As a data analyst the best data repositories are the ones with the least features

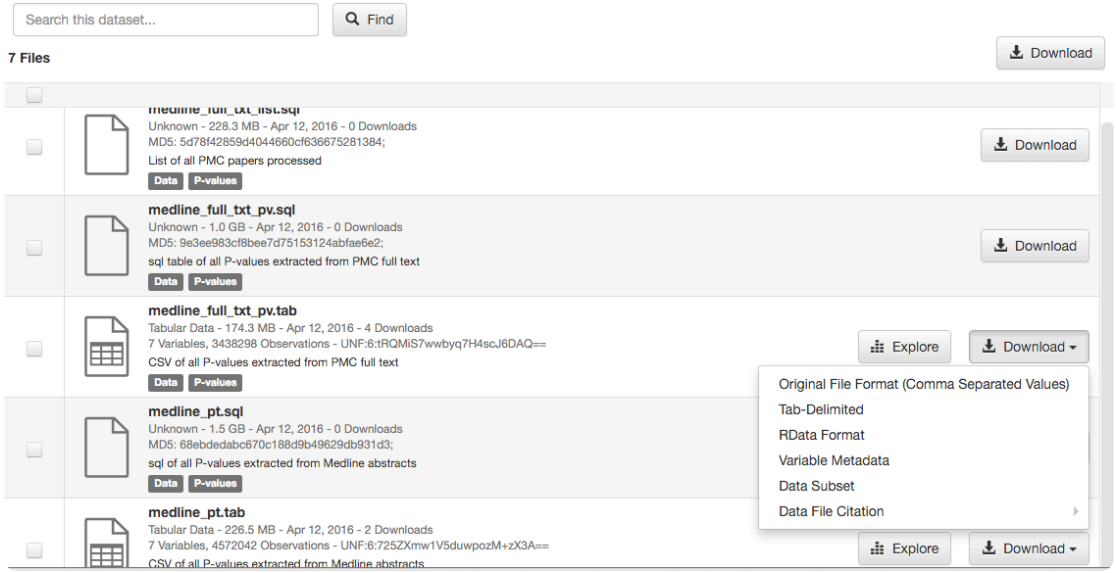
20 Apr 2016

By Jeff Leek

Lately, for a range of projects I have been working on I have needed to obtain data from previous publications. There is a growing list of data repositories where data is made available. General purpose data sharing sites include:

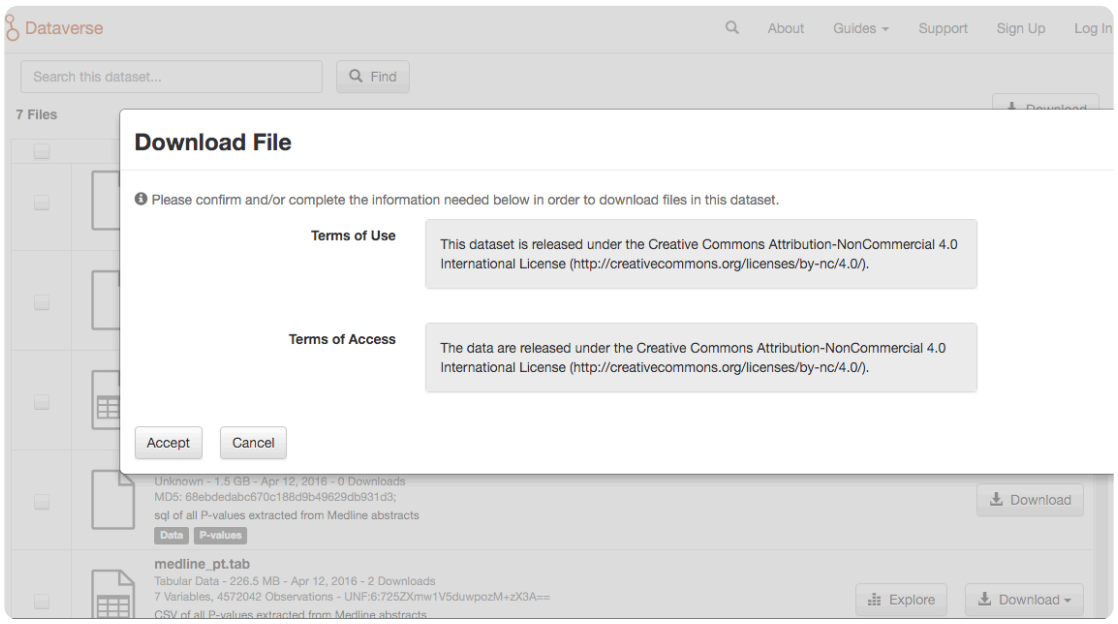
- The open science framework
- The Harvard Dataverse
- Figshare
- Datadryad

There are also a host of field-specific data sharing sites.One thing that I find a little frustrating about these sites is that they add a lot of bells and whistles. For example I wanted to download a p-value data set from Dataverse (just to pick on one, but most repositories have similar issues). I go to the page and click `Download` on the data set I want.



Then I have to accept terms:

Then I have to



Then the data set is downloaded. But it comes from a button that doesn't allow me to get the direct link. There is an R package that you can use to download dataverse data, but again not with direct links to the data sets.

This is a similar system to many data repositories where there is a multi-step process to downloading data rather than direct links.

But as a data analyst I often find that I want:

- To be able to find a data set with some minimal search terms
- Find the data set in .csv or tab delimited format via a direct link
- Have the data set be available both as raw and

processed versions
- The processed version will either be one or many tidy data sets.

As a data analyst I would rather have all of the data stored as direct links and ideally as csv files. Then you don't need to figure out a specialized package, an API, or anything else. You just use `read.csv` directly using the URL in R and you are off to the races. It also makes it easier to point to which data set you are using. So I find the best data repositories are the ones with the least features.

Related Posts

Not So Standard Deviations Episode 24 - 50 Minutes of Blathering 16 Oct 2016

Should I make a chatbot or a better FAQ? 14 Oct 2016

The Dangers of Weighting Up a Sample 12 Oct 2016