Towards Visualizing Oceanographic Bibliometric Data Across Canada

Colin Orian[1], Poppy Riddle[2], Zaiyang Zhong[1], Geoff Krause[2], Stephen Brooks[1], Philippe Mongeon[2]
[1]Department of Computer Science, Dalhousie University
[2]Department of Information Science, Dalhousie University

Keywords: research location extraction, spatial bibliometrics, named entity recognition, ocean science, OpenAlex

Abstract

Global Environmental Assessments like the United Nations World Ocean Assessment II (WOA II) draw on thousands of scientific articles, but the bibliographic metadata for those articles records where researchers work — not where they conducted their research. A study of Arctic sea ice by a Dalhousie researcher is catalogued as Halifax. This gap makes it difficult to ask basic questions about which marine regions are being studied. We built a prototype system to extract and map research locations from articles referenced in WOA II. Using OpenAlex for bibliographic metadata and spaCy for named entity recognition, we identify place names in article abstracts, geocode them using the Canadian Geographical Names Database and GeoNames, and display them alongside institutional affiliations on an interactive map. Manual review of 68 articles shows the system performs well for studies referencing named geological or physical features in northern Canada but struggles with administrative place names in southern regions where ambiguity is higher. The tool is a first step toward spatial bibliometric analysis that maps where ocean science is done, not just who does it.

1. Introduction

1.1 Background

The World Ocean Assessments are a series of global environmental assessment (GEA) reports published by the United Nations (UN) to document the status of the world’s oceans. The references of the World Ocean Assessment II (WOAII) report have been used in prior work to create a corpus of ocean research (Toupin et al., 2025) as each chapter contained many references which had their own free and open bibliographic metadata.

The metadata from these references are relevant for scientists to identify gaps in research and for journalists to communicate scientific information to the public. This metadata can be explored by a user through interactive visualizations. A specific metadata that has been rarely explored is the research location(s), the specific areas of study that have been investigated by a researcher. This is different from the geographic locations of the institutional affiliations of the authors. As an example, a researcher from Dalhousie University (the affiliated institution) studied how tides in the Bay of Fundy (the research location) were impacted by new infrastructure (Figure 1.1).

By extracting the affiliation and research location from the metadata, both may be displayed on a map using latitude and longitude, linking institutions with the sites of research. To the best of our knowledge, finding and mapping research location(s) of a scientific paper is an under-explored topic. This purpose of this study was to develop a tool to extract and visualize research locations from scientific journal articles.

Figure 1.1: An example of the difference between research location and institutional affiliation.

2. Related Works

Several bodies of prior work inform this study. Researchers have used GEAs as a basis for constructing literature corpora, developed methods for extracting geographic information from scholarly texts, and built tools for visualizing bibliometric data geographically. Each of these areas has advanced independently, but none has directly addressed how to extract and map where research is actually conducted at scale.

Reports such as those from the UN or the Intergovernmental Panel on Climate Change (IPCC) are created using carefully selected lead authors who are authority figures in their field to synthesize published work. Reports like these are invaluable for governance and policy decision-making based on authoritative sources (Evans et al., 2021). As a more readily available and digestible format, GEAs like the WOAII report make information available that is typically inaccessible (Cossarini et al., 2014). In the case of the WOAII, interdisciplinary teams of 300 authors and 780 experts synthesized the “best available data and information” (United Nations Office of Legal Affairs, 2021) along with workshops, reviews, and input from members of the UN.

https://doi.org/10.5194/isprs-archives-XLIX-B4-2026-653-2026

GEAs have been used previously to create corpora of works centred on a research area. The purpose of creating such corpora can be to identify the cited research outputs and also new works that cite the same references, linking them topically (Boyack & Klavans, 2010; Kessler, 1963) and expanding the body of literature. Toupin et al. (2025) used WOAII to create a set of core works based on the references from the synthesis report and an extended set of works based on citation relationships. They took this approach in contrast to query-based approaches using bibliographic databases as those methods have limitations on topic assignment and search precision and accuracy (Kashnitsky et al., 2024). GEAs have also been analysed for their impact on governance and policy to increase transparency on knowledge choices (Jacquemont et al., 2025). However, while GEAs such as WOA II facilitate topical synthesis and citation-based corpus construction, they do not provide mechanisms for analyzing the geographic distribution of the underlying research sites, limiting insights into which marine regions and environments are being studied.

2.1 Scholarly Metadata

Each scholarly work, such as a conference paper, journal article, or poster, has metadata (Table 2.1) that can be used by visualization systems such as ResearchRabbit or recommendation systems such as Google Scholar. For example, Smith et al., (2024) used open bibliographic metadata to construct a social network of authors at Georgia Tech. Bibliographic databases have been used as a data source to analyse current ocean research (Guyot-Téphany et al., 2024). Bibliographic metadata, also called Big Scholarly Data (BSD), is difficult for a human to parse manually due to the volume of existing data, the velocity of new data, quality variations, variety of types, and veracity (Xia et al., 2017).

Past studies have examined problems of quality, types, and veracity of both closed and open databases (Céspedes et al., 2024; Chavarro et al., 2025; Culbert et al., 2024; Delgado-Quirós & Ortega, 2024; Haupka et al., 2024). OpenAlex as an open database, has over 400 million data points of BSD (Priem et al., 2022) and has been found to be appropriate as a source for analysis tasks (Alperin et al., 2024) when compared with closed databases. OpenAlex data may be collected from their interface, REST API, or by downloading the data snapshot. OpenAlex has limitations such as its source data, ingest processes, and API rate-limiting.

Metadata	Description
Authors	The name of the author(s) of the work
Citations	The works that cite the given work
References	Works that the given work cited
Publication Year	Year that the article was published

Table 2.1: A non-exhaustive list of scholarly metadata

While bibliographic databases contain data describing the scientific text and their relationships with authors, institutions, funding, they generally do not include location or geospatial coordinates for research sites. To gather this data from full text, past research has used manual and automated processes. Howell et al., (2019) investigated locations of study areas and manually assigned three labels to extracted place names: provided, estimated, and generalized. Coordinates were manually sourced from gazetteers for each place. Niers & Nüst (2020) approached the problem from the publishing production direction and proposed an OJS plugin – software to produce scientific articles used by publishers – for capturing geospatial data from text using a human-in-the-loop approach. Data was stored as text and coordinates and used to add a layer to the publisher landing page for the article enhancing search engine optimization.

Hauschke et al., (2021) built upon Niers & Nüst’s proposal for an OJS plugin to capture geospatial data related to the author’s research using manual and undisclosed automated methods for extraction of geographical text and use of gazetteers for coordinates. Yet, this too was a speculative design proposal. While manual processes described in these two examples have high accuracy, they are unrealistic and unscalable with current knowledge production rates.

Scalable deep learning networks have been utilized in prior work where automation improved accuracy over older machine learning methods. Zhang et al., (2025) improved upon rule-based extraction due to its limitations for term variants. They used a three-stage extraction method of full-text, sentences, entities using Bidirectional Long-Short Term Memory with Conditional Random Fields (BiLSTM-CRF). A fine-tuned SciBERT model was used for extraction of resources, such as software, data, and services, reported in scientific texts along with the relational terminology to create resource citation networks (Zheng et al., 2021). While scalable, these methods for extraction require the development of manually tagged training datasets.

While bibliographic metadata has been used by the scientometric community for the study of research production, geographic analyses has usually been limited to the author’s institutional affiliations resulting in country or region-level aggregation. Additionally, maps are typically simplified omitting areas without institutions (Guler et al., 2016) or aggregated into economic zones or continents. Geographic mapping using bibliographic data has explored citations and co-author relationships as social space (Rotolo et al., 2017). Frenken et al. (2009) described this as “spatial scientometrics” in their proposed framework for incorporating geographic proximity into analyses of scientific production and impact. Fu et al., (2025) developed GeoBM to visualize geographical bibliographic data using Python. GeoBM shows bibliographic data for countries and the relationships between the publications per country visually. Although GeoBM is easy to use due to being written in Python and provides country-level information, it lacks fine details about geographical data and the ability to show details within a country. Institution geo-locations have been used to identify hot and cold spots for research across the USA (Bornmann & De Moya Angeon, 2019). By visually displaying institution locations that have published high-impact papers, a user can identify which locations in a country are producing scientific information and can contribute to the formation of science polices.

