general. To increase our training data for the machine learning (ML) models used by Exparser, we extracted references from 170 German geography research papers. These articles were randomly chosen from our digital corpus and were published between 1952 and 2019. The EXCITE dataset<sup>7</sup> contains 125 German articles. We further annotated and segmented these references according to the specified EXCITE requirements [11]. As test set we combined 10% of our dataset and 10% of the EXCITE German Goldstandard. Training parameters were set identical as reported by Hosseini et al. [7]. We trained one model with our data (GEOcite model) and the EXCITE Goldstandard (EXCITE model) respectively, as well as one model with both

<sup>7</sup><https://github.com/exciteproject/EXgoldstandard>

training sets combined (Combined model).

### 3.2. Results

The results of all three models are shown in Table 2.

Label	F1 GEOcite model	F1 EXCITE model	F1 Combined model
publisher	0.82	0.85	<b>0.87</b>
last page	0.93	<b>0.94</b>	<b>0.94</b>
surname	0.75	0.80	<b>0.84</b>
article-title	0.90	<b>0.91</b>	0.90
url	<b>0.89</b>	0.88	0.87
volume	<b>0.89</b>	0.82	0.86
source	<b>0.83</b>	0.81	0.82
given-names	0.82	0.85	<b>0.86</b>
editor	<b>0.81</b>	0.80	<b>0.81</b>
first page	0.94	<b>0.95</b>	<b>0.95</b>
year	0.86	0.90	<b>0.93</b>
identifier	0.73	0.75	<b>0.76</b>
issue	0.77	0.79	<b>0.80</b>
other	0.75	<b>0.78</b>	0.77

**Table 2**

Results of three models, trained on different datasets. The F1-score is used as evaluation metric.

Our evaluation shows that a domain-specific dataset does not necessarily improve the output of the Exparsar segmentation model. In particular we can observe that important tags for our work, like surname, given-names, and article-title achieve lower F1-scores than the EXCITE model. For several tags, we notice slightly improved results with the combined model. Our results indicate that more general data and training data from our target domain can improve the Exparsar segmentation model. Considering the predominating use of English language in the scientific community, it might be no surprise that the majority of datasets in this domain were collected from English publications. This unfortunately limits the usage of large scale datasets like PMC Open Access [12] or DocBank [13].

One subject of our future research is the utilization of more sophisticated ML models. In recent years a paradigm shift for ML can be observed [14]. Models like BERT [15] or GPT-3 [16] trained on broad data at scale and used as foundation models show exceeding results in NLP or Computer Vision tasks. We experiment with multilingual models (e.g. XLM-R [17]) for text features and instance segmentation models (e.g. Mask R-CNN [18]) for structural features. The underlying idea is to use these available large scale datasets for language independent models and circumvent the sparsity of data in different languages.

## 4. Outlook

With the completion of the project, we will release a dataset of our citation network, as well as all extracted citations from our corpus. Additionally, we will provide a REST API for all members

of the scientific community to query data from different actors, corresponding citations, and various other attributes.

Similar to our *GEOprof* dataset, we will provide an interactive website where our citation network and research results are visualized. Furthermore our software platform *GEOcite* will be entirely Open Source.

Initial empirical analyses based on the *GEOcite* corpus are also already planned. For example, there will be further investigations on the question of the unity of geography (see above). In addition, work is planned on paradigm genesis and evolution in German speaking geography as well as specific bibliometric studies on the disadvantage of female geographers in the sense of the so called Matilda effect [19, 20] in the course of the discipline's history. As another example, self-citation behavior [21] of this special community covered in the *Geocite* corpus can be analysed over the covered period.

## Acknowledgments

This work was funded by DFG under grant 249237273, **Die Säulen der Einheit und die Brücken im Fach: Geographische Forschung zwischen Rhetorik und Praxis (*GEOcite*)** project, <https://geographische-netzwerkstatt.uni-passau.de/geocite/>.

