Leveraging open source resources to preserve research outputs from a large interdisciplinary water science project: case study of peer review publications

Morteza Behbooei, University of Waterloo; Kamalloo Ehsan, Persaud Bhaleka D., Eager Sara, Goucher Nancy, Grant Julie, Van Cappellen Philippe, Lin Jimmy

The Canada First Research Excellence Fund provided CDN $77.8 million to the Global Water Futures Programme (GWF) to generate practical scientific knowledge on how to forecast, prepare for, and manage water futures in Canada, given the anticipated risks associated with climate change. Between 2017 and 2021, GWF has produced thousands of research outputs including peer-reviewed publications, books chapters, articles in media, conference presentations, and datasets. To make these findings more accessible, we are leveraging artificial intelligence and other open access computing resources to create a user-friendly, searchable, and accessible one stop shop interface.

This research tested the feasibility of adapting the ACL Anthology Network, a popular resource in the field of Computational Linguistics, to promote GWF peer-reviewed publications, as this is one key category of research outputs. The GWF anthology output consists of over 1000 peer-reviewed publications, each accessible via a unique identifier, and includes various statistics about individual authors and publications. Our team also utilized cutting-edge technology to develop a highly efficient publication clustering project. Our approach involved implementing a BERT-based model to generate embeddings for each publication using both the title and abstract of the publication. This allowed us to capture both the broad themes and specific details of each piece of work, ensuring that the clustering model would have a robust set of data to work with. In addition, we utilized a k-means clustering model to group together similar publications based on their subject matter, making it easier for users to find articles and papers that are relevant to their interests. With this tool, users can easily filter through publications on a particular topic, saving them valuable time and effort.

By leveraging these open sourced resources, we hope to share our unique approach towards publication accessibility with other large interdisciplinary projects, who could replicate this approach, thereby saving significant time and resources. This approach is particularly relevant given the significant investment in such projects by Canada and other countries. By adapting these techniques, researchers and project managers can build on the success of past projects and make further advancements in data accessibility and open access resources.