Although institution geo-locations are beneficial in identifying locations that are contributing to science, it does not provide information on which locations are being studied. Wang et al. (2023) represented geospatial information extracted from news articles to find the location discussed in the article and developed a web tool to visualize the locations on a map. They then ran a case study by uploading news articles about dam failures across the USA and the tool was able to show the locations on the map. Acheson and Purves (2021) used a combination of a rules-based approach and machine learning to extract toponyms from scientific articles. The system first used machine learning to detect the locations in the text and then manually written rules filtered at any false positives. They used the Google Geocoding API to retrieve spatial information about the texts that were extracted. They also used the system on two different corpuses of research articles: a corpus focused on orchards and a corpus focused on cancer. They tested the system on a sample of a large amount of locations from the orchards corpus and a sample with a low amount from the cancer corpus. The system had high precision for identifying and resolving locations for both corpuses.

Despite this progress, most bibliographic databases emphasize author affiliation rather than the geographic setting of the research itself. For ocean science, this is a meaningful gap: a study of Arctic sea ice conducted by a researcher at Dalhousie is recorded as Halifax, not the Beaufort Sea. Addressing this requires going beyond institutional affiliation found in the metadata to extracting location data about the study location from the abstract text.

2.2 NER & Geo-Location

Named Entity Recognition (NER) is a natural language processing (NLP) method to identify signifiers attributed to predefined types, usually including such types as person, location, organization, or object (Li et al., 2022). In this context, we are interested in identifying place names for the purpose of locating them geographically. Orduña et al. (2024) developed a framework to extract meaningful location data from the text and metadata of research articles. They found that author affiliation and references are often used to geolocate the location of research articles. However, the content of the article, such as the abstract, title, and body were not explored.

Karl (2019) developed a lexical parser that used regular expressions as a rule-based approach to extract latitude and longitudes from research articles. The parser was able to successfully find and extract latitude/longitude pairs and then plot the locations onto a map. Although the parser was able to successfully find locations in research articles, this method can only work when a latitude and longitude is explicitly defined in the article and matched by a rule, not when a place is mentioned and the coordinates need to be determined via another dataset like a gazetteer. Acheson and Purves (2021) used Stanford NER as an unsupervised machine learning approach followed by rules-based filtering for relevancy to identify locations from research articles[1], improving upon past rule-based methods of location extraction such as the approach used by Karl.

Unlike most prior work, which relied on manual annotation, proprietary tools, or explicit coordinate extraction, our approach uses open-source NER and open gazetteers applied to article abstracts. We are not trying to precisely locate every study — the goal is to identify patterns across a large corpus, which requires methods that are open, repeatable, and fast enough to run at scale.

2.3 Data Visualization

Visualization of spatial data has been explored extensively, yet challenges remain with heterogeneous data, overdispersion, or distortion due to applying the wrong map type. In our context, we're looking for high-fidelity maps in an attempt to observe patterns or disparities. Ai et al. (2022) developed a tool that displayed spatial-temporal changes in published geoscience articles. The tool displayed the geographical centroid of research papers from 1994 to 2018 and a trend of the centroid drifting South-Eastward from the Atlantic Ocean to over Spain. Since the tool used the institution as the geographical data, it can track trends of *who* is doing research. Likewise, Fu et al., (2025) developed a system for displaying country-level but also higher fidelity locations for research collaboration analysis. Bornmann & De Moya Angeon (2019) had a similar limitation when determining research hotspots across Europe since the geographical coordinates the system used the author's affiliation's coordinates for the system, not the research site.

However, since these approaches were not tracking research locations, they are not able to track trends of *where* research was being conducted. In short, existing tools track where science comes from, not where it takes place. For fields like oceanography and coastal management, the study location is central to interpreting the findings. The current lack of methods to identify this is a significant gap.

3. Methodology

To address the research gap identified above, our approach combines process steps for NER using spaCy, open gazetteers CGND and GeoNames for geo-location, and bibliographic metadata from OpenAlex to extract and map research locations from a large corpus (Figure 3.1).

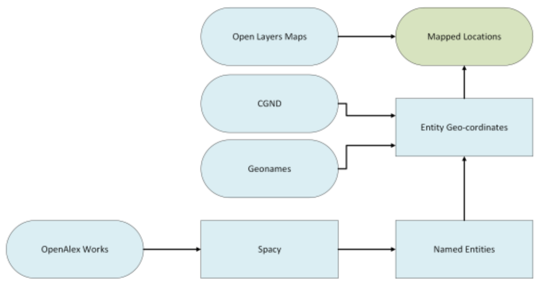

Figure 3.1 Process diagram for our methods using metadata from OpenAlex with additional inputs from Open Layers Maps, CGND, and GeoNames.

3.1 Named Entity Recognition

Using deep learning as our approach for text extraction (Li et al., 2022), we considered many libraries for NER and selected spaCy (2020) since it utilizes Python, can run on a local computer, and is free to use. SpaCy is also able to visualize the recognized text, which may be a useful feature for future development of the system. SpaCy provides an automated structuring of the install command and the parameters used in our system can be seen in Figure 3.2.

[1] https://github.com/eacheson/pyscine

https://doi.org/10.5194/isprs-archives-XLIX-B4-2026-653-2026 | © Author(s) 2026. CC BY 4.0 License.

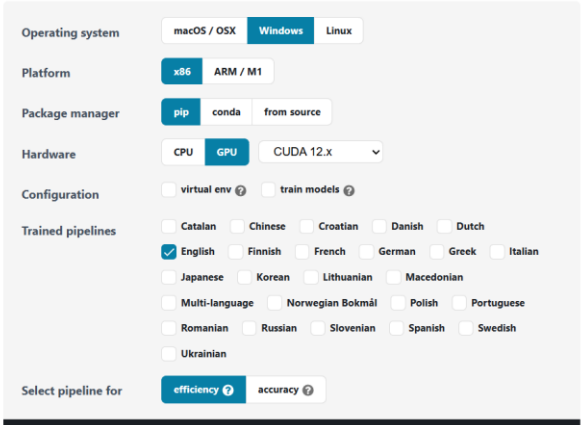

Figure 3.2 SpaCy can identify a variety of named entities, such as persons and locations.

Out of the named entities, we only included locations (LOCs) and geopolitical entities (GPEs). We then used spaCy to identify places from the abstracts of articles cited by the WOA II. Although abstracts lack the full details of the body of the article and are not standardized, they were suitable alternatives to closed source papers and less computationally expensive than full text parsing.

We used article abstracts rather than full text for two practical reasons: they are openly available for most articles in our corpus, and they are short enough to process at scale. Full text would give more spatial detail, but most articles in WOA II are paywalled, making consistent access impossible. For our purposes — identifying broad geographic patterns rather than precise study boundaries — abstracts contain enough information to be useful.

3.2 Geocoding

Many place names appear in multiple locations across Canada, which creates ambiguity during geocoding. Rather than automatically picking one match and discarding the others, we kept all Canadian candidates for each entity. This enabled us to examine the ambiguity directly, rather than hiding it behind an automated choice that may be wrong.

To determine the latitude and longitude of each location, we utilized two gazetteers. The first gazetteer used was the Canadian Geographical Names Database (CGND)[2]. Within the CGND are many different types of locations such as cities, administrative areas, and geographical features. Each location has relevant information, such as the type of location, province or territory it is in, and latitude and longitude. In addition to the CGND, we used the Canadian subset from Geonames[3]. Like the CGND, the Geonames dataset has information that we used to geocode the entities. Places that were not found in either gazetteer were assumed to be not in Canada and were removed.

3.3 Institution Geocoding

We utilized the OpenAlex dataset to collect geographical information about the institutions. Each work within OpenAlex has zero to many institutions associated with them. We extracted all the institutions from each work’s metadata and then accessed the institution information via the OpenAlex API to collect the institution’s country code, latitude, and longitude. Works were then associated with the institution that produced them.

3.4 Visualization

We used OpenLayers (openlayers.org, n.d.), NodeJS (OpenJS Foundation, n.d.), and D3 (Bostock, n.d.) to visualize the extracted data. OpenLayers offers ready access to OpenStreetMaps and D3 enables easy deployment for our map[4] . The latitude and longitude[5] for the study locations and the institutions were plotted on the map as red and blue circles respectively. Clicking on either a study location or institution will display a line linking the location to all institutions that have studied that location or all locations that the given institution has studied. (Figure 3.3).

Clicking on a study location will list all works that are associated with that location and all authors that have studied that location (Figure 3.4). Both works and authors have relevant information such as the work’s abstract and the authors ORCID (if available).

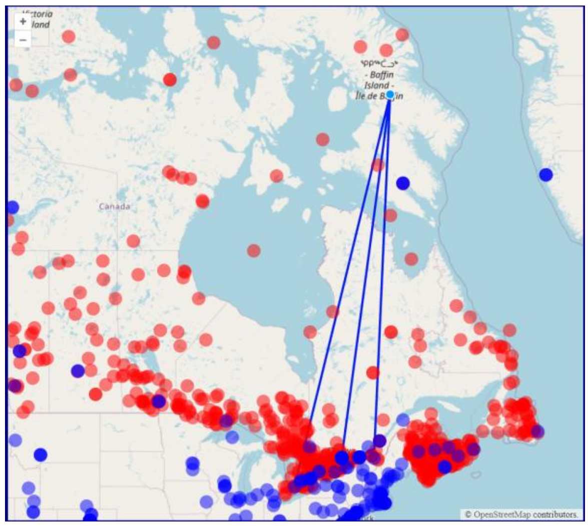

Figure 3.3 Example of institutions (blue) linked with study location (red).

