FREIE UNIVERSITÄT BERLIN

GEOWISSENSCHAFTLICHE BIBLIOTHEK

Google Scholar
Strengths and Limitations

Graphic: DALL-E

Coffee lecture, winter semester 2024/2025

Published at: https://www.geo.fu-berlin.de/bibliotheken/Aktuelles/Coffee-Lectures-2024-25-WiSe.html
Presentation based on a coffee lecture of Pierre-Emmanuel Roy (2024)

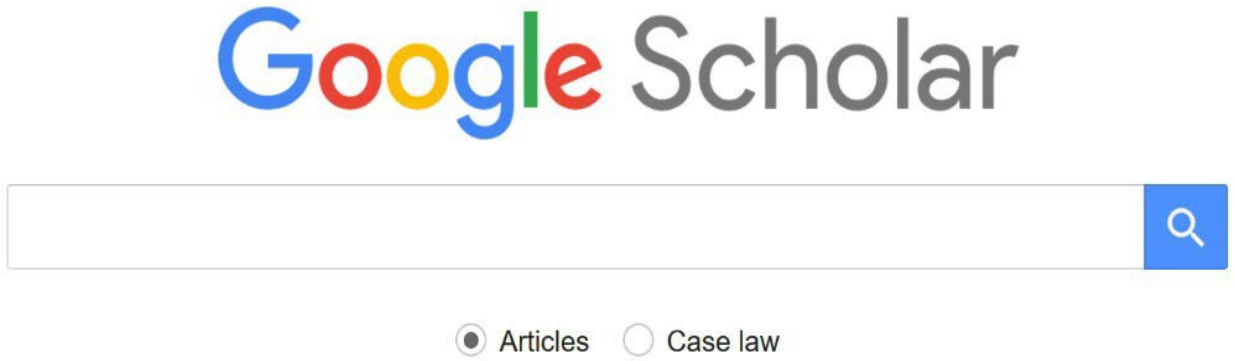
What is Google Scholar?

- Crawler-based search engine for scholarly publications
  - ➢ Sources: academic publisher websites, online journals, university repositories, scholarly websites, etc.
  - ➢ File types: PDF and HTML
  - ➢ Indexed content: articles, books, theses and dissertations, reports, posters, PowerPoint presentations, etc.
  - ➢ What you (probably) won't find: journals (only articles are indexed!), datasets, book reviews, news sections, editorials, announcements…

Source: DALL-E

Why it is so popular?

- Wide coverage: around 389 million records (Gusenbauer, 2019)
  - ProQuest: ca 280 M; Scopus: more than 90 M; Web of Science Core Collection: ca 96 M
  - Take these numbers with a grain of salt!
  - Google Scholar's coverage is not perfect!
- Simple and familiar interface, similar to Google Search



Limitations

1. Noise
2. Bad metadata
3. Limited search options
4. Inconsistent search results

Noise

[HTML] **Groundwater** recharge: an overview of processes and challenges

JJ De Vries, I Simmers - Hydrogeology Journal, 2002 - Springer

… is greatest in those areas – **groundwater** is often the only water … **groundwater** recharge in temperate and humid zones, because recharge is normally included in regional **groundwater** …

☆ Save 99 Cite Cited by 1157 Related articles All 17 versions Web of Science: 565 »

Class Common Name Species

M **Chimpanzee**, M Gorilla - SciELO Brasil

Mammalia Macaque Macaca mulatta NM_001266321. 1a NM_001033021. 1a NM_001032918. 1a NM_001266091. 1a Mammalia Baboon Papio anubis XM_009200345 …

☆ Save 99 Cite Related articles »

FOR THE FAMILY

L **Potatoes**, V Salad, B Cheesecake, CC Cookie - cal, 1977 - staging.thesimplegreek.com

Protein — and — Sauce Pita — or — Bowl Toppings — and — Finishes Page 1 1 2 3

Protein cal. 190-410 — and — Sauce cal. 10-70 Pita cal. 190-200 — or — Bowl cal. 10-280 …

