A Model Workflow for Locating Pliocene and Pleistocene Ice-Rafted Debris using GeoDeepDive

Simon J. Goring[[1]](#footnote-20)

Jeremiah Marsicek[[2]](#footnote-21)

Shan Ye[[3]](#footnote-22)

John W. Williams[[4]](#footnote-23)

Stephen E. Meyers[[5]](#footnote-24)

Shanan E. Peters[[6]](#footnote-25)

Daven Quinn[[7]](#footnote-26)

Allen Schaen[[8]](#footnote-27)

Brad S. Singer[[9]](#footnote-28)

Shaun A. Marcott[[10]](#footnote-29)

Machine learning technology promises a more efficient and scalable approach to locating and aggregating data and information from the burgeoning scientific literature. Realizing this promise requires provision of applications, data resources, and the documentation of key tools. GeoDeepDive provides a digital library comprising over 11 million peer-reviewed documents and the computing infrastructure upon which to build and deploy search and text-extraction capabilities using regular expressions and natural language processing. Here we present a model GeoDeepDive workflow and accompanying R package to extract spatiotemporal information about site-level records in the geoscientific literature. We apply these capabilities to in a case study to generate a preliminary distribution of ice-rafted debris (IRD) records in both space and time. We use regular expressions and natural language-processing utilities to extract and plot reliable latitude-longitude pairs from publications containing IRD, as well extract age estimates from those publications. This workflow and R package provides researchers from the geosciences and allied disciplines a general set of tools for querying spatiotemporal information from GeoDeepDive for their own science questions.

## Introduction

Peer-reviewed papers communicate knowledge to their audiences through figures, text and tables. These elements have been refined over generations to efficiently communicate scientific insight to human readers, but the volume and variety of the peer-reviewed literature has challenged meta-analysis and efficient extraction of the underlying primary data. Efforts to improve the practice of data archiving in structured, sustainable data repositories are increasing (Sansone et al., 2019) as individuals and groups recognize the importance of data sharing and curation (Steering Committee, 2018). Despite these efforts, however, a large volume of data and information still exists exclusively in published form as text within manuscripts, embedded in tables, or graphically within figures. In response, new automated software tools are being developed to extract information directly from the scientific literature. Various fields are developing tools for automated extraction of meaningful information from the scientific literature, including natural language processing (NLP) and other forms of machine learning (ML), the vast majority of which is being developed and deployed to extract information from general and freely available content, like Twitter feeds and publication abstracts. Currently, the development of these new software tools has rapidly outpaced their deployment in the geosciences to digital libraries and infrastructures that can can address questions at scales not available to traditional meta-analyses.

GeoDeepDive (http://geodeepdive.org) is a digital library and computing system that currently contains over 11 million publications from multiple commercial and open-access content providers. Early versions of GeoDeepDive have been used to extract fossil occurrences from the scientific literature (Peters et al., 2014), e.g. to better understand the temporal patterns and possible drivers of stromatolite resurgences in the geological past (Peters et al., 2017). However, the newness of GeoDeepDive as a platform and the few available software tools to leverage it has limited its impact. Here we provide a sample workflow and accompanying R package that leads the reader through key, public elements of GeoDeepDive. As a case study, we retrieve documents on the distribution of ice-rafted debris (IRD) from the Pliocene to present and extract both geographic coordinates and temporal information. We choose to focus our effort on IRD because of the near uniqueness of the acronym in the geoscience literature and because IRD is almost exclusively restricted to ocean settings, thus simplifying our identification of false positives – occurrences of IRD that do not refer to ice-rafted debris – in our training dataset. IRD distribution in marine sediments provides a key constraint on cryosphere development, yielding insight into past climate evolution (*e.g.*, Hemming, 2004).

One implementation of GeoDeepDive uses sentences as the atomic unit, managing the sentence-level data within a PostgreSQL database. Each sentence within a paper is identified by a unique document id (gddid, an internally unique identifier to accommodate publications that may or may not have their own formal digital object identifier [DOI]) and a sentence number that is assigned and unique within the paper. A separate table relates gddids to publication metadata (e.g. title, journal, authors, etc.). Hence, GeoDeepDive workflows and the individual steps within these workflows effectively operate at two distinct levels: the sentence level and the document level. Because GeoDeepDive (GDD) also provides unique IDs for each journal and links these to the sentence IDs, journal-level analytics are possible. GDD makes use of Stanford NLP (Manning et al., 2014), so it is also possible to obtain word-level analysis using indexing within sentences. For this paper, we focus on “sentence”- and “document”-level analytics, because of our focus on extracting geospatial and temporal information for individual records.



**Figure 1**: *Workflow used to go from a list of documents that mention ice-rafted debris (IRD; IRD is the actual search string in this case) and (Pliocene or Pleistocene or Holocene) to a vetted set of the documents, and finally a summary of the documents and relevant information.*

This paper presents a sample workflow, intended to provide meaningful but preliminary results, with the main goal of illustrating the potential of GDD, and showing potential users how GDD can be used to extract information from text. In this example workflow, we will identify papers with mentions of ice-rafted debris (IRD) in the Pliocene and Pleistocene (Fig. 1), extract space and time coordinates using an R Toolkit called geodiveR (http://github.com/EarthCubeGeochron/geodiveR), and store the data and code in a GitHub repository (http://github.com/EarthCubeGeoChron).

Many publications document the existence of IRD at the level of individual marine drilling sites, but assembling this information across publications into large-scale mapped syntheses is a non-trivial task that has traditionally taken years of painstaking literature compilation (Heinrich, 1988; Hemming, 2004; W. Ruddiman, 1977; Stern and Lisiecki, 2013). A comprehensive, accurate database of IRD deposits and their spatial distribution extracted with software from the published scientific literature can help the scientific community better understand and characterize ice sheet dynamics over the last 5.3 million years, ideally leading to a better understanding of how glaciers respond to changes in climate and ocean circulation.

The goal of this paper is not to remove the expert (*e.g*., sedimentologist, paleoceanographer) from decision-making processes, but to provide an initial framework to use the DeepDive infrastructure to more efficiently perform research tasks through machine learning. Various subtleties and complexities persist that are not yet tractable to machine learning. For example, ice-rafted debris is part of a complex of sedimentary deposits within ocean sediments, including iceberg, ice shelf, and sea ice rafted debris (Powell, 1984). Differences between the processes resulting in sediment entrainment and deposition between these types of sediment may result in issues of interpretation (Andrews, 1998). Sedimentological features can be used to differentiate particle sizes and shapes to offer a better understanding of the sources of sediments identified as ice-sourced (St John et al., 2015), and thus provide a more complete picture of the processes that led to deposition. The geospatial and temporal information retrieved by GDD require vetting.

Nonetheless, by providing a comprehensive corpus of documents, with identified publications, timings, and locations for identified deposits it becomes possible to identify potential outliers or misidentified samples, to re-sample existing physical cores where there may be uncertainty in identification, and ultimately, to generate a more complete model of sea-ice patterns during the Pleistocene. As a first step forward, this paper provides a model workflow to be carried out with the DeepDive infrastructure. We begin with a general walkthrough of the analytical steps and various considerations that arise at each stage, then move to a specific walkthrough that focuses on identifying papers with IRD records from the Pliocene and Pleistocene, extracts spatial and temporal coordinates, and maps the returned results. We show that a relatively simple framework is already able to recover a substantial body of useful information that can inform further data processing, cleaning, and interpretation that would be part of a more formal analysis of patterns of paleo-oceanographic patterns during the Pleistocene.

## Workflow Overview

### Initial Returns and RegEx

Processing the entirety of the documents within GDD is time consuming because the body of accesible papers (the corpus) contains over 11 million peer-reviewed publications. For any given goal, such as finding all instances of IRD, only a small fraction of the total corpus is relevant. Keywords are a common first solution to reducing data volume. Particular keywords within a document can be used to identify the subset of all documents that are potentially relevant. Straightforward string matching using keywords can take advantage of the GDD public API (https://geodeepdive.org/api). Here, we chose terms that would return a sufficient breadth of documents for this trial study. We used the acronym “IRD”, as well as constraints on geologic time intervals, “Holocene”, “Pleistocene” and “Pliocene”. In our example dataset, this returned 5,315 total documents; Figure 2.

Useful information about these terms can be derived from the “snippets” route of GeoDeepDive’s API. Snippets harnesses an ElasticSearch index spanning the full text of all PDFs that have a “native” text layer (*i.e*., PDFs with searchable text in them already, which constitutes the vast majority of PDFs distributed by journal publishers):

https://geodeepdive.org/api/snippets?term=IRD&full\_results=true

The response to this API call is a JSON object that indicates the total number of “hits” of the term (n = 35754 as of 2020-03-26), basic bibliographic citation information for each document containing the term, including a link to the original PDF distributed by the publisher, and a “snippet” of text around mentions of the term in the full text of the document:

## [  
## {  
## "pubname": "South African Geographical Journal",  
## "publisher": "Taylor and Francis",  
## "\_gddid": "5946b1c8cf58f13cac0191e7",  
## "title": "Reviews/Resensies",  
## "doi": "10.1080/03736245.1976.10559569",  
## "coverDate": "1976 04",  
## "URL": "http://www.tandfonline.com/doi/abs/10.1080/03736245.1976.10559569",  
## "authors": "Young, Bruce; Davies, R. J.; Hart, G. H. T.",  
## "highlight": [  
## "materials for course work in the second part. The th,<em class=\"hl\">ird<\/em> and fourth parl~ of the book. 0n the other hand,"  
## ]  
## }  
## ]  
##

By default, the matching terms are highlighted with HTML tags (to remove the tags, the parameter &clean=true can be added to the URL for the snippet API). Because of the large number of results in the case above, the results include a link to the next page of documents containing the term, allowing the user to scroll through large sets of results.

The GeoDeepDive snippet API also supports searches that combine multiple terms. The following API call, which returns only papers that contain both “IRD” and “Pleistocene” returns 4208 hits as of 2020-03-26, far fewer than the more expansive search.

https://geodeepdive.org/api/snippets?term=IRD,Pleistocene&inclusive=TRUE&full\_results=true

The GeoDeepDive API, however, is not designed to provide full functionality, but rather is designed to be deployed in user-constructed applications. Hence, the GDD API is suitable only for limited data extraction tasks. More powerful analyses require analysis of the PostgreSQL representation of the document text data.

Here, text matching and data extraction from the retrieved body of papers in PostgreSQL uses existing PostgreSQL text and string functions, plus regular expression matching in R using the stringr package (Wickham, 2019a). The use of the Stanford NLP library also allows GDD workflows to take advantage of parts-of-speech tagging, and more advanced NLP tools, but these capabilities are not employed in this demonstration workflow.

### Subsetting and Cleaning

We begin our analysis with a subset of the data, consisting of 150 papers, that was pared down from the total set of the corpus, using the keyword constraints described above. The subset of papers may still include papers that are not appropriate (i.e. IRD may refer to something other than ice-rafted debris). To obtain a training dataset, we run a second round of same text matching, using the same keywords and rules used at the document level, but now enforced at the sentence level. For example, “IRD” must be located in a sentence with another term (e.g., “IRD” and “Holocene”). These additional rules restrict the total list returned to 81 documents for which any sentence contains a match to the keyword.

Searching for IRD as a keyword retrieves articles that use IRD as an acronym for Ice-Rafted Debris, but it also, for instance, retrieves articles mentioning the French Research **I**nstitute of **R**esearch for **D**evelopment. Throughout this paper we will refer to *rules*; generally these are statements that can resolve to a boolean (TRUE/FALSE) output. So for example, within our subset we could search for all occurrences of IRD and CNRS:

sentence <- "this,is,a,IRD,and,CNRS,sentence,we,didnt,want,."  
stringr::str\_detect(sentence, "IRD") & !stringr::str\_detect(sentence, "CNRS")

This statement will evaluate to TRUE if IRD appears in a sentence without CNRS. If we apply this sentence-level test at the document level (any(test == TRUE)) we can estimate which papers are most likely to have the correct mention of IRD for our purposes. This then further reduces the number of papers (and sentences) for our training dataset.

## Extracting Data

After cleaning and subsetting, we develop a series of tests and workflows to iteratively extract information. In many cases this requires further text matching, and packages in R such as stringr were useful for accomplishing this task. Additional support can come from the NLP output that can be generated for the data. In all of these cases, we generate clear rules to be tested, and then apply them to the document.

Because understanding both the IRD distribution in ocean sediments and the timing of the deposition of IRD through the Pliocene and Pleistocene is critical for interpreting past ice dynamics, spatial coordinates and geochronologic constraints of the IRD deposits need to be identified within the paper. As with the earlier cleaning process, any paper that contains neither spatial coordinates or ages, or one but not the other, is here deemed not of interest. Less restrictive searches could be defined.

Extracting the spatial location and age of an IRD occurrence within a paper that contains “IRD”, however, is not sufficient. We need to be able to distinguish between an age related to the event we are interested in versus an age reported in a paper for some other reason. So, again, we must develop general rules that allow distinguishing of all ages from ages of interest, and all spatial locations from spatial locations of interest.

