Auditing Open Data Publications with BITSS

Client: Temina Madon

About BITSS

The Berkeley Initiative for Transparency in the Social Sciences (BITSS) is a network of researchers and institutions committed to improving the standards of openness and integrity in economics, political science, psychology, and related disciplines. BITSS is headquartered at Berkeley, where its faculty staff are working to identify tools and strategies for increasing transparency and reproducibility in the social sciences. This includes the use of study registries, pre-analysis plans, version control, data sharing platforms, disclosure standards, and replications.

Project Description

This project aims to understand whether researchers actually publish and maintain their data and code when required by a peer-reviewed journal. To answer the question, we need to pull and analyze information from a randomly selected sample of articles published online in the American Economic Review (or another high profile field journal). Given a random sample, how many articles link to original data and code that is actually still available online? If posted, do the study's data sets provide raw material (i.e. data as they were measured), or just the final processed data used to produce tables in the print articles? Finally, can the available materials be used to replicate results? If it is possible to automate this last process, the project becomes scalable and could be applied to journals in other subjects. Co-authorship of a paper with several other researchers may be feasible as a result of this project; the other authors involved in this work would contribute qualitative & theoretical aspects of the analysis. A part of the project would involve writing a program that automatically runs a researcher's code on their posted data, and compares the outputs (ie results tables) with published versions. This could be tricky, since there is variation in the way statistical results are formatted, even within the same journal.

The problem to be solved includes: 1) writing a program to randomly select a sample of papers from an online repository; 2) scraping information from these articles (stored as html or PDFs), in particular the links to publicly posted data and code used to produce the article's results; 3) retrieving the online data and code, if available (or recording failure to download); 4) if feasible, writing a program that automatically runs the code on the data provided, to determine if the output (ie. "results" or data tables) matches whatever is published in the original article.