Contributions of the group members:

Aishwarya and Mounica organized the project and worked together on this project.

Each of us took one part of the project as follows:

Aishwarya Rasal: figured out how to do the crawling; extracted first 3 fields (DOI, title and authors)

Mounica Sreedhara: figured out parsing; extracted next 3 fields (Author, corresponding_author, corresponding_author_email)

Aishwarya and Mounica: figured out how to do extracting first; and extracted rest of the fields (published_date, abstract, keywords, fullText)

Challenges that we faced:

Problem faced in reading the html page: http://www.g3journal.org

>html <- getURL(input, followlocation=TRUE)

Error in function (type, msg, asError = TRUE) :

Unknown SSL protocol error in connection to journals.plos.org:443

Fixed it:

library(xml2)

html <- read_html(input)

The website is obviously trustworthy. Unfortunately, both Rcurl and httr break at SSL verification. My web browser doesn't give any sort of warning. But, it was fixed by using the library(xml2) and using the R command read_html(input). Supposedly, the new package xml2 provides a wrapper around Rcurl /httr to make it easier to scrape all kinds of pages.