doi.org/10.22498/pages.26.2.70

Automated extraction of spatiotemporal geoscientific data from the literature using GeoDeepDive

Jeremiah Marsicek[1], S.J. Goring[2], S.A. Marcott[1], S.R. Meyers[1], S.E. Peters[1], I.A. Ross[3], B.S. Singer[1] and J.W. Williams[2]

Although open data resources are growing, most scientific data remain "dark" (Heidorn 2008), available only in peer-reviewed literature, where the volume and lack of structure for these data resources presents challenges to data retrieval. GeoDeepDive is an expanding digital library with toolkits that allow automated acquisition and management of published and unpublished documents, supporting large-scale text and data mining of published, peer-reviewed journal articles (Peters et al. 2014; geodeepdive.org). Initial projects have demonstrated the utility of GeoDeepDive's large-scale synthetic geoscientific research (Peters et al. 2017), with new efforts underway.

GeoDeepDive provides a corpus of documents that contain a set of user-prescribed keywords (e.g. 'IRD' and 'Pliocene' or 'Pleistocene' or 'Holocene'). Users develop a set of rules to define the kinds of data they wish to retrieve (coordinates, measurements, etc.) from a subset of the matching publications, and write a test application. The application is deployed against the full GeoDeepDive corpus once a user has developed and tested their workflow on the data subset.

Initial work with GeoDeepDive – studying the dynamics of Northern and Southern Hemisphere ice sheets during the Quaternary – has allowed us to leverage publications focusing on ice-rafted debris (IRD). Assembling information from publications documenting IRD at marine drilling sites is a non-trivial task that has traditionally involved painstaking literature compilation (Hemming 2004). GeoDeepDive allowed us to discover and extract information by searching through 7.5 million publications across a range of publishers using an R workflow based on regular expressions and natural language-processing utilities. This work also allows us to develop a general workflow for GeoDeepDive, supporting others who might use it in their future research (Fig. 1).

Future directions

Our GeoDeepDive workflow allows us to extract and plot reliable latitude-longitude pairs from publications reporting IRD events (Fig. 1). We are building a spatial database of IRD events and beginning to extract event ages from the papers. Extracting temporal information from the unstructured peer-reviewed literature is a non-trivial but tractable task using regular expressions and string matching. We are also differentiating primary, original sources from secondary studies that include previously published data, and building a GitHub repository for open code development and sharing (github.com/EarthCubeGeoChron). Next steps include building summary maps of the location, finding specific named IRD events or the timing of IRD deposits found in cores, and continuing development of an R package (github.com/EarthcubeGeochron/geodiveR). The project will result in an IRD database that can provide a better characterization of Northern and Southern Hemisphere ice sheets over the last 5.3 million years. The R package that results from this work will consist of a general set of tools for querying space and time information from GeoDeepDive, allowing other researchers to simply import their own data using their own search logic and output coordinates and subsets of the text relevant to a researcher's particular questions.

An ongoing question in this broad-scale, data-mining project is to determine the appropriate points for human intervention and interpretation, one of many questions discussed at a recent GeoDeepDive user workshop in Madison, USA (geodeepdive.org/workshop2018). These points should be minimized for reasons of scalability, but some features may not be readily automated. Future advances will likely be powered by "centaur" systems combining the relative strengths of human- and machine-learning approaches, which will then provide the basis for new applications of machine-learning methods. We view the GitHub Repository and the R package as building blocks that will serve researchers across the geosciences and allied disciplines.

AFFILIATIONS

[1]Department of Geoscience, University of Wisconsin-Madison, USA
[2]Department of Geography, University of Wisconsin-Madison, USA
[3]Department of Computer Sciences, University of Wisconsin-Madison, USA

CONTACT

Jeremiah Marsicek: jmarsicek@wisc.edu

REFERENCES

Heidorn PB (2008) Library Trends 57: 280-299

Hemming SR (2004) Rev Geophys 42: RG1005

Peters et al. (2014) PLoS ONE 9: e113523

Peters et al. (2017) Geology 45: 487-490

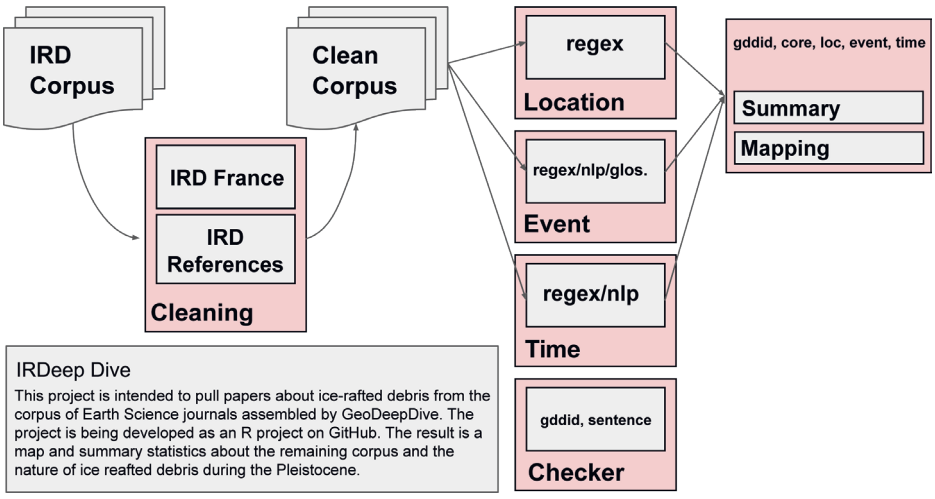

Figure 1: GeoDeepDive workflow used to build a corpus of documents that mention "ice-rafted debris" (IRD), screen a vetted set of the documents, and summarize the documents and relevant information (github.com/EarthcubeGeochron/geodiveR). 'Cleaning' = removing instances of IRD in the affiliations and references sections; regex = regular expression; nlp = natural language processing; glos. = glossary; Checker = a step to ensure sentences contain relevant IRD information; gddid = GeoDeepDive identification key.