[2]https://natural-resources.canada.ca/maps-tools-publications/tools-applications/canadian-geographical-names-database

[3] https://www.geonames.org/countries/CA/canada.html

[5] OpenLayer orders coordinates “longitude-latitude” instead of “latitude-longitude”, which may cause confusion when developing this work further.

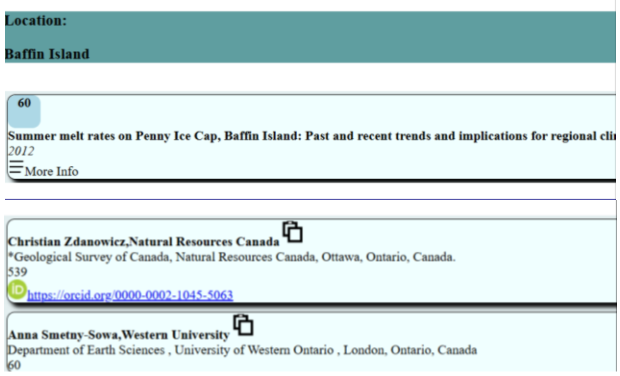

Figure 3.4 Prototype visualization showing extracted research locations and publishing institutions linked across Canada.

4. Results

We evaluated results manually, counting a geocoded location as a true positive if it was explicitly named in the abstract and accurately corresponded to where the study was conducted. False positives were locations that were either mapped to the wrong place or weren't relevant to the study at all. The system was able to successfully geo-locate study sites from the WOAII dataset throughout Canada.
From a manual review of 36 articles in northern Canada, 24 of them were true positives and 12 were false positives. For example, the system was able to identify Baffin Island, Nunavut. However, the geocoding falsely assigned locations, such as Long Island, to locations within Canada when the location was in another country. When investigating 32 locations within Southern Ontario, the only true positive was Toronto, Ontario. Nearly all the locations in Southern Ontario were administrative locations (cities, towns, regions, etc.). However, many of the true positives in northern Canada were geological features, such as the Foxe Basin. Although our examination was not exhaustive and may have human bias, we suspect that our tool is able to detect geographical features better than administrative locations.

The north/south split made sense in hindsight. Studies in northern Canada tend to name specific geological features — Baffin Island, Foxe Basin — that appear clearly in abstracts and have unambiguous gazetteer entries. Studies in southern Ontario more often reference cities and administrative regions, where the same name can refer to multiple places and the study location may not be stated directly. The tool is better at finding where research happen when researchers name the landscape rather than the nearest town. We initially expected administrative place names to be easier to resolve given their prevalence in gazetteers, but manual review showed the opposite in practice due to name reuse and vague spatial references.

Observed extraction errors fall into two types: (1) incorrect resolution, where a named entity is geocoded to the wrong place due to shared names (e.g., Long Island mapped to Canada instead of New York), and (2) over-resolution, where one name matches multiple valid gazetteer entries. Both are gazetteer ambiguity problems which has been previously discussed (Acheson & Purves, 2021; Leidner & Lieberman, 2011) — the NER itself is identifying the right entities.

The results show the system is useful for exploratory analysis, particularly in regions where studies reference named physical features. The limitations are real but well-defined, which gives clear direction for future work.

5. Conclusion

The goal of this research was to show that research location — where studies are specifically conducted — can be extracted and visualized at scale using entirely open tools. We built a prototype system that extracts locations from metadata and maps research locations from ocean science articles referenced in WOA II, using OpenAlex metadata, spaCy for named entity recognition, and open Canadian gazetteers.

Overall, the results were promising but highlighted challenges other researchers have also faced with ambiguous or multiple place names (Acheson & Purves, 2021) and the limitations of gazetteers. While our map only focused on place names in Canada, the system worked well for studies in northern Canada that reference named geological or physical features, which appeared explicitly in abstracts. It struggled in southern regions where place names were administrative and ambiguous. That contrast turned out to be informative. While it reflected clear technical limitations of the tool, it also highlighted differences in how location was described in scientific writing across regions and study types.

In contrast with prior work with maps used in bibliometric studies which typically reflect institutional geolocations (Bornmann & De Moya Angeon, 2019), this work contributes by showing how to utilize the abstract to extract and map study locations and proposing a pipeline of open-source software and libraries. This work is a step toward showing where research happens and contributes to providing evidence of where research occurs in the field beyond the institutional affiliations of the authors.

5.1 Limitations

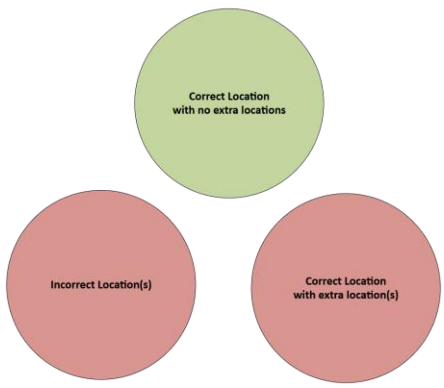

Figure 4.1 Example of ambiguous geocoding outcomes resulting from place name ambiguity in Canadian gazetteers.

There are two major limitations with the system currently (Figure 4.1). First, a named entity or entities are recognized in the abstracts and then geocoded to a place in Canada. However, the abstract is not referring to any of those places (incorrect location). For example, Alaska (The USA state) gets geocoded to Alaska, Prince Edward Island (PEI).

Second, the NER system recognizes a named entity and then the named entity is geocoded to multiple places in Canada. Only one of these places are the correct location geocoded locations (correct location with extra locations). For example, Victoria Island is correctly geocoded to the island in the North-West Territories and Nunavut, but is incorrectly geocoded to an island in Ottawa, Ontario.

This limitation is caused by ambiguity in the gazetteer. In future work, we will counteract this limitation by introducing a user-in-the-loop as a manual operation to correct the inaccurate locations or to append variations in names, such as Indigenous names. Another approach is an automated confidence score — the more places a name could refer to, the lower the confidence that any single match is correct, and ambiguous results could be flagged or filtered like the approach used by Acheson & Purves (2021).

Abstracts are also inconsistent as a data source. They vary in availability, length, language , and level of geographic detail — some name specific field sites, others describe regions only vaguely. Without access to full text, we may miss locations that are described in the body of a paper but not mentioned in the abstract. These variations make NER harder and limit how precisely we can compare spatial scope across different articles.

And lastly, we only used data from the World Ocean Assessment II, which limits our dataset to articles published before 2021. The World Ocean Assessment III will contain information up to 2025. While we used an ocean sciences domain as a test corpus, our approach should work on collections of works from any scientific domain.

5.2 Future Works

We're currently exploring how this tool can aid journalists in finding potential scientists to interview. The user study is currently in process, and we will publish our results in the future.

To alleviate the limitations of the system, a human-in-the-loop system can be developed to annotate article metadata with geocoded information. We could also use the WOA I and eventually WOA III to make a spatial-temporal mapping of research articles.

Finally, we only used gazetteers focused on Canada because global gazetteers would be significantly larger. Future work could generalize the system to a global scale.

References

Acheson, E., & Purves, R. S. (2021). Extracting and modeling geographic information from scientific articles. *PLOS ONE*, *16*(1), e0244918. https://doi.org/10.1371/journal.pone.0244918

Ai, X., Ma, M., Wang, X., & Kuang, H. (2022). A novel bibliometric and visual analysis of global geoscience research using landscape indices. *Frontiers of Earth Science*, *16*(2), 340–351. https://doi.org/10.1007/s11707-021-0875-z

Alperin, J. P., Portenoy, J., Demes, K., Larivière, V., & Haustein, S. (2024). *An analysis of the suitability of OpenAlex for bibliometric analyses* (arXiv:2404.17663). arXiv. https://doi.org/10.48550/arXiv.2404.17663

Bornmann, L., & De Moya Angeon, F. (2019). Hot and cold spots in the US research: A spatial analysis of bibliometric data on the institutional level. *Journal of Information Science*, *45*(1), 84–91. https://doi.org/10.1177/0165551518782829

Bostock, M. (n.d.). *D3* [JavaScript]. Retrieved https://github.com/d3/d3

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, *61*(12), 2389–2404. https://doi.org/10.1002/asi.21419