## Exploratory Itteration

There are a number of reasons to continue to refine the rules used in this workflow to discover data. First, extraction of text from the PDF and optical character recognition (OCR) is not always accurate, so that some sentences and words are parsed incorrectly. This problem is particularly acute for geographic coordinates. Second, many words have multiple meanings, leading to false positives if only string-matching is used, as the IRD example illustrates. Third, semantic terms and concepts often vary subtly within and among disciplines and journals. As a simple example, if we were interested in retrieving paleoecological information we would need to know that *paleoecology* and *palaeoecology* refer to essentially identical concepts. Similarly, ice rafted debris may also be referred to as sand sized layers in the marine context (e.g. W. F. Ruddiman, 1977), while a paleoceanographer might want careful separation among different kinds of IRD, e.g. iceberg-rafted debris (IBRD), ice shelf rafted debris (ISRD), or sea ice rafted debris (SIRD). Fourth, the context and placement of words matters. For example, temporal information like ‘Holocene’ and ‘Pliocene’ may be found in the Methods, where they refer to marine core locations, or in the Discussion, where they might refer to global climate trends.

Some potential pitfalls include:

* OCR matching - commonly mistaken letters (O, Q, o)
* Age reporting variety in units or time datum (e.g., kyr vs. ka)
* Age basis variety (e.g. radiocarbon years vs. calendar years)
* GDD sentence construction

Repeatedly reviewing matches at the sentence level and at the document level (i.e., “Why did this match or why didn’t this paper return a match?”), then refining the workflow rule-sets accordingly, is critical to developing a clear workflow and high-value corpus. In many cases, beginning with very broad tests and slowly paring down to more precise tests is an appropriate approach. In this case, tools like RMarkdown are very helpful for interactive data exploration, using packages like DT (Xie et al., 2020) and leaflet (Cheng et al., 2019). We can assess the distribution of age-like elements within a paper and determine if they match with our initial expectations (e.g. “Why does Debris fields in the Miocene contain Holocene-aged matches?”; “Why does a paper about Korea report locations in Belize?”). Depending on the success of the algorithm, the tests can be revised and the process repeated until an acceptable match is found.

### Reproducible and Validated Workflows for a Dynamic Literature

As the workflow develops and refines, we can begin to report on patterns and findings. Some of these may be semi-qualitative (e.g. “The majority of sites are dated to the LGM”), while others may involve statistical analysis (e.g., “The presence of IRD increases after the Mid-Pliocene Transition (*p<0.05*)”). In an analysis where the underlying dataset is static or a version has been frozen, it is reasonable to develop a paper and report these findings.

However, the publication database in GDD is far from static; the GDD infrastructure acquires more than 10,000 papers per day from multiple sources. Given this, it is likely that some patterns will change over time as more information is brought to bear. For example, a new ocean drilling campaign might reveal new insights into the spatiotemporal distribution of IRD, or the addition of new records may reveal previously undiscovered search artifacts within the publication record. For this reason the use of assertions or testable statements that can be evaluated to TRUE or FALSE within the workflow become critically important. In R we can use the assertthat package (Wickham, 2019b) to provide support for test-driven development and assertions within the workflow.

Test-driven development is common in software development. As developers create new features, a good practice is to first develop tests for the features, to ensure that feature behavior matches expectations. The analogy in our scientific workflow is that findings are features, and as we report them, we want to be assured that those findings are valid. In R the assertthat package provides a tool for testing statements, and providing robust feedback through custom error messages (Wickham, 2019b).

howmany\_dates <- all\_sentences %>%  
 mutate(hasAge = stringr::str\_detect(words, "regular expression for dates")) %>%  
 group\_by(gddid) %>%  
 summarise(age\_sentences = any(hasAge),  
 n = n())  
  
# We initially find that less than 10% of papers have dates in them, and report that as an important finding in the paper.  
  
percent\_ages <- sum(howmany\_dates$age\_sentences) / nrow(howmany\_dates)  
  
assertthat::assert\_that(percent\_ages < 0.1, msg = "More than 10% of papers have ages.")

**Text Box 1**. *From sentences returned by GDD (all\_sentences) create a new column called ‘hasAge’ that is a boolean variable to test whether an age is found. Once the variable has been created, group the data to test whether any of the papers (each with a unique gddid) has an age reported in the paper, and then count the total number of reported ages within the paper. If we expect that most papers won’t have reported ages, we can create an assertion, in this case that only 10% of papers have a reported age. If this assertion fails, then a custom error message is returned.*

With this workflow overview we now have mapped out an iterative process that is also responsive to the underlying data. We have developed clear tests under which our findings are valid. We can create a document that combines our code and text in an integrated manner, supporting FAIR Principles (Wilkinson et al., 2016), and supporting the next generation of reproducible research. In the following section we run through this workflow in detail.

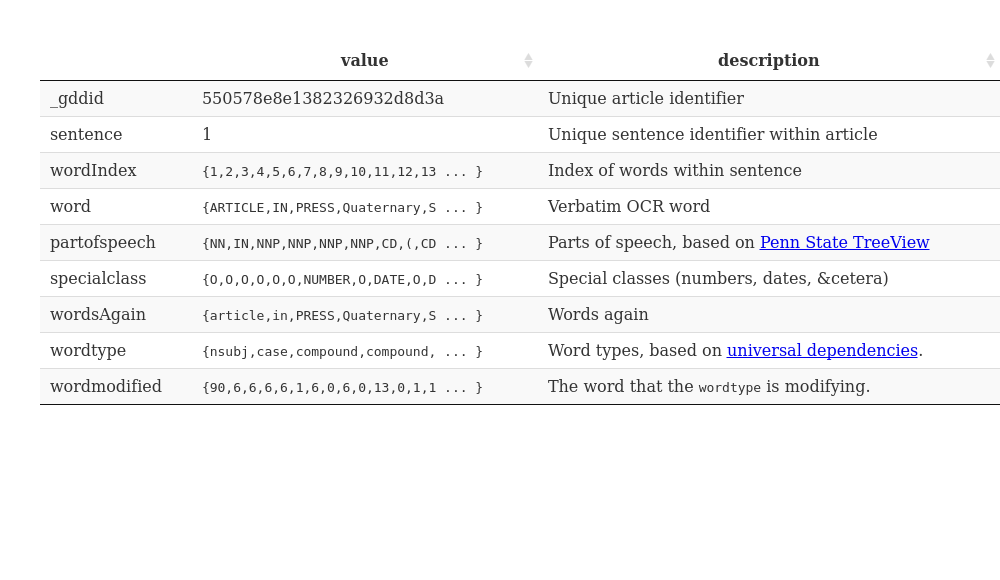
## Ice Rafted Debris Case Study and Discussion

### Finding Spatial Matches

To begin, we load the packages that will be used, and then import the data:

library(geodiveR)  
library(ggplot2)  
library(jsonlite) # (Ooms et al., 2018)  
library(readr) # (Wickham et al., 2018)  
library(dplyr) # (Wickham et al., 2019)  
library(stringr)  
library(leaflet)  
library(purrr) # (Henry et al., 2019)  
library(DT)  
library(assertthat)  
  
sourcing <- list.files('R', pattern = ".R$", full.names = TRUE) %>%  
 map(source, echo = FALSE, print = FALSE, verbose = FALSE)  
  
publications <- fromJSON(txt = 'input/bibjson', flatten = TRUE)  
  
if(!file.exists('input/sentences\_nlp352')){  
 data(nlp)  
 full\_nlp <- nlp  
 rm(nlp)  
 colnames(full\_nlp) <- c('\_gddid', 'sentence', 'wordIndex',  
 'word', 'partofspeech', 'specialclass',  
 'wordsAgain', 'wordtype', 'wordmodified')  
} else {  
  
 full\_nlp <- readr::read\_tsv('input/sentences\_nlp352',  
 trim\_ws = TRUE,  
 col\_names = c('\_gddid', 'sentence', 'wordIndex',  
 'word', 'partofspeech', 'specialclass',  
 'wordsAgain', 'wordtype', 'wordmodified'))  
}  
  
# uses the clean\_corpus.R function within geodiveR  
# (implements the IRD search-and-refinement rules mentioned  
# above in 'Exploratory Iteration')  
  
nlp\_clean <- clean\_corpus(x = full\_nlp, pubs = publications)  
  
nlp <- nlp\_clean$nlp

This code produces an output object that includes a key for the publication (\_gddid, linking to the publications variable), the sentence number of the parsed text, and then both the parsed text and some results from natural language processing. We also obtain a list of gddids to keep or drop given the regular expressions we used to find instances of IRD in the affiliations or references sections of the papers. This analysis returns 1 publications:



**Table 1.** Columns returned by GeoDeepDive for use in text matching and extraction, along with an example of each column entry.

### Extracting Spatial Coordinates

We are interested in using GDD to obtain site coordinates for locations that contain IRD data spanning the Pliocene and Pleistocene. The goal is to provide relevant site information for use in meta-analysis, or for comparing results to existing geographic locations from the relevant geocoded publications and then linking back to the publications using DOIs.

To obtain geographical coordinates from the paper we must consider several potential issues. The first is that not all coordinates will necessarily refer to an actual ocean core. We may also, inadvertently, find numeric objects that appear to be coordinates, but are in fact simply numbers. Therefore, we must identify what exactly we think coordinates might look like and build a regular expression (or set of regular expressions) to accurately extract these values. Since we will process DMS coordinates differently than DD coordinates, we generate two regular expressions:

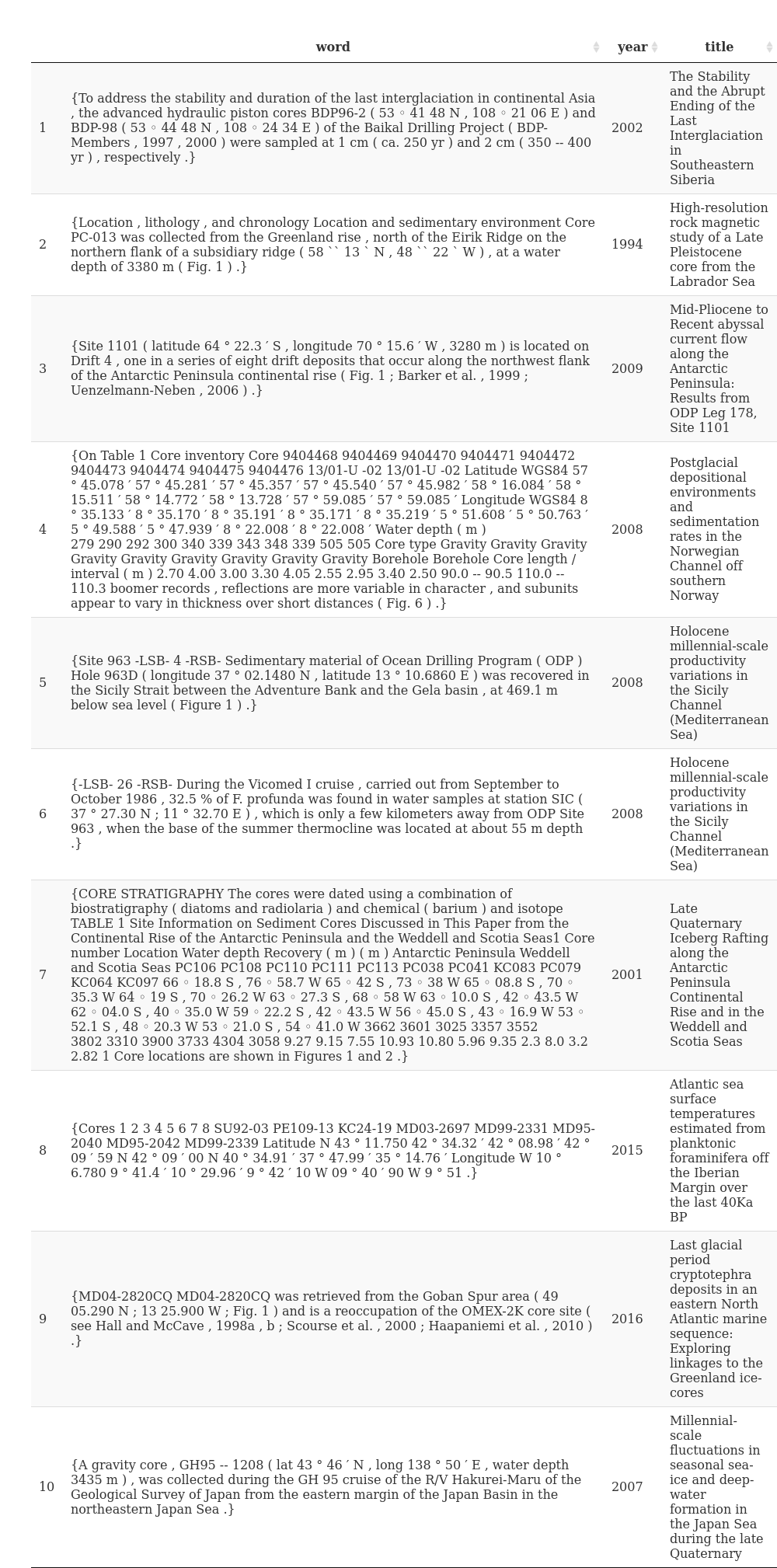
dms\_regex <- "[\\{,]([-]?[1]?[0-9]{1,2}?)(?:(?:,[°◦o],)|(?:[O])|(?:,`{2},))([1]?[0-9]{1,2}(?:.[0-9]\*)),[′'`]?[,]?([[0-9]{0,2}]?)[\"]?[,]?([NESWnesw]?),"  
  
dd\_regex <- "[\\{,][-]?[1]?[0-9]{1,2}\\.[0-9]{1,}[,]?[NESWnesw],"

