The Repository as Data (Re) User: Hand Curating for Replication

Limor Peer

Associate Director, Yale University Institution for Social and Policy Studies

This poster describes the tools and workflow used by the ISPS Data Archive to enhance the usability and usefulness of its research data. Specifically, it explains how replication of published results drives the curation work undertaken at the ISPS Data Archive and offers researchers a re-use test case.

The ISPS Data Archive is a small archive for a specialized community, consisting of replication materials for research in the social sciences which involves field experiments (that is, fully-randomized research designs in which observations found in a naturalistic setting -- voters, patients, welfare recipients, community organizations, government entities, and the like -- are assigned to treatment and control conditions). All files are available in open format and with minimal restrictions and they are linked directly to the scientific papers describing the research.

Curating digital research data properly involves many steps. These include adding labels, standardizing missing values, creating documentation, assessing and minimizing disclosure risk, assigning DOIs, and long-term preservation. Over time, standards and best practices have developed, helping guide the research data community as a whole (see, for example, ICPSR's portfolio of data enhancements: https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/lifecycle/ingest/enhance.html). There is consensus that each of these actions adds value to the data, making it more usable, and more useful.

More recently, there have been calls to further the usability and usefulness of data by including additional materials that will allow others to reproduce the research, in the service of advancing science. This most often involves the statistical code used to analyze the data, but may also include workflows and other computational tools necessary to reproduce the results. Repositories and services, such as Dataverse, have responded by making it increasingly possible to include such materials alongside the data. New initiatives, such as RunMyCode, offer a platform for disseminating the necessary pieces required to submit the research to scrutiny by fellow scientists and implement the methodology presented in a given scientific paper.

The ISPS Data Archive, however, strives to be a single access-point for users who wish to re-use data and code associated with published scientific papers. Moreover, it is also able to offer researchers a key service: verifying analