Ocean Science Visualization Via the World Ocean Assessment

Colin Orian
Faculty of Computer Science
Dalhousie University
Halifax, Canada
colin.orian@dal.ca

Poppy Riddle
Department of Information Science
Dalhousie University
Halifax, Canada
pnriddle@dal.ca

Remi Toupin
Department of Information Science
Dalhousie University
Halifax, Canada
remi.toupin@dal.ca

Geoff Krause
Department of Information Science
Dalhousie University
Halifax, Canada
gkrause@dal.ca

Philippe Mongeon
Department of Information Science
Dalhousie University
Halifax, Canada
pmongeon@dal.ca

Stephen Brooks
Faculty of Computer Science
Dalhousie University
Halifax, Canada
sbrooks@cs.dal.ca

***Abstract*—The World Ocean Assessment is a comprehensive investigation of oceanography. The assessment contains over 3500 bibliographic references, however due to the size of the assessment it can be difficult for a human to effectively use this as a starting point for literature reviews. Therefore, we implemented a system that visualizes the relationships between the assessment's chapters, keywords, and bibliographic metadata. Filters and a tagging system were also implemented to make finding papers easier. Finally, we demonstrate the system's features through a case study.**

***Index Terms*—Scholarly Data Visualization, Oceanography, Big Scholarly Data**

I. Introduction

In 2015 the United Nations (UN) published the World Ocean Assessment 1 (WOA I) to report on the current status of the ocean and updated the report in 2020 with the World Ocean Assessment 2 (WOA II) [1], [2]. These reports contain a vast amount of scholarly information, such as authors, publication dates, and topics. In addition, the individual articles within WOA I & II have been cited and have references outside of both WOAs. Most of this data has be extracted and parsed by Tupin et al. through OpenAlex [3], [4] consisting of 3,532 works as the core dataset with a further 419,079 in the full dataset. Although this data is relevant for oceanographic research, the size and complexity can be too burdensome for a human to parse through to extract relevant information [5]. This paper presents a scholarly data visualization (SDV) system that visualizes metadata elements from the core dataset extracted by Tupin et al. to make the data more understandable and accessible for humans.

The paper is structured in the following way: A literature review is conducted that summarizes big scholarly data (BSD) and how large amounts of scholarly data can be visualized. Next, the system is described by explaining the data structure and the different components of the system. We discussed the findings of the system and a case study is presented on a potential application of the system. Finally, the system's limitations and future work is discussed.

II. Related Works

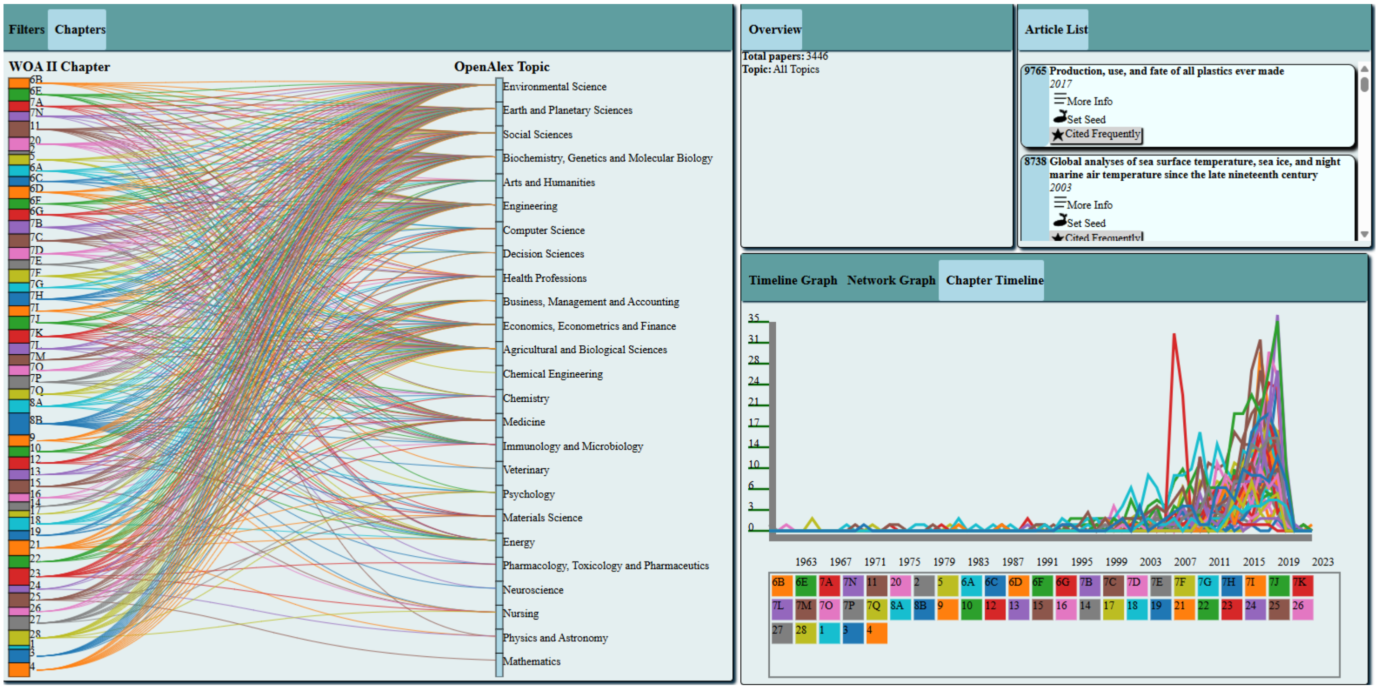
A. Big Scholarly Data & Scholarly Data Visualization

The amount of scientific literature has grown dramatically over the past several decades, benefiting humanity, but has resulted in the problem of big scholarly data [5]. Finding relevant information out of BSD can be difficult due to what Xia et al. calls the "5Vs of Big Scholarly Data". Due to the large *volume* of articles, it can be difficult for a human to find articles. Even when a human finds a paper, it may not have the desired *value* because it may not be relevant to the researcher's goal or be of poor quality. The number of papers published every year is growing, thus increasing the *velocity* of publications. There is a large *variety* of data points such as authors and articles.

OpenAlex is a free open source scholarly knowledge graph (SKG) containing more than 266 million metadata records structured in a graph network [4]. OpenAlex uses several types of entities in the graph and therefore creates a large variety of data. Of these types of entities, our system uses the *works*, *authors*, and *topics* entities and they are described in greater detail in the **data schema** section of the paper.

He et al. analyzed interactive recommendation systems to identify research gaps [6]. They found that providing a more transparent system and users with more control was well explored. They identified many systems do not understand the user's current emotional state and were not context aware. Furthermore, recommendation systems had difficulties in providing recommendations when the user first starts using the system.

This paper was funded by the Canada First Research Excellence Fund



Total papers: 3446
Topic: All Topics

9765 **Production, use, and fate of all plastics ever made**
2017
More Info
Set Seed
★Cited Frequently

8738 **Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century**
2003
More Info
Set Seed
★Cited Frequently

Fig. 1. Key features of the system includes the flow of chapter to topic and the timeline for the chapters.

Dattolo et al. developed a visualization system called VisualBib[1] to manage a user's bibliography [7]. Through an extensive literature review, they identified 24 system requirements important for a visualization system. The following are those specific to visualization.

- Provide an overview of papers, authors, and relationships
- Visualize networks of citations, subject areas, keywords, and tags
- Provide different data ordering criteria
- Identify most frequent items
- Visualize paper chronology
- Visualize collaboration networks
- Visualize self-citations

III. System Overview

We had two objectives when developing our visualization system:

1) Visualize relationships that WOA II has with other metadata.
2) Help the users easily access works within the dataset.

The system has two tabs that provide details about WOA II's chapters' relationships between topics and time. A filter was implemented so a user can filter out unwanted articles from a list of articles. The list of articles uses an tagging system to identify potentially relevant articles.

The data set from Toupin et al., including the core set of 3,532 works with metadata from 49 chapters, was cleaned, processed, and structured using R and Python for use with the Javascript library D3. The visualization system[2] was made using D3 because it is often used in similar scholarly data visualization systems [7], [8] and is easily deployed. When cleaning the data, one article from 1874[3] was removed because it skewed timeline visualizations. The system has four sections (Figure 1) and each of these sections have one or more components within them.

A. Data Schema

The dataset contains over 3500 articles that are stored in a JSON file or in a CSV[4] file. Each article has a variety of relationships (Figure 3).

1) Woaii Chapter: *woaii_chapter* is an identifier for each chapter in WOA II (ex. 1, 7G) while *chapter_title* is the title of the chapter.

2) Core Works: The core works is a collection of the articles that are in WOA II. *id* is an unique OpenAlex URL that can be used to retrieve articles. *woaii_chapter* is the chapter the article is in. *source_id* and *source_name* are where the article was published. *Type* is what type of work it is (ex. paper). *Cited_by_count* and *reference_count* is how many works cited the given work and how many works the given work cited, respectively.

3) Topic Data: The topic data is represented as a hierarchy (Figure 2) that becomes more specific as it progresses from **Domain ->Fields ->Subfields ->Topics** [4]. Topics are

[1] https://visualbib.uniud.it/en/vb3-en/

[2] The code for the system, Python code, and data can be found at https://github.com/Colin-Orian/Visualization-Final-Project

[3] The Structure and Distribution of Coral Reefs by Charles Darwin

[4] The csv files used the pipe character (|) to delimitate the columns because some topics had commas in the text.

Child	Parent	Paper Count
All Topics	*empty*	*empty*
Impacts of Climate Change on Marine Fisheries	Global and Planetary Change	699
Global and Planetary Change	Environmental Science	*empty*
Environmental Science	Physical Science	*empty*
Physical Science	All Topics	*empty*

TABLE I
AN EXAMPLE OF HOW THE HIERARCHY DATA IS REPRESENTED IN THE CSV FILE.

algorithmically applied by OpenAlex and derived from the title, abstract, source, and citations[5]. Since each topic has a hierarchical relationship to three other topics in the array, the *i* property is used to define which topics are in the same hierarchy. The *name* property contains how high the topic is on the hierarchy and the *display_name* property is the name of the topic.

Some topics were not initially unique. For example, the topic *Social Science* is both a **domain** topic and **field** topic. To disambiguate topics, the parent topic was appended to children topic. Social Science is an exception to this rule because the **field** Social Science would be converted into *Social Science (Social Science)*. Instead, the social science topics had their respective hierarchy appended to the text. The leaves of the hierarchy is the number of papers that fall into that topic. The topic hierarchy was converted into a csv file so that the *d3.stratify()* function could structure the hierarchy in a format that could be visualized. A root parent *All Topics* was included to satisfy the requirements for the *straify()* function.

4) Article Data: The article data contains an *id* which is an URL to the OpenAlex website and acts as a unique identifier and retriever for all data points in the OpenAlex database. The *display_name* property is the name of the article and the *ab* property contains the abstract, if available. The *publication_year* property is the year the article was published and the *doi* property contains the doi of the article if the article has one. The *cited_by_count* property is the number of articles that have cited the article. Each article also has several relationships. First, an article has a relationship to one or many authors through the *author* property. An article has a relationship with the core works through the *id*. An article also has a relationship with 12 topics through the *topics* property. Finally, an article can have a relationship to the articles that it referenced through the *referenced_works* property.

5) Author Data: The author data has two properties. The *au_id* is an URL that uniquely identifies authors and the *au_display_name* property is the author's name.

B. Time Filter Component

The time filter component allows the user to filter the entire corpus based on the publication date by dragging bars that represent the start and end years. Making changes to the filter will update the data in the other sections of the visualization system. A potential use case may be for the user to explore the impact of a major global event (such as the Fukushima nuclear accident in 2011).

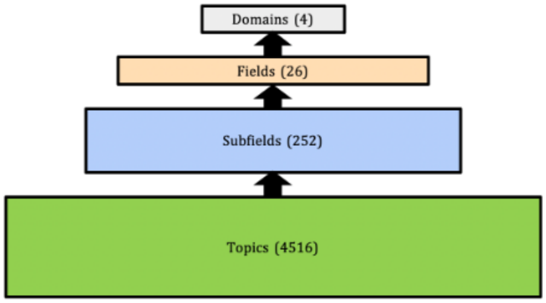


Fig. 2. The OpenAlex hierarchy as defined by [4].

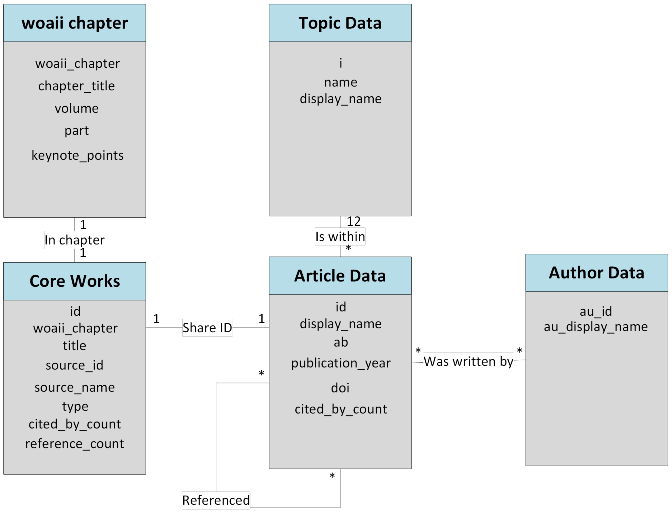


Fig. 3. The data schema for the system contains 5 data points and have a variety of relationships between them.

C. Topic Filter Component

The topic filter allows the user to select relevant topics from the corpus and was derived from an example from the D3 website [9]. Like the time filter component, adjusting the filter will update the other sections of the visualization. To reduce clutter on the screen, the name of the wedge is only visible when hovered over. When the user clicks on a wedge, the circle zooms into that wedge to provide better clarity for the child wedges. A sunburst diagram (Figure 4) was decided because it was able to demonstrate the percentage of the papers that fall into each topics as a pie chart. The colours for the sunburst diagram were chosen based on the research domains.

D. Article List Component

The article list contains all articles that fall within the time filter and the hierarchy filter. Each article within the list is sorted by the number of times the given article has been cited. Each entry also has a button that will display the article's abstract, authors, and the DOI and a button that sets that article

[5] https://docs.openalex.org/api-entities/topics

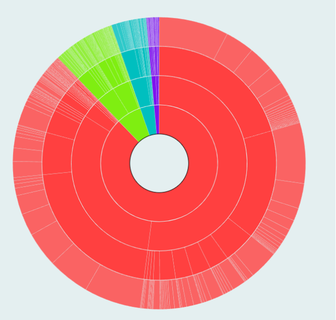
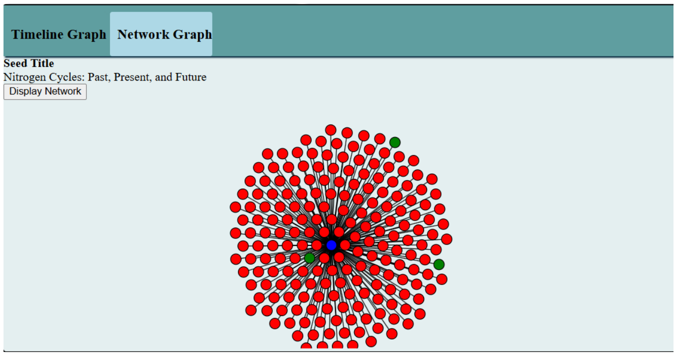


Fig. 4. The sunburst diagram displays the OpenAlex hierarchies and become increasingly refined the farther the wedge is from the center. The red wedges represent topics within *environmental science* and the large amount of red wedges is expected given the corpus' focus.