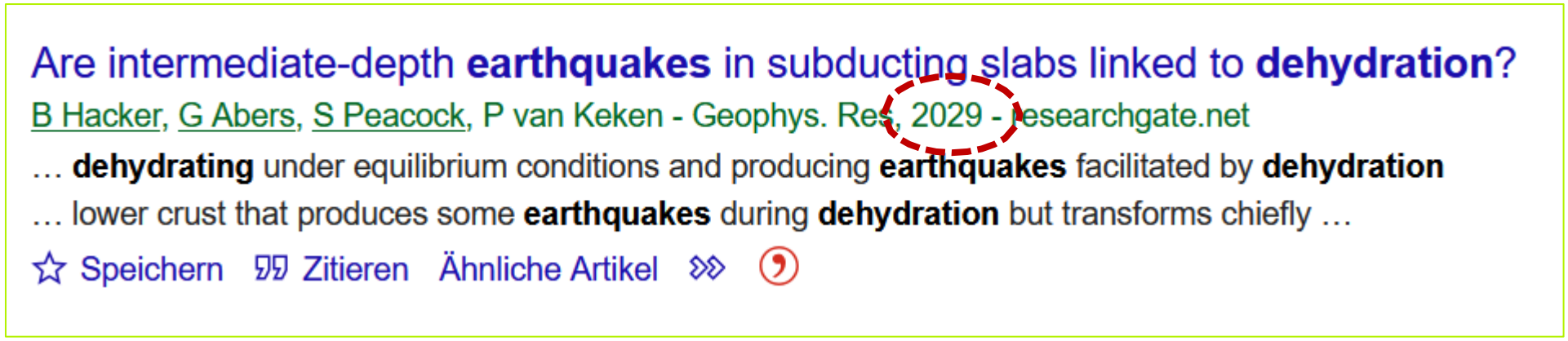
☆ Save 99 Cite Cited by 2 Related articles All 4 versions »

Noise

- Irrelevant records
- Non-peer-reviewed publications
- Obsolete versions
- Duplicates — ca 5% in Google Scholar, almost 0% in Web of Science (Gusenbauer, 2022)

→ **Noise makes it harder to find relevant articles!**

Bad metadata



Bad metadata

Rard II: The 94 million related-article recommendation dataset
J Beel, B Smyth, A Collins - arXiv preprint arXiv:1807.06918, 2018 - arxiv.org
The main contribution of this paper is to introduce and describe a new recommender-systems dataset (RARD II). It is based on data from Mr. DLib, a recommender-system as-a ...
☆ Speichern 99 Zitieren Zitiert von: 12 Ähnliche Artikel »

[PDF] Example Document
J Doe, A Bow - ACM Transactions on Interactive Intelligent Systems ..., 2016 - ceur-ws.org
The main contribution of this paper is to introduce and describe a new recommender-systems dataset (RARD II). It is based on data from Mr. DLib, a recommender-system as-a ...
99 Zitieren

RARD II: The 94 Million Related-Article Recommendation Dataset
J Beel, B Smyth, A Collins - arXiv e-prints, 2018 - ui.adsabs.harvard.edu
The main contribution of this paper is to introduce and describe a new recommender-systems dataset (RARD II). It is based on data from Mr. DLib, a recommender-system as-a ...
99 Zitieren

[PDF] Example Document
J Doe, A Bow - ACM Transactions on Interactive Intelligent ..., 2016 - amir-workshop.org
The main contribution of this paper is to introduce and describe a new recommender-systems dataset (RARD II). It is based on data from Mr. DLib, a recommender-system as-a ...
99 Zitieren

Google Scholar parsed the entire PDF and believed the meta-data of the current file was a mix of strings occurring
a) in the middle of the document in a vector graphic (title and author) and
b) in one of the entries in the reference list (Journal and Year).

The (sometimes) weird indexing practice of Google Scholar
Joeran Beel (2019), https://isg.beel.org/blog/2019/08/19/the-sometimes-weird-indexing-practice-of-google-scholar/

