https://doi.org/10.51408/issi2025_207

Open Citations in German Educational Research–Identifying Disciplinary Practices to Train Data Extraction

Verena Weimer[1], Tamara Heck[2], Christoph Schindler[3]

[1]*v.weimer@dipf.de*, [2]*t.heck@dipf.de*, [3]*c.schindler@dipf.de*
DIPF | Leibniz Institute for Research and Information in Education; Rostocker Straße 6, 60323 Frankfurt am Main (Germany)

Introduction

Citations are an important element of scientific communication, as they transparently show relationships between scientific publications, research data and their authors within the scientific community. Citation data is used in bibliometric and scientometric studies as evidence of internal scientific communication for the self-reflection of a discipline, for the evaluation and control of research performance and for research management (van Raan, 2019; Ball, 2020). In the past, the collection, processing and provision of scientific citation data was in the hands of a few commercial providers, such as Clarivate (Web of Science) and Elsevier (Scopus). Their use is based on licenses, which results in two major problems: Firstly, the commercial citation databases are subject to a fee and are not openly accessible. Secondly, those citation databases do not cover all disciplines to the same extent. As a result, these citation databases are only suitable for searching for literature and evaluating research to a very limited extent. This applies above all to the social sciences and humanities, which include almost all disciplines doing research about education, such as educational research, psychology, economics, and sociology (Moed, 2005; Singleton et al., 2015). Studies also show that reference lists in those databases are missing or are insufficient (Martín-Martín et al., 2018; Visser, van Eck, Waltman, 2021; Chi, 2014). In summary, educational research lacks exhaustive and high-quality citation data to improve literature search and disciplinary bibliometric studies.

Current research projects and network activities aim to contribute to open and networked citation data in science (Backes et al., 2024). Two examples of such approaches are the Initiative for Open Citations (I4OC) and OpenAlex. Our project Open Citation Data for Educational Research (OFFZIB) aligns with those initiatives and aims to extract citation data from open access publications in educational research and make them available via the central national German Education Index (FIS Bildung) (Botte, 2017). This meets the need for a more optimized literature search in the form of a semantic research graph in the database (Hocker et al., 2019) and at the same time offers the possibility of more detailed citation analyses in educational research. To reach this goal, we need to adapt an extraction algorithm to best perform with educational literature data and to establish new workflows to maintain the provision of the extracted data when the project has ended. To develop this extraction algorithm, knowledge must first be gained about how German education researchers cite, specifically in-text citations (Burbules, 2014). The specific research question is: Which citation styles (including special cases) exist in German educational research and are and there sub-disciplinary and document type-based differences?

Method

To investigate this question, a dataset was developed that represents the educational science publication landscape in Germany. The sample considers the different sub-disciplines of German educational research as well as the document types (data collection) and is coded regarding generally valid citation styles (coding).

Data Collection

The dataset shall represent the educational research publication landscape in Germany and thus is based on publications in the largest

disciplinary national open access repository peDOCS (Schindler & Butz, 2023). We aim to analyse at least 1% of the database peDOCS (~ 25,000 documents), thus determining a dataset of 400 documents. In the dataset, the ratio between the sub-disciplines (e.g. developmental psychology, educational sociology) and the existing three document types articles, books and collections (e.g. proceedings) are balanced according to the overall ratio in peDOCS. In addition, it was considered to ensure that the ratio of older and more recent publications as well as German and English documents in the peDOCS database is reflected.

Coding

The citation practices applied in the 400 documents are coded and analysed regarding common and standardised citation styles (e.g. APA citation style), but above all also with regard to styles specific for educational research. For example, special cases that cannot be assigned to a standardised citation style are citations of legal texts, which are then coded as an individual style. The documentation of the styles will be provided in an interoperable format to enable others to compare and reuse the collection for their own citation extraction.

Discussion

The citation practices of educational research are presented, compared and discussed against the background of other disciplines. Similarities and differences are highlighted. The result of the analysis is a comprehensive presentation of citation styles in educational research in Germany and their special formats. Furthermore, the results are discussed regarding challenges for citation extraction.

Outlook

Building on the results, the OFFZIB project will train the OUTCITE algorithm (Hosseini et al., 2019; Backes et al., 2024) to extract citations from educational open access publications. To make an active contribution to the development of a transdisciplinary and transnational citation inventory beyond the specific subject communities of educational research, the citation data will be given to the Open Citations Initiative. Therefore, a maintainable workflow will be established, which will also consider the workflows of the 30 partner institutes, which index and provide the literature for the German Education Index.

References

Backes, T., Iurshina, A., Shahid, M. A., & Mayr, P. (2024). Comparing Free Reference Extraction Pipelines. *International Journal on Digital Libraries*, 25(4), pp. 841–853. https://doi.org/10.1007/s00799-024-00404-6

Ball, R. (Hrsg.). (2020). *Handbook Bibliometrics*. De Gruyter Saur. https://doi.org/10.1515/9783110646610

Botte, A. (2017). 25 Jahre Fachinformationssystem (FIS) Bildung – eine einzigartige Kooperation. *Bibliotheksdienst* 51(8), pp. 651–663. https://doi.org/10.1515/bd-2017-0071

Burbules, N.C. (2014). The Paradigmatic Differences Between Name/Date and Footnote Styles of Citation. In: P. Smeyers, M. Depaepe (Eds.), *Educational Research: Material Culture and Its Representation*. Educational Research, vol 8. Springer, Cham. https://doi.org/10.1007/978-3-319-03083-8_13

Chi, P.-S. (2014). Which role do non-source items play in the social sciences? A case study in political science in Germany. *Scientometrics*, 101(2), pp. 1195–1213. https://doi.org/10.1007/s11192-014-1433-1

Hocker, J.; Veja, C., Schindler, C.; Rittberger, M., (2019). Establishing semantic research graphs in humanities' research practice. In: C. Draude, M. Lange & B. Sick (Eds.), *INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft* (Workshop-Beiträge). Bonn: Gesellschaft für Informatik e.V. (pp. 169-174). https://doi.org/10.18420/inf2019_ws18

Hosseini, A., Ghavimi, B., Boukhers, Z., & Mayr, P. (2019). EXCITE - A toolchain to extract, match and publish open literature references. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries 2019*, pp. 432–433. https://doi.org/10.1109/JCDL.2019.00105

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, 12(4), pp. 1160-1177. https://doi.org/10.1016/j.joi.2018.09.002

Moed, H. F. (2005). *Citation analysis in research evaluation*. Information Science and Knowledge Management: Bd. 9. Springer. https://doi.org/10.1007/1-4020-3714-7

Schindler, C. & Butz, A. (2023). peDOCS - ein Fachrepositorium in der Bildungsforschung mit Kooperationsnetzwerk für Open Access. In H. Ertl & B. Rödel (Eds.), *Offene Zusammenhänge: Open Access in der Berufsbildungsforschung* (Berichte zur beruflichen Bildung, pp. 236-242). Bonn: Bundesinstitut für Berufsbildung. URL: https://www.bibb.de/dienst/publikationen/de/18249

Singleton, K., Kuhberg-Lasson, V., Sondergeld, U., & Schultheiß, J. (2015). Publikationen der Bildungsforschung. In A. Botte, U. Sondergeld, & M. Rittberger (Eds.), *Monitoring Bildungsforschung: Befunde aus dem Forschungsprojekt "Entwicklung und Veränderungsdynamik eines heterogenen sozialwissenschaftlichen Feldes am Beispiel der Bildungsforschung"* (pp. 69–106). Klinkhardt.

Van Raan, A. (2019). Measuring Science: Basic Principles and Application of Advanced Bibliometrics. In: W. Glänzel, H.F. Moed, U. Schmoch & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators*. Switzerland: Springer.

Visser, M., van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, 2(1), pp. 20–41. https://doi.org/10.1162/qss_a_00112