These regular expressions allow for negative or positive coordinate systems, that may start with a 1, and then are followed by one or two digits ({1,2}). From there we see differences in the structure, reflecting the need to capture the degree symbols, or, in the case of decimal degrees, the decimal component of the coordinates. The regular expressions are more rigorous (i.e., have fewer matching options) for the decimal degrees than for DMS coordinates. DMS coordinates reported in documents may be interpreted in non-standard ways by OCR.

The regex commands were constructed to work with the stringr package (Wickham, 2019c), so that we obtain five elements from any match, including the full match, the degrees, the minutes, the seconds (which may be an empty string), and the quadrant (NESW).

degmin <- str\_match\_all(nlp$word, dms\_regex)  
decdeg <- str\_match\_all(nlp$word, dd\_regex)

We expect that all coordinates are reported as pairs within sentences, and so we are most interested in finding all sentences that contain pairs of coordinates. We start by finding the publications with sentences that have coordinate pairs:



**Table 2**. Sample sentences from the IRD corpus that contain matches to the coordinate regular expression rule.

Even here, we can see that many of these matches work, but that some of the matches are incomplete. There appears to be a much lower proportion of sites returned than we might otherwise expect. Given that there are 81 articles in the NLP dataset with matches to IRD related terms, it is surprising that only 20 appear to support regex matches to coordinate pairs. We would expect that the description of sites or locations using coordinate pairs should be common practice. The observed outcome is likely to be, in part, an issue with the OCR/regex processing. A next iterative step would be to review potential matches more thoroughly to find additional methods of detecting the coordinate systems.

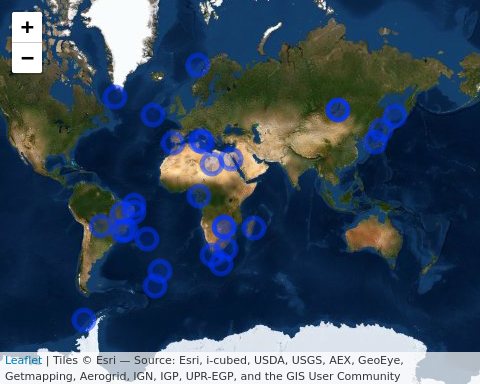
### Converting Coordinates

Given the geographic coordinate strings, we need to be able to transform them to reliable latitude and longitude pairs with sufficient confidence to actually map the records. These two functions convert the GeoDeepDive (GDD) word elements pulled out by the regular expression searches into decimal degrees that can account for reported locations.

convert\_dec <- function(x, i) {  
  
 drop\_comma <- gsub(',', '', x) %>%  
 substr(., c(1,1), nchar(.) - 1) %>%  
 as.numeric %>%  
 unlist  
  
 domain <- (str\_detect(x, 'N') \* 1 +  
 str\_detect(x, 'E') \* 1 +  
 str\_detect(x, 'W') \* -1 +  
 str\_detect(x, 'S') \* -1) \*  
 drop\_comma  
  
 publ <- match(nlp$`\_gddid`[i], publications$`\_gddid`)  
  
 point\_pairs <- data.frame(sentence = nlp$sentence[i],  
 string = nlp$word[i],  
 lat = domain[str\_detect(x, 'N') | str\_detect(x, 'S')],  
 lng = domain[str\_detect(x, 'E') | str\_detect(x, 'W')],  
 publications[publ,],  
 stringsAsFactors = FALSE) %>%  
 rename(gddid = X\_gddid)  
  
 return(point\_pairs)  
}  
  
convert\_dm <- function(x, i) {  
  
 # We use the `i` index so that we can keep the coordinate outputs from the  
 # regex in a smaller list.  
 dms <- data.frame(deg = as.numeric(x[,2]),  
 min = as.numeric(x[,3]) / 60,  
 sec = as.numeric(x[,4]) / 60 ^ 2,  
 stringsAsFactors = FALSE)  
  
 dms <- rowSums(dms, na.rm = TRUE)  
  
 domain <- (str\_detect(x[,5], 'N') \* 1 +  
 str\_detect(x[,5], 'E') \* 1 +  
 str\_detect(x[,5], 'W') \* -1 +  
 str\_detect(x[,5], 'S') \* -1) \*  
 dms  
  
 publ <- match(nlp$`\_gddid`[i], publications$`\_gddid`)  
  
 point\_pairs <- data.frame(sentence = nlp$sentence[i],  
 string = nlp$word[i],  
 lat = domain[x[,5] %in% c('N', 'S')],  
 lng = domain[x[,5] %in% c('E', 'W')],  
 publications[publ,],  
 stringsAsFactors = FALSE) %>%  
 rename(gddid = X\_gddid)  
  
 return(point\_pairs)  
}

Once we are done converting coordinate strings to reliable latitude and longitude pairs, we need to apply those functions to the set of records we’ve extracted to build a composite table:

coordinates <- list()  
coord\_idx <- 1  
  
for (i in 1:length(decdeg)) {  
 if ((length(decdeg[[i]]) %% 2 == 0 |  
 length(degmin[[i]]) %% 2 == 0) & length(degmin[[i]]) > 0) {  
  
 if (any(str\_detect(decdeg[[i]], '[NS]')) &  
 sum(str\_detect(decdeg[[i]], '[EW]')) == sum(str\_detect(decdeg[[i]], '[NS]'))) {  
 coordinates[[coord\_idx]] <- convert\_dec(decdeg[[i]], i)  
 coord\_idx <- coord\_idx + 1  
 }  
 if (any(str\_detect(degmin[[i]], '[NS]')) &  
 sum(str\_detect(degmin[[i]], '[EW]')) == sum(str\_detect(degmin[[i]], '[NS]'))) {  
 coordinates[[coord\_idx]] <- convert\_dm(degmin[[i]], i)  
 coord\_idx <- coord\_idx + 1  
 }  
 }  
}  
  
coordinates\_df <- coordinates %>% bind\_rows %>%  
 mutate(sentence = gsub(',', ' ', sentence)) %>%  
 mutate(sentence = str\_replace\_all(sentence, '-LRB-', '(')) %>%  
 mutate(sentence = str\_replace\_all(sentence, '-RRB-', ')')) %>%  
 mutate(sentence = str\_replace\_all(sentence, '" "', ','))  
  
coordinates\_df$doi <- coordinates\_df$identifier %>% map(function(x) x$id) %>% unlist  
  
leaflet(coordinates\_df) %>%  
 addProviderTiles(providers$Esri.WorldImagery) %>%  
 addCircleMarkers(popup = paste0('<b>', coordinates\_df$title, '</b><br>',  
 '<a href=https://doi.org/',  
 coordinates\_df$doi,'>Publication Link</a><br>',  
 '<b>Sentence:</b><br>',  
 '<small>',gsub(',', ' ', coordinates\_df$string),  
 '</small>'))



**Figure 3**: *Leaflet map depicting sites that mention IRD and contain coordinate information and an IRD event. Each dot can be clicked to pull up: the title of the paper, the link to the publication, the sentence(s) containing spatial coordinates and other relevant IRD information. Out of the 150 papers in the test dataset, we found 11 papers with extractable spatial coordinates, comprising a total of 30 coordinate pairs (References included in leaflet map include: [Baeten et al. (2014);Cofaigh et al. (2001);Stoner et al. (1994);Simon et al. (2013);Seki et al. (2005);Rosell-Melé et al. (1997);Rea et al. (2016);Prokopenko et al. (2002);Kaboth et al. (2016);Ikehara and Itaki (2007);incarbona2008holocene].*

After cleaning and subsetting the corpus, we find 11 papers with 30 coordinate pairs out of 150 documents in the IRDDive test dataset (Figure 3). This test suggests further improvements that can be made to the current methods. First, in some cases, we find papers where IRD is simply referenced, and the paper does not report primary data. Second, in some cases we find IRD but no coordinates or other core metadata; some papers simply do not contain coordinate information. Third, some papers mention IRD in the core data for continental cores (see Fig. 3 Central Asia location). These might possibly be valid instances of IRD in lacustrine deposits (Smith, 2000), or might represent papers that mention IRD without representing primary data. One solution would be to cross-reference the returned IRD coordinates with the location of continents (past or present) and remove coordinate pairs that fall within the continental boundaries. A fourth and last step would be to further refine the regex to obtain additional mentions of IRD, e.g. as ‘IRD-rich layers’, ‘IRD-rich deposits’, ‘Heinrich layers’, etc.

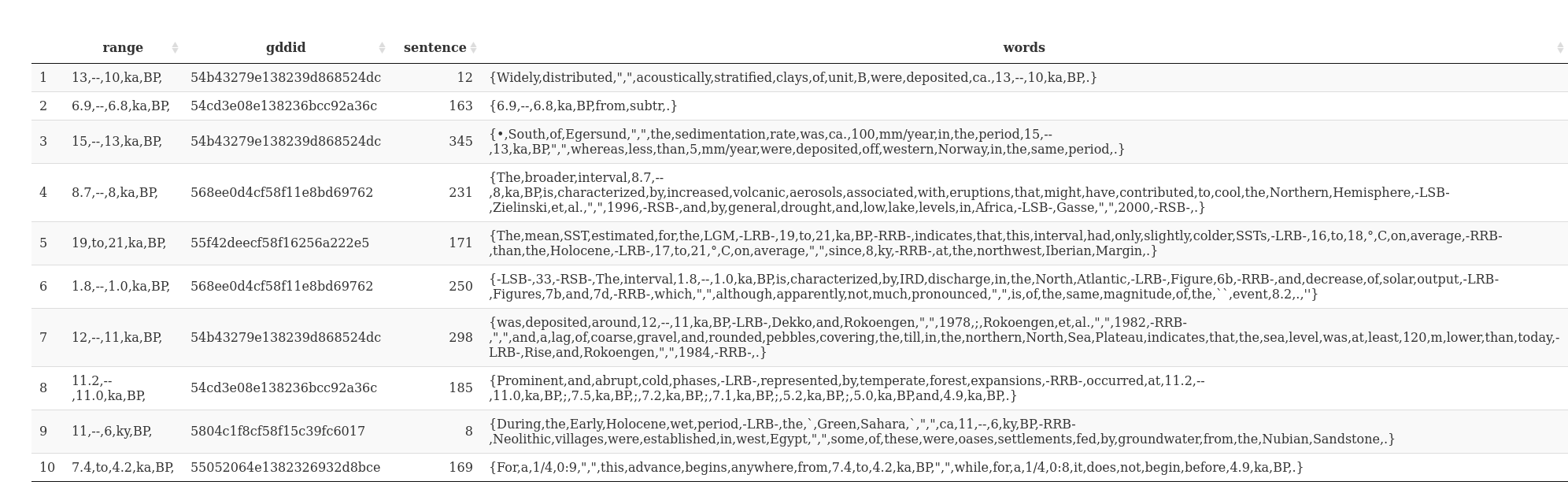
### Extracting ages and age ranges

The next step is to extract ages and age ranges that may be associated with IRD events. This requires building regular expressions that pull dates with many different naming conventions (e.g., years BP, kyr BP, ka BP, a BP, Ma BP, *etc*.)

is\_date <- str\_detect(nlp$word, ",BP,")  
is\_range <- str\_detect(nlp$word, "(\\d+(?:[.]\\d+)\*),((?:-{1,2})|(?:to)),(\\d+(?:[.]\\d+)\*),([a-zA-Z]+,BP),")  
  
date\_range <- str\_extract\_all(nlp$word,  
 "(\\d+(?:[.]\\d+)\*),((?:-{1,2})|(?:to)),(\\d+(?:[.]\\d+)\*),([a-zA-Z]+,BP),") %>%  
 map(function(x) {  
 data.frame(range = paste(x, collapse = ";"),  
 stringsAsFactors = FALSE)  
 }) %>%  
 bind\_rows() %>%  
 mutate(gddid = nlp$`\_gddid`,  
 sentence = nlp$sentence,  
 words = nlp$word)  
  
# Extract the ranges within sentences. If there are multiple ranges then  
# collapse them into a single string using a semi-colon.  
number <- str\_extract\_all(nlp$word,  
 ",(\\d+(?:[\\.\\s]\\d+){0,1}),.\*?,yr.\*?,") %>%  
 map(function(x) {  
 data.frame(age = paste(x, collapse = ";"),  
 stringsAsFactors = FALSE)  
 }) %>%  
 bind\_rows() %>%  
 mutate(gddid = nlp$`\_gddid`,  
 sentence = nlp$sentence) %>%  
 left\_join(x = ., y = date\_range, by = c('gddid', 'sentence'))

When we apply the code to the documents in the IRD corpus we begin to see patterns of dates (Table 3). A limitation of the regular expression presented here is that the regular expression for the variable is\_range, would only match BP ages prepended by a scale (a, ka, Ma). As currently defined, is\_range only allows one term between matched numbers and the string BP. Thus more detailed methods would need to be used to capture all age descriptors and types. Ultimately, multiple matching terms are likely required to find the breadth of age terms.

date\_range %>%  
 arrange(sample(1:nrow(date\_range))) %>%  
 filter(!range == "") %>%  
 distinct(range, .keep\_all=TRUE) %>%  
 datatable(escape = FALSE, rownames = TRUE, options = list(dom='t'))

 **Table 3**. *A sample of recovered age ranges reported in the NLP corpus that is associated with IRD records based on regular expression captures. The sentence number within a paper provides some indication of whether the age range is identifying broader global patterns of change (age ranges defined in the paper’s Introduction), or elements specific to the current study.*

At this stage of the workflow, we have successfully identified instances of dates in the papers where there are references to IRD (Table 1), but we do not know if these ages are associated with specific IRD events. Nonetheless, when we map the data (Figure 3), we observe clusters that are consistent with IRD distributions found in the primary literature but with some notable exceptions (e.g., over land). A next step would be to begin to match the dates to specific units.



**Figure 4**. *Incidence of time periods covered by ranges described in papers (e.g., “5 - 7 ka BP”). Individual papers may report multiple time ranges, and not all time ranges reported may be directly related to Ice Rafted Debris events.*

One avenue for further exploration is to use the sentence position within the document to provide context for a retrieved geographic coordinate or age. For example, the distribution of coordinates within papers shows a well defined pattern (Figure 4). Coordinates are generally presented in the Abstract, Introduction or Methods section, but rarely elsewhere. This differs from the distribution of ages and age ranges within papers, which appear throughout papers, although age ranges tend to be most frequent in the results, discussion and conclusions. Age information also often appears in the titles of papers presented in the References section. This is a promising avenue for further study.

Of course, while the methods presented here do extract ages and geographic locations, it should be clear that this alone is not sufficient to fully understand process. Rather, this demonstration shows how these new capabilities enable an unprecedentedly detailed and powerful searches into the scientific literature, accelerate the pace and scale of scientific synthesis and insight.



**Figure 5**: *Relative location of ages and spatial coordinates reported in GDD documents (i.e., are ages or spatial coordinates generally located at the beginning, middle, or end of a paper?). The line segments above represent the approximate position of the term “Abstract”, “Methods”, etc. in papers, not the extent of those terms.*

## 8. Conclusions

Here we have provided a model workflow to obtain coordinate and age information associated with IRD deposits, a GitHub Repository for open code development and sharing, and an R toolkit (http://github.com/EarthCubeGeochron/geodiveR). The specific case study showcased here is designed to help researchers study the dynamics of ice sheets over the last 5 million years, while GeoDeepDive and the post-processing analytical workflows shown here will be helpful to those wanting to query for space and time information using GeoDeepDive. This will allow other researchers to import their own data from their own search logic, and output text and coordinates relevant to a researcher’s question. The GitHub Repository and R package can act as building blocks that serve researchers not only in the geosciences, but allied disciplines as well.

## Data Availability

All data are publicly available at http://github.com/EarthCubeGeoChron/IRDDive and http://github.com/EarthCubeGeochron/geodiveR.

## Acknowledgements

Funding for this research was provided by the National Science Foundation (#EAR-1740694 to Goring, Marcott, Meyers, Peters, Singer, and Williams). The authors also acknowledge I. Ross, J. Czaplewski, A. Zaffos, E. Ito, and J. Husson for helpful discussions.

## Author Contributions

All authors contributed to study design. J.M., S.Y., and S.J.G. performed analysis and developed the GitHub repository and R package. J.M., S.J.G., and S.A.M. wrote the paper with input from the rest of the authors.

Andrews, J.T., 1998. Abrupt changes (Heinrich events) in late Quaternary North Atlantic marine environments: A history and review of data and concepts. Journal of Quaternary Science, 13, 3–16.

Baeten, N.J., Laberg, J.S., Vanneste, M., Forsberg, C.F., Kvalstad, T.J., Forwick, M., Vorren, T.O., Haflidason, H., 2014. Origin of shallow submarine mass movements and their glide planes—sedimentological and geotechnical analyses from the continental slope off northern norway. Journal of Geophysical Research: Earth Surface 119, 2335–2360.

Cheng, J., Karambelkar, B., Xie, Y., 2019. leaflet: Create interactive web maps with the JavaScript ’leaflet’ library.

Cofaigh, C.Ó., Dowdeswell, J.A., Pudsey, C.J., 2001. Late Quaternary iceberg rafting along the Antarctic Peninsula continental rise and in the Weddell and Scotia Seas. Quaternary Research 56, 308–321.

Heinrich, H., 1988. Origin and consequences of cyclic ice rafting in the northeast atlantic ocean during the past 130,000 years. Quaternary research 29, 142–152. doi:[10.1016/0033-5894(88)90057-9](https://doi.org/10.1016/0033-5894(88)90057-9)

Hemming, S.R., 2004. Heinrich events: Massive late pleistocene detritus layers of the north atlantic and their global climate imprint. Reviews of Geophysics 42. doi:[10.1029/2003rg000128](https://doi.org/10.1029/2003rg000128)

Ikehara, K., Itaki, T., 2007. Millennial-scale fluctuations in seasonal sea-ice and deep-water formation in the japan sea during the late quaternary. Palaeogeography, Palaeoclimatology, Palaeoecology 247, 131–143. doi:[10.1016/j.palaeo.2006.11.026](https://doi.org/10.1016/j.palaeo.2006.11.026)

Kaboth, S., Bahr, A., Reichart, G.-J., Jacobs, B., Lourens, L.J., 2016. New insights into upper mow variability over the last 150 kyr from iodp 339 site u1386 in the gulf of cadiz. Marine geology 377, 136–145. doi:[10.1016/j.margeo.2015.08.014](https://doi.org/10.1016/j.margeo.2015.08.014)

Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D., 2014. The stanford corenlp natural language processing toolkit, in: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 55–60. doi:[10.3115/v1/P14-5010](https://doi.org/10.3115/v1/P14-5010)

Peters, S.E., Husson, J.M., Wilcots, J., 2017. The rise and fall of stromatolites in shallow marine environments. Geology 45, 487–490. doi:[10.1130/G38931.1](https://doi.org/10.1130/G38931.1)

Peters, S.E., Zhang, C., Livny, M., Ré, C., 2014. A machine reading system for assembling synthetic paleontological databases. PLoS one 9. doi:[10.1371/journal.pone.0113523](https://doi.org/10.1371/journal.pone.0113523)

Powell, R.D., 1984. Glacimarine processes and inductive lithofacies modelling of ice shelf and tidewater glacier sediments based on Quaternary examples. Marine Geology 57, 1–52. doi:[10.1016/0025-3227(84)90194-4](https://doi.org/10.1016/0025-3227(84)90194-4)

Prokopenko, A.A., Karabanov, E.B., Williams, D.F., Khursevich, G.K., 2002. The stability and the abrupt ending of the last interglaciation in southeastern siberia. Quaternary Research 58, 56–59. doi:[10.1006/qres.2002.2329](https://doi.org/10.1006/qres.2002.2329)

Rea, D.K., Moore, T.C., Hassold, N., Pluijm, B. van der, 2016. Do magnetic fabrics of marine deposits preserve orbital forcing? A test case in the southern ocean, antarctic peninsula. Lithosphere 8, 751–756. doi:[10.1130/L566.1](https://doi.org/10.1130/L566.1)

Rosell-Melé, A., Maslin, M.A., Maxwell, J.R., Schaeffer, P., 1997. Biomarker evidence for “heinrich” events. Geochimica et Cosmochimica Acta 61, 1671–1678. doi:[10.1016/S0016-7037(97)00046-X](https://doi.org/10.1016/S0016-7037(97)00046-X)

Ruddiman, W., 1977. Late Quaternary deposition of ice-rafted sand in the subpolar North Atlantic (lat 40° to 65°N). Geological Society of America Bulletin 88, 1813–1827.

Ruddiman, W.F., 1977. North Atlantic ice-rafting: A major change at 75,000 years before present. Science, 196, 1208–1211.

Sansone, S.-A., McQuilton, P., Rocca-Serra, P., Gonzalez-Beltran, A., Izzo, M., Lister, A.L., Thurston, M., 2019. FAIRsharing as a community approach to standards, repositories and policies. Nature biotechnology 37, 358–367.

Seki, O., Kawamura, K., Sakamoto, T., Ikehara, M., Nakatsuka, T., Wakatsuchi, M., 2005. Decreased surface salinity in the sea of okhotsk during the last glacial period estimated from alkenones. Geophysical research letters 32. doi:[10.1029/2004GL022177](https://doi.org/10.1029/2004GL022177)

Simon, M.H., Arthur, K.L., Hall, I.R., Peeters, F.J., Loveday, B.R., Barker, S., Ziegler, M., Zahn, R., 2013. Millennial-scale agulhas current variability and its implications for salt-leakage through the indian–atlantic ocean gateway. Earth and Planetary Science Letters 383, 101–112. doi:[10.1016/j.epsl.2013.09.035](https://doi.org/10.1016/j.epsl.2013.09.035)

Smith, I.R., 2000. Diamictic sediments within high Arctic lake sediment cores: Evidence for lake ice rafting along the lateral glacial margin. Sedimentology 47, 1157–1179.

Steering Committee, P. Scientific, 2018. New data-availability procedures echo PAGES’ long-standing commitment. PAGES Magazine 56, 48.

Stern, J.V., Lisiecki, L.E., 2013. North atlantic circulation and reservoir age changes over the past 41,000 years. Geophysical Research Letters 40, 3693–3697. doi:[10.1002/grl.50679](https://doi.org/10.1002/grl.50679)

St John, K., Passchier, S., Tantillo, B., Darby, D., Kearns, L., 2015. Microfeatures of modern sea-ice-rafted sediment and implications for paleo-sea-ice reconstructions. Annals of Glaciology 56, 83–93.

Stoner, J.S., Channell, J.E., Hillaire-Marcel, C., Mareschal, J.-C., 1994. High-resolution rock magnetic study of a late pleistocene core from the labrador sea. Canadian Journal of Earth Sciences 31, 104–114. doi:[10.1139/e94-009](https://doi.org/10.1139/e94-009)

Wickham, H., 2019a. Stringr: Simple, consistent wrappers for common string operations.

Wickham, H., 2019b. Assertthat: Easy pre and post assertions.

Wickham, H., 2019c. Stringr: Simple, consistent wrappers for common string operations.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Silva Santos, L.B. da, Bourne, P.E., Bouwman, J., Brookes, A., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C., Finkers, R., Gonzalez-Beltran, A., Gray, A., Groth, P., Goble, C., Grethe, J., Heringa, J., Hoen, P. ’t, Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S., Martone, M., Mons, A., Packer, A., Persson, B., Rocca-Serra, P., Roos, M., Schaik, R. van, Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M., Thompson, M., Lei, J. van der, Mulligen, E. van, Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The fair Guiding Principles for scientific data management and stewardship. Scientific Data 3. doi:[10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18)

Xie, Y., Cheng, J., Tan, X., 2020. DT: A wrapper of the JavaScript library ’DataTables’.

1. Department of Geography, University of Wisconsin-Madison, USA [↑](#footnote-ref-20)
2. Department of Geoscience, University of Wisconsin-Madison, USA [↑](#footnote-ref-21)
3. Department of Geoscience, University of Wisconsin-Madison, USA [↑](#footnote-ref-22)
4. Department of Geography, University of Wisconsin-Madison, USA [↑](#footnote-ref-23)
5. Department of Geoscience, University of Wisconsin-Madison, USA [↑](#footnote-ref-24)
6. Department of Geoscience, University of Wisconsin-Madison, USA [↑](#footnote-ref-25)
7. Department of Geoscience, University of Wisconsin-Madison, USA [↑](#footnote-ref-26)
8. Department of Geoscience, University of Wisconsin-Madison, USA [↑](#footnote-ref-27)
9. Department of Geoscience, University of Wisconsin-Madison, USA [↑](#footnote-ref-28)
10. Department of Geoscience, University of Wisconsin-Madison, USA [↑](#footnote-ref-29)