Bad metadata

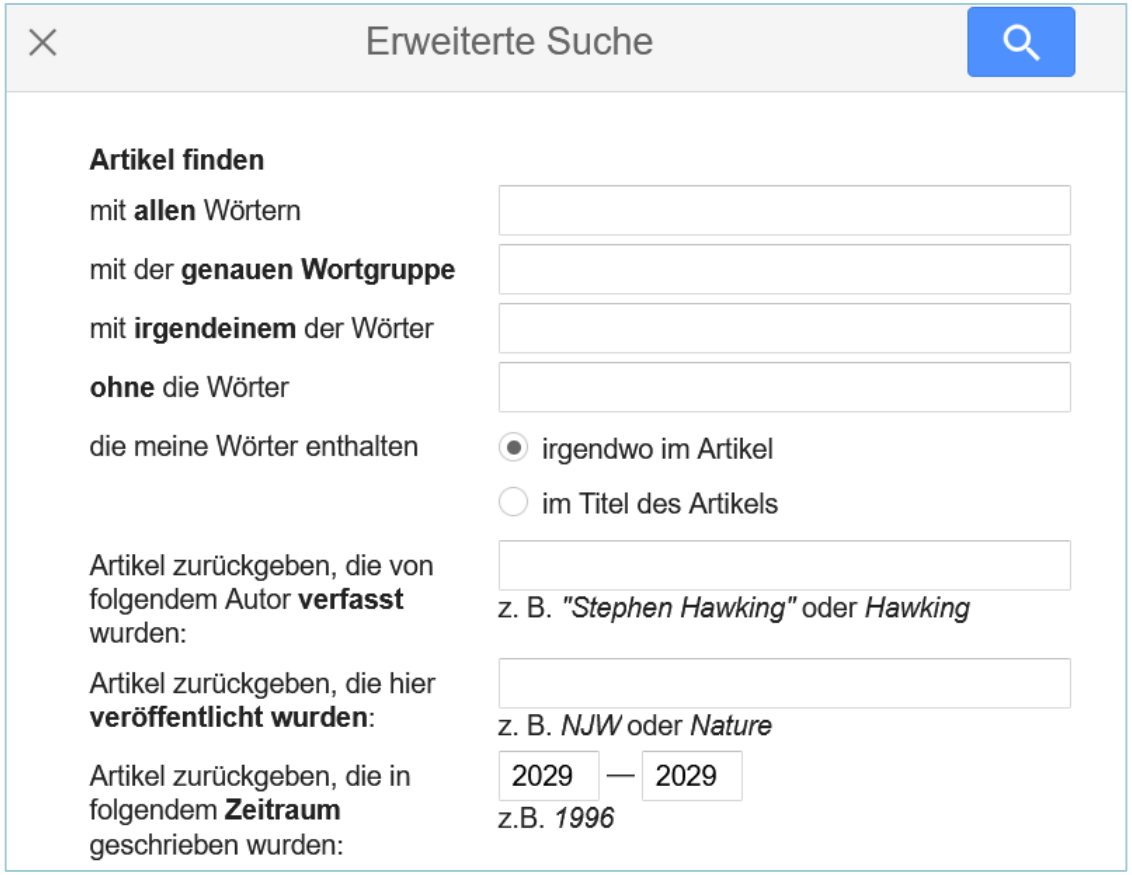
Common metadata errors in Google Scholar (Sauvayre, 2022):

- Author errors
- Citation errors (phantom citations)
- Title errors
- Publication year errors
- Publication source errors (journal name, etc.)

“only 2 of 281 (0.71%) references collected from GS were free from errors.”

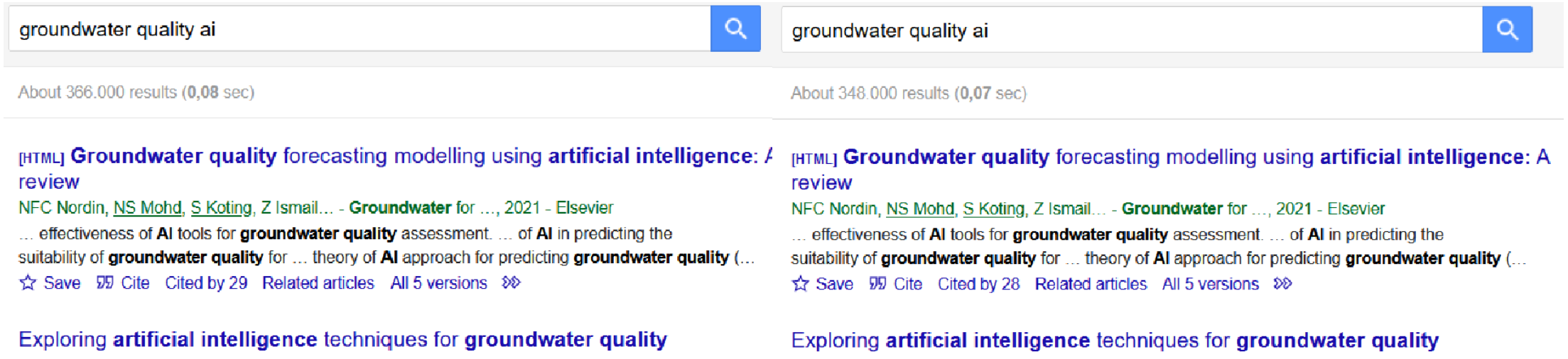
Limited search options

- Boolean operators (AND, OR, -) and quotation marks work
  (Ex. "deep drainage" OR "deep percolation")
- No truncation (geolog*)
- No filters (document type, peer reviewed, language, field, etc.)
- No subject indexing



Inconsistent search results

- Same search query, large fluctuations in number of hits (Gusenbauer, 2019; Bramer, 2016)
- Limits Google Scholar's usefulness for systematic searches



Best use of Google Scholar

- Do not limit yourself to Google Scholar, also use…
  - other multidisciplinary databases with fewer erroneous records and better search functions (Web of Science, BASE, ProQuest…)
  - specialized databases (Geo-Leo, GEODOK, etc.)
  - Datenbank-Infosystem
- Google Scholar might be better suited for “lookup searching” than for “exploratory and systematic searching” (Gusenbauer, 2021)

Thank you for your attention!

Any questions?

Source: DALL-E

Sources

Bramer W. M. (2016). Variation in number of hits for complex searches in Google Scholar. *Journal of the Medical Library Association*, *104*(2), 143-145. https://doi.org/10.3163/1536- 5050.104.2.009

Dion, M.L., Sumner, J.L., Mitchell, S.M. (2018). Gendered Citation Patterns across Political Science and Social Science Methodology Fields. *Political Analysis, 26*(3), 312-327. http://doi.org/10.1017/pan.2018.12

Gasparyan, A. Y., Yessirkepov, M., Voronov, A. A., Trukhachev, V. I., Kostyukova, E. I., Gerasimov, A. N., & Kitas, G. D. (2016). Specialist Bibliographic Databases. *Journal of Korean medical science, 31*(5), 660-673. https://doi.org/10.3346/jkms.2016.31.5.660

Georgas, H. (2014). Google vs. the library (part II): Student search patterns and behaviors when using Google and a federated search tool. *Portal: Libraries and the Academy, 14*(4), 503-532. Google Scholar. (n.d.) Search Help: Content Coverage. https://scholar.google.com/intl/fr/scholar/help.html#coverage

Gusenbauer, M. (2022.) Search where you will find most: Comparing the disciplinary coverage of 56 bibliographic databases, *Scientometrics 127*(5), 2683-2745, https://doi.org/10.1007/s11192-022-04289-7

Gusenbauer M. (2021). The age of abundant scholarly information and its synthesis- A time when 'just google it' is no longer enough. *Research synthesis methods*, *12*(6), 684-691. https://doi.org/10.1002/jrsm.1520

Gusenbauer, M. (2019). Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics,* 118, 177-214. https://doi.org/10.1007/s11192-018-2958-5

Gusenbauer, M., Haddaway, N.R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. Research Synthesis Methods 11(2), 181-217.

Larivière, V., Sugimoto, C.R. (2018). Measuring Research. Oxford University Press.

Lerman, K., Yu, Y., Morstatter, F., Pujara, J. (2022). Gendered citation patterns among the scientific elite. *Proceedings of the National Academy of Sciences.* https://doi.org/10.1073/pnas.2206070119

Nicholas, D., Boukacem-Zeghmouri, C., Rodríguez-Bravo, B., Xu, J., Watkinson, A., Abrizah, A., et al. (2017). Where and how early career researchers find scholarly information. *Learned Publishing,* 30, 19-29. https://doi.org/10.1002/leap.1087

Rovira, C., Codina, L., Guerrero-Solé, F. & Lopezosa, C. (2019). Ranking by Relevance and Citation Counts, a Comparative Study: Google Scholar, Microsoft Academic, WoS and Scopus. Future Internet, 11(9), 202. http://dx.doi.org/10.3390/fi11090202

Sauvayre R. (2022). Types of Errors Hiding in Google Scholar Data. Journal of medical Internet research, 24(5). https://doi.org/10.2196/28354 Van Noorden, R. (2014). Online collaboration: Scientists and the social network. Nature, 512, 126-129. https://doi.org/10.1038/512126a

Across all institutional types, the *h*-index on Google Scholar is higher than that of Scopus.

Marsicano, C.R., Braxton, J.M. & Nichols, A.R.K. The Use of Google Scholar for Tenure and Promotion Decisions. *Innov High Educ* **47**, 639–660 (2022). https://doi.org/10.1007/s10755-022-09592-y