Céspedes, L., Kozlowski, D., Pradier, C., Sainte-Marie, M. H., Shokida, N. S., Benz, P., Poitras, C., Ninkov, A. B., Ebrahimy, S., Ayeni, P., Filali, S., Li, B., & Larivière, V. (2024). *Evaluating the Linguistic Coverage of OpenAlex: An Assessment of Metadata Accuracy and Completeness* (arXiv:2409.10633). arXiv. https://doi.org/10.48550/arXiv.2409.10633

Chavarro, D., Alperin, J. P., & Willinsky, J. (2025). *On the Open Road to Universal Indexing: OpenAlex and Open Journal Systems*. https://doi.org/10.1590/SciELOPreprints.11205

Cossarini, D. M., MacDonald, B. H., & Wells, P. G. (2014). Communicating marine environmental information to decision makers: Enablers and barriers to use of publications (grey literature) of the Gulf of Maine Council on the Marine Environment. *Ocean & Coastal Management*, *96*, 163–172. https://doi.org/10.1016/j.ocecoaman.2014.05.015

Culbert, J., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., & Mayr, P. (2024). *Reference Coverage Analysis of OpenAlex compared to Web of Science and Scopus* (arXiv:2401.16359). arXiv. https://doi.org/10.48550/arXiv.2401.16359

Delgado-Quirós, L., & Ortega, J. L. (2024). Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies*, *5*(1), 31–49. https://doi.org/10.1162/qss_a_00286

Evans, K., Zielinski, T., Chiba, S., Garcia-Soto, C., Ojaveer, H., Park, C., Ruwa, R., Schmidt, J. O., Simcock, A., Strati, A., & Vu, C. T. (2021). Transferring Complex Scientific Knowledge to Useable Products for Society: The Role of the Global Integrated Ocean Assessment and Challenges in the Effective Delivery of Ocean Knowledge. *Frontiers in Environmental Science*, *9*, 626532. https://doi.org/10.3389/fenvs.2021.626532

Frenken, K., Hardeman, S., & Hoekman, J. (2009). Spatial scientometrics: Towards a cumulative research program. *Journal of Informetrics, Science of Science: Conceptualizations and Models of Science*, *3*(3), 222–232. https://doi.org/10.1016/j.joi.2009.03.005

Fu, C. C., Fleta-Asín, J., Muñoz, F., Sáenz-Royo, C., & Wei, L. K. (2025). GeoBM: A Python-based tool for integrated visualization of global bibliometric data. *MethodsX*, *15*, 103497. https://doi.org/10.1016/j.mex.2025.103497

Guler, A. T., Waaijer, C. J. F., & Palmblad, M. (2016). Scientific workflows for bibliometrics. *Scientometrics*, *107*(2), 385–398. https://doi.org/10.1007/s11192-016-1885-6

Guyot-Téphany, J., Trouillet, B., Diederichsen, S., Juell-Skielse, E., Thomas, J.-B. E., McCann, J., Rebours, C., Scherer, M., Freeman, P., Gröndahl, F., Walsh, J. P., & Lukic, I. (2024). Two decades of research on ocean multi-use: Achievements, challenges and the need for transdisciplinarity. *Npj Ocean Sustainability*, *3*(1), Article 1. https://doi.org/10.1038/s44183-024-00043-z

Haupka, N., Culbert, J. H., Schniedermann, A., Jahn, N., & Mayr, P. (2024). *Analysis of the Publication and Document Types in OpenAlex, Web of Science, Scopus, Pubmed and Semantic Scholar* (arXiv:2406.15154). arXiv. https://doi.org/10.48550/arXiv.2406.15154

Hauschke, C., Nüst, D., Cordts, A., & Lilienthal, S. (2021). OPTIMETA – Strengthening the Open Access publishing system through open citations and spatiotemporal metadata. *Research Ideas and Outcomes*, *7*, e66264. https://doi.org/10.3897/rio.7.e66264

Howell, R. G., Petersen, S. L., Balzotti, C. S., Rogers, P. C., Jackson, M. W., & Hedrich, A. E. (2019). Using WebGIS to Develop a Spatial Bibliography for Organizing, Mapping, and Disseminating Research Information: A Case Study of Quaking Aspen. *Rangelands*, *41*(6), 244–247. https://doi.org/10.1016/j.rala.2019.10.001

Jacquemont, J., Brodie-Rudolph, T., Gerhardinger, L. C., Claudet, J., Pörtner, H.-O., & Gaill, F. (2025). Strengthening the seascape of global environmental assessments to support ocean sustainability. *Npj Ocean Sustainability*, *4*(1), 9. https://doi.org/10.1038/s44183-025-00108-7

Karl, J. W. (2019). Mining location information from life- and earth-sciences studies to facilitate knowledge discovery. *Journal of Librarianship and Information Science*, *51*(4), 1007–1021. https://doi.org/10.1177/0961000618759413

Kashnitsky, Y., Roberge, G., Mu, J., Kang, K., Wang, W., Vanderfeesten, M., Rivest, M., Chamezopoulos, S., Jaworek, R., Vignes, M., Jayabalasingham, B., Boonen, F., James, C., Doornenbal, M., & Labrosse, I. (2024). Evaluating approaches to identifying research supporting the United Nations Sustainable Development Goals. *Quantitative Science Studies*, *5*(2), 408–425. https://doi.org/10.1162/qss_a_00304

Kessler, M. M. (1963). Bibliographic Coupling Between Scientific Papers. *American Documentation*, *14*(1), 10–25.

Leidner, J. L., & Lieberman, M. D. (2011). Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, *3*(2), 5–11. https://doi.org/10.1145/2047296.2047298

Li, J., Sun, A., Han, J., & Li, C. (2022). A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, *34*(1), 50–70. https://doi.org/10.1109/TKDE.2020.2981314

Niers, T., & Nüst, D. (2020). Geospatial Metadata for Discovery in Scholarly Publishing. *Septentrio Conference Series*, (4). https://doi.org/10.7557/5.5590

OpenJS Foundation. (n.d.). *NodeJS* [Computer software]. Retrieved https://nodejs.org

openlayers.org. (n.d.). *Open Layers Maps* [Computer software]. Retrieved https://github.com/openlayers/openlayers

Orduña-Malea, E., Miguel, S., González, C. M., Arias, R. R., & Ortiz-Jaureguizar, E. (2024). An analytical framework for the study of geographical places in the scientific literature. *Revista Española de Documentación Científica*, *47*(3), e396. https://doi.org/10.3989/redc.2024.3.1571

Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2205.01833

Rotolo, D., Rafols, I., Hopkins, M. M., & Leydesdorff, L. (2017). Strategic intelligence on emerging technologies: Scientometric overlay mapping. *Journal of the Association for Information Science and Technology*, *68*(1), 214–233. https://doi.org/10.1002/asi.23631

Smith, K., Liu, F., Phanish, D., Chen, H.-J., Chen, R., & Contis, D. (2024). Research Collaboration Discovery through Neo4j Knowledge Graph. *Practice and Experience in Advanced Research Computing 2024: Human Powered Computing*, 1–7. https://doi.org/10.1145/3626203.3670539

spaCy: Natural Language Processing in Python. (2020). [Computer software]. https://doi.org/10.5281/zenodo.1212303

Toupin, R., Krause, G., Riddle, P. N., Hare, M., & Mongeon, P. (2025). Identifying Ocean-Related Literature Using the UN Second World Ocean Assessment Report. *Ocean and Society*, *2*, 8924. https://doi.org/10.17645/oas.8924

United Nations Office of Legal Affairs, (2021). *The Second World Ocean Assessment: World Ocean Assessment II - Volume I & II*. United Nations. https://doi.org/10.18356/9789216040062

Wang, Z., Jin, B., Hu, W., Jiang, M., Kang, S., Li, Z., Zhou, S., Han, J., & Wang, S. (2023). Geospatial Knowledge Hypercube. *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, 1–4. https://doi.org/10.1145/3589132.3625629

Xia, F., Wang, W., Bekele, T. M., & Liu, H. (2017). Big Scholarly Data: A Survey. *IEEE Transactions on Big Data*, *3*(1), 18–35. https://doi.org/10.1109/TBDATA.2016.2641460

Zhang, M., Wang, J., & Zhang, X. (2025). Mapping a Knowledge Graph of Flooding in Academic Literature Through Full-Text Entity Extraction. *Transactions in GIS*, *29*(2), e70045. https://doi.org/10.1111/tgis.70045

Zheng, A., Zhao, H., Luo, Z., Feng, C., Liu, X., & Ye, Y. (2021). Improving On-line Scientific Resource Profiling by Exploiting Resource Citation Information in the Literature. *Information Processing & Management*, *58*(5), 102638. https://doi.org/10.1016/j.ipm.2021.102638