Geophysical Research Abstracts
Vol. 20, EGU2018-14542, 2018
EGU General Assembly 2018

Geo-coding geoscience articles for spatial search index with FOSS

Alexander Kmoch (1), Evelyn Uuemaa (1), and Hermann Klug (2)

(1) University of Tartu, Institute of Ecology and Earth Sciences, Department of Geography, Tartu, Estonia (alexander.kmoch@ut.ee), (2) Interfaculty Department of Geoinformatics - Z_GIS, University of Salzburg, Austria

We analyzed the corpus of three geoscientific journals to investigate if there are enough locational references in research articles to apply a geographical search method. For that we used the 'GNU parallel' library and 'Tesseract OCR' to digitize and transcode PDFs into plain text. We searched title, abstracts and full texts for place name occurrences that match records from a gazetteer with the help of the Python 'pandas' library and stored results in an 'sqlite' database. The results indicate that the use of well well-crafted abstracts for journal articles with carefully chosen place names of relevance for the article provides a guideline for geographically referencing unstructured information like journal articles and reports in order to make such resources discoverable through geographical queries. We used the Term Frequency – Inverse Document Frequency (TF-IDF) algorithm from the Python 'Scikit-learn' library to generate keywords for articles that didn't provide any. Finally, we generated ISO 19115 and 19139 standards-compliant metadata records for each article including the spatial references and make them available via the open source Python-based 'pycsw' catalogue server. Pycsw is an Open Geospatial Consortium (OGC) Catalogue Service for Web (CSW) and provides capabilities to store such metadata and make it searchable via spatial and standard metadata queries. For most parts the work was conducted on a Linux operating system, and with the support of 'Jupyter' Python notebooks.

We developed an exemplary web application that uses a plethora of open source JavaScript libraries, amongst them 'openlayers' and 'angular', that can query CSW-compatible catalogues. A user can now query and retrieve metadata records for journal articles and provide spatial search constraints via a map. The case study journals' websites provided textual search access to their more than 5000 research articles, but they could not be searched via spatial queries. Based on the demonstration for these journals we could show that in principle the automated detection of place based keywords is working and they can now be searched via spatial (e.g. BBOX) search queries, too. The web application was released on GitHub as open source, too.