## References

- [1] Z. Boukhers, S. Ambhore, S. Staab, An end-to-end approach for extracting and segmenting high-variance references from pdf documents, in: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries 2019, 2019, pp. 186–195. doi:10.1109/JCDL.2019.00035.
- [2] P. Aufenvenne, M. Steinbrink, Brüche und Brücken: Netzwerk- und zitationsanalytische Beobachtungen zur Einheit der Geographie, *Geographie und Landeskunde* 88 (2014) 257–292.
- [3] C. Kesteloot, L. Bagnoli, Human and physical geography: Can we learn something from the history of their relations?, *BELGEO* (2021). doi:10.4000/belgeo.52627.
- [4] D. Demeritt, Dictionaries, disciplines and the future of geography, *Geoforum* 39 (2008) 1811–1813. doi:10.1016/j.geoforum.2008.09.008.
- [5] M. Steinbrink, P. Aufenvenne, Integrative Geographiedidaktik? Versuch einer Positionsbestimmung der Fachdidaktik innerhalb der deutschsprachigen Geographie 142/143 (2016). URL: [http://hw.oeaw.ac.at/?arp=7887-0inhalt/gwu142-143\\_03\\_Steinbrink-Aufenvenne.pdf](http://hw.oeaw.ac.at/?arp=7887-0inhalt/gwu142-143_03_Steinbrink-Aufenvenne.pdf). doi:10.1553/gw-unterricht142/143s5.
- [6] M. Steinbrink, P. Aufenvenne, On othering and mainstreamisation of new cultural geography. some scientometric observations, *Mitteilungen der Österreichischen Geographischen Gesellschaft* 159 (2017) 83–104. doi:10.1553/moegg159s83.
- [7] A. Hosseini, B. Ghavimi, Z. Boukhers, P. Mayr, EXCITE - A toolchain to extract, match and publish open literature references, in: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries 2019, ACM, 2019, pp. 432–433. doi:10.1109/JCDL.2019.00105.
- [8] M. Steinbrink, P. Aufenvenne, M. Köhler, B. Birkeneder, *GEOprof-Database: Datenbank*

- der geographischen ProfessorInnenschaft im deutschsprachigen Raum ab 1949, 2021. doi:10.5880/FIDGEO.2021.018.
- [9] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek, Ł. Bolikowski, CERMINE: automatic extraction of structured metadata from scientific literature, *Int. J. Doc. Anal. Recognit.* 18 (2015) 317–335.
- [10] Grobid, <https://github.com/kermitt2/grobid>, 2008–2022. arXiv:1:dir:dab86b296e3c3216e2241968f0d63b68e8209d3c.
- [11] Excite documentation, <https://exparser.readthedocs.io/en/latest/ReferenceParsing/>, 2019. [Online; accessed 1-May-2022].
- [12] Pmc open access subset [internet]. Bethesda (md): National library of medicine, <https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/>, 2003. [Online; accessed 1-May-2022].
- [13] M. Li, Y. Xu, L. Cui, S. Huang, F. Wei, Z. Li, M. Zhou, Docbank: A benchmark dataset for document layout analysis, 2020. arXiv:2006.01038.
- [14] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., On the opportunities and risks of foundation models, arXiv preprint arXiv:2108.07258 (2021).
- [15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
- [16] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, arXiv preprint arXiv:2005.14165 (2020).
- [17] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale (2019). arXiv:1911.02116.
- [18] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.
- [19] M. W. Rossiter, The matthew matilda effect in science, *Social Studies of Science* 23 (1993) 325 – 341. doi:10.1177/030631293023002004.
- [20] P. Aufvenne, C. Haase, F. Meixner, M. Steinbrink, Participation and communication behaviour at academic conferences – an empirical gender study at the german congress of geography 2019, *Geoforum* 126 (2021) 192–204. URL: <https://www.sciencedirect.com/science/article/pii/S0016718521001986>. doi:<https://doi.org/10.1016/j.geoforum.2021.07.002>.
- [21] A. Kacem, J. W. Flatt, P. Mayr, Tracking self-citations in academic publishing, *Scientometrics* 123 (2020) 1157–1165. doi:10.1007/s11192-020-03413-9.

## A. Online Resources

The sources for GEOcite project will be available via

- GitHub.