Tag	Description
Highly Cited	Has at least 10 citations per year on average.
Literature Review	Has at least 100 references
New Paper	Has been published within the last 4 years
Unnoted	Has not been cited before

TABLE II
EACH ARTICLE IN THE ARTICLE LIST MAY HAVE THESE TAGS AND THE TAGS ARE DEFINED PER THE DESCRIPTION.



Seed Title
Nitrogen Cycles: Past, Present, and Future

Fig. 5. The citation network for a literature review has over 100 references.

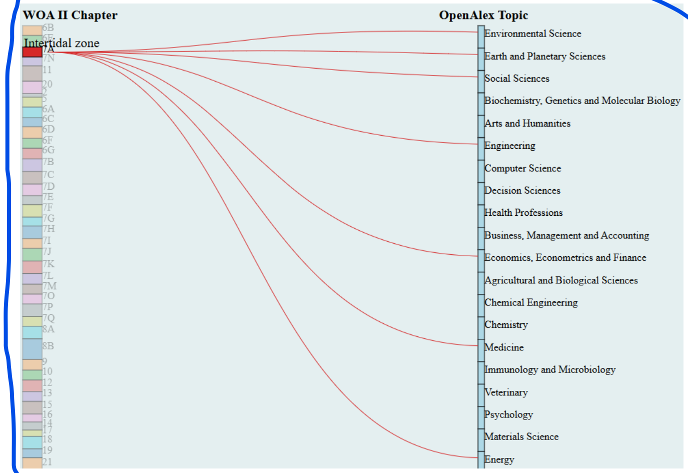


Fig. 6. All the topics associated with the chapter *7A*, also titled *Intertidal Zone*.

as the seed for a network diagram. Similar to PureSuggest, a tagging system was made to visualize potentially relevant papers [8] (Table II).

E. Chapter-Topic Relationship

The Chapter-Topic Relationship visualizes which topics are covered in each chapter. This was done by identifying which works were within a chapter and then associating those works to an OpenAlex subfield. All the OpenAlex topics are *subfields*. *Subfields* were chosen arbitrarily because *subfields* were not very broad or very specific, striking a balance between the two. Hovering over a chapter or topic will make links not associated with the chapter or topic disappear for added clarity (Figure 6). Hovering over a chapter will annotate the chapter with it's title. Two chapters were not in the core set of works, but were included in the visualization for completeness.

F. Publication Timeline Component

The publication timeline counts how many papers within the filtered corpus were published in each year and then constructs a line graph of the results. Therefore, the user can see how the number of publications changes over time. The start and end of the publication timeline depend on the timeline filter.

G. Chapter Timeline

The chapter timeline visualizes how many papers were published each year for each of the chapters and plots them on a timeline. Hovering over a chapter in the legend will hide the other chapter timelines and display the chapter's title.

H. Citation Network Component

If a seed paper is selected from the article list, the citation network will display all the papers that the selected article cited (Figure 5) [10]. This is done by iterating through the OpenAlex *referenced_works* property. Each of the nodes in the network are coloured in three different colours: Blue represents the seed article, green represents an article already within the corpus, and red represents an article outside of the corpus. Hovering over the nodes will display the OpenAlex ID and clicking on the nodes will open another tab that shows the article on the OpenAlex website.

IV. DISCUSSION

In the system's current state, it is possible to identify relevant information from the visualization. Although the tagging system uses heuristics in defining the tags, the heuristics are able to provide significant benefits to a user. For example, a user can rely on the literature review tag in the early stages of researching a topic to find relevant papers.

The chapter-topic relationship visualization provides expected insights for an oceanographic corpus, such as many of the chapters have a relationship to *Environmental Science*. However, some relationships are not expected and should be investigated in more detail. There is a relationship between the topic *Neuroscience* and the chapter *Biogenic reefs and sandy, muddy, and rocky shore substrates*. Filtering the papers into the *Neuroscience* topic shows three out of the four papers are about toxins. These results therefore raise questions. Are toxins in reef and shores having a negative impact on the human brain and how? Or is this an ambiguity in the automatically assigned topics by OpenAlex?

The current dataset is very broad and may contain articles that are not relevant to users research goals. Therefore, the hierarchy filter allows the user to focus on their specific goals. For example, the paper *Assessing the Economic Value of Beach Restoration: Case of Song-do Beach, Korea* has the following topic hierarchy: **Social Science (Domian) ->Social Science (Field) ->Cultural Studies ->Fashion Trends among Generation Z**. This paper may be relevant for social sciences, however may not be relevant to an oceanographer trying to find how climate change impacts aquatic life.

Since the dataset is local and static, there is no *velocity* to the dataset. *Veracity* was improved by disambiguating the hierarchy topics. The system is currently limited in *variety* to the core dataset and is not as diverse as the full dataset. The filtering system was able to quickly reduce the *volume* within the dataset in an interactive way and the tagging system helps users find highly *valuable* articles, such as literature reviews.

The chapter timeline shows that the number of papers related to oceanography grows. A large amount of articles were published between 2015 and 2019, which coincides with when the UN decided to compose WOA II [2] and is expected. However, the chapter *Marine Life and Macroalgae* does not follow this trend and instead has a major spike in publications around 2007. This outlier can be checked by looking at the chapter's references and many of the references in 2007 come from a single person's multiple reports of threatened species [2].

We implemented several of the requirements found by Dattolo et al [7]. The most frequently cited articles have been identified through the tagging system and have also be ordered by the most cited. The citation network provides a visualization on a small network of citations. Although papers were not visualized chronologically, the system is able to visualize how the corpus of articles changes over time. Finally, the article list component provides a quick overview for each article and can be clicked on to provide a greater amount of information.

Our system and GraphDetective [11] utilizes OpenAlex for climate related research; however, there are some key differences. GraphDetective focuses on making database queries easier to do through a visual approach, while our system does not use database queries. Our system is biased towards scholarly data, while GraphDetective draws articles from multiple sources. Although making decisions based on scholarly data is important, using exclusively academic articles leaves our system blind to opinions of non-scholars.

A. Case Study

Suppose an epidemiologist is starting their research into the relationship between oceanographic research and the spread of diseases. The user selects the purple *health science* wedge of the hierarchy filter because they assume that epidemiology would fall into health science. This filter results in 126 papers that the epidemiologist needs to look through.

The epidemiologist finds the paper *Shellfish-Borne Viral Outbreaks: A Systematic Review*. The epidemiologist selects the paper to bring up more info. Although the abstract was not found in the dataset, the visualization system provides a DOI to the article. The epidemiologist opens the link in a separate tab and sets the article aside to read after they are done exploring the dataset.

The epidemiologist then populates the network display using *Shellfish-Borne Viral Outbreaks: A Systematic Review* as the seed article. When the network is displayed, the user finds that the paper cited another article in the dataset. The user clicks on the article to get more information from the OpenAlex website. The user finds out that the paper cited is *Selective Accumulation May Account for Shellfish-Associated Viral Illness* and can be found under the *Viral gastroenteritis research and epidemiology* topic.

The epidemiologist can also use the network graph to find papers that are out of their domain of research. For example, using the paper *The Marine Viromes of Four Oceanic Regions* as a seed paper, the user can find the paper *Community structure and metabolism through reconstruction of microbial genomes from the environment*. The latter article is in the following topic chain: **Physical Sciences ->Engineering->Biomedical Engineering ->Metal Extraction and Bioleaching**.

The epidemiologist looks at the chapter-topic relationship diagram and hovers over the *Immunology and Microbiology* topic and finds that the topic has a relationship with 12 of the 52 chapters in WOA II. The epidemiologist could then later look at those chapters the see if the chapters can contribute to the epidemiologist's research.

Through the visualization tool, the epidemiologist was able to achieve several goals. The user was able to find a literature survey with relative ease that could used as a seed to grow their knowledge over the domain. Also, the user was able to use the tool to expand their knowledge outside of their domain specialization, potentially finding a paper that may not have been discovered otherwise.

B. Limitations

Although the core dataset used is broad, it is still missing a large amount of oceanographic papers. It is very likely that relevant articles are not included in the dataset and future work should use the full dataset. Furthermore, some metadata elements are missing, such as referenced papers or abstracts. The system could potentially use an online database to provide the missing properties or other methods to enhance the metadata.

The citation network also does not expand the network beyond the seed node because it would require multiple API calls to the OpenAlex servers and OpenAlex only allows 10 API calls per second [12]. A wait time of 2 seconds can be tolerable for simple information retrieval tasks [13] so a large citation network, such as a literature review, would likely cause the user to be frustrated. The full dataset would help expand the network but would require a database connection.

C. Future Work

The system mainly focuses on article data, but OpenAlex has multiple data points that might be relevant to BSD. The system only uses author data to display the author(s) of an article. However, co-authorship can be used to construct a social network and be used to identify scientific communities [5]. Identifying scientific communities may allow the user to find potential collaborators when exploring a novel topic. Similarly, OpenAlex has data pertaining to the institution that the paper originated from and can used to identify which institutions are collaborating with each other [14]. Using both author and institution data, a future system could potentially make a heatmap of research around the world.

The system could also use other scholarly databases to address metadata gaps, however, as OpenAlex already aggregates metadata from multiple sources, this risks duplication of efforts. Metadata enhancement through other automated means should be explored, particularly if moving from the core to the full dataset.

The system currently uses the number of citations as a way to rank the articles within the list. A more complicated algorithm, such as PageRank [15], or other normalized citation measures [16] could be used to rank the articles within the list.

The system only utilizes data from WOA II. Extracting article data from WOA I in the same way as the data from WOA II could allow comparison between the two versions to identify changes in focus between the two assessments or find other relevant information. The third World Ocean Report is currently being written and will likely provide similar insights when it is fully released [17]. Comparing all three reports will provide a full decade of oceanographic research and could be used to highlight major changes in focuses throughout the years.

In addition to further work needed on topic refinement, as each article falls into three different hierarchies, the hierarchy may be used to identify interdisciplinary communities in a way similar to ScholarNode [18].

We intend to do a user study to evaluate the system and identify potential new features for the system.

V. Conclusion

This paper presents a potential solution to visualize several thousand research articles. A filtering system lets users hide unwanted articles. The relationships between WOA II chapters and OpenAlex topics helps identify what topics are relevant to climate change research. Finally, the system presented can eventually be expanded to include WOA I and WOA III so that users can see the growth of research over time.

References

[1] "The first global integrated marine assessment : world ocean assessment i," p. 973 p. :, 2017, annexes (p. 953-969): 1. List of contributors and commentators – 2. Glossary – 3. Acronyms. [Online]. Available: http://digitallibrary.un.org/record/1291159

[2] U. N. O. of Legal Affairs, *The Second World Ocean Assessment*. United Nations, 2021. [Online]. Available: https://www.un-ilibrary.org/content/books/9789216040062

[3] R. Toupin, G. Krause, P. N. Riddle, M. Hare, and P. Mongeon, "Identifying Ocean-Related Literature Using the UN Second World Ocean Assessment Report," *Ocean and Society*, vol. 2, no. 0, Jan. 2025. [Online]. Available: https://www.cogitatiopress.com/oceanandsociety/article/view/8924

[4] J. Priem, H. Piwowar, and R. Orr, "Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts," 2022. [Online]. Available: https://arxiv.org/abs/2205.01833

[5] F. Xia, W. Wang, T. M. Bekele, and H. Liu, "Big scholarly data: A survey," *IEEE Transactions on Big Data*, vol. 3, no. 1, pp. 18–35, 2017.

[6] C. He, D. Parra, and K. Verbert, "Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities," *Expert Systems with Applications*, vol. 56, pp. 9–27, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417416300367

[7] A. Dattolo, M. Corbatto, and M. Angelini, "Authoring and Reviewing Bibliographies: Design and Development of a Visual Analytics Online Platform," *IEEE Access*, vol. 10, pp. 21 631–21 645, 2022, conference Name: IEEE Access. [Online]. Available: https://ieeexplore.ieee.org/document/9718060

[8] F. Beck, "PUREsuggest: Citation-Based Literature Search and Visual Exploration with Keyword-Controlled Rankings," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 1, pp. 316–326, Jan. 2025, conference Name: IEEE Transactions on Visualization and Computer Graphics. [Online]. Available: https://ieeexplore.ieee.org/document/10670434

[9] Observable, "Zoomable sunburst / D3 | Observable," 2025. [Online]. Available: https://observablehq.com/@d3/zoomable-sunburst, https://observablehq.com/@d3/zoomable-sunburst

[10] "Force-directed graph / D3 | Observable," 2023. [Online]. Available: https://observablehq.com/@d3/force-directed-graph/2,https://observablehq.com/@d3/force-directed-graph/2

[11] D. Opitz, A. Hamm, R. El Baff, J. Korte, and T. Hecking, "Graph Detective: A User Interface for Intuitive Graph Exploration Through Visualized Queries," in *Proceedings of the ACM Symposium on Document Engineering 2024*, ser. DocEng '24. New York, NY, USA: Association for Computing Machinery, Sep. 2024, pp. 1–9. [Online]. Available: https://dl.acm.org/doi/10.1145/3685650.3685660

[12] "Rate limits and authentication | OpenAlex technical documentation," Feb. 2025. [Online]. Available: https://docs.openalex.org/how-to-use-the-api/rate-limits-and-authentication

[13] F. F. H. Nah, "A study on tolerable waiting time: how long are web users willing to wait?" *Behaviour & Information Technology*, vol. 23, no. 3, pp. 153–163, 2004. [Online]. Available: https://doi.org/10.1080/01449290410001669914

[14] K. Smith, F. Liu, D. Phanish, H.-J. Chen, R. Chen, and D. Contis, "Research Collaboration Discovery through Neo4j Knowledge Graph," in *Practice and Experience in Advanced Research Computing 2024: Human Powered Computing*, ser. PEARC '24. New York, NY, USA: Association for Computing Machinery, Jul. 2024, pp. 1–7. [Online]. Available: https://dl.acm.org/doi/10.1145/3626203.3670539

[15] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1, pp. 107–117, Apr. 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016975529800110X

[16] L. Bornmann, "How can citation impact in bibliometrics be normalized? A new approach combining citing-side normalization and citation percentiles," *Quantitative Science Studies*, vol. 1, no. 4, pp. 1553–1569, Dec. 2020, _eprint: https://direct.mit.edu/qss/article-pdf/1/4/1553/1871000/qss_a_00089.pdf. [Online]. Available: https://doi.org/10.1162/qss_a_00089

[17] "Third cycle of the Regular Process | Division for Ocean Affairs and the Law of the Sea." [Online]. Available: https://www.un.org/regularprocess/cycle3

[18] M. A. Noor, J. A. Clark, and J. W. Sheppard, "ScholarNodes: Applying Content-based Filtering to Recommend Interdisciplinary Communities within Scholarly Social Networks," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '24. New York, NY, USA: Association for Computing Machinery, Jul. 2024, pp. 2791–2795. [Online]. Available: https://dl.acm.org/doi/10.1145/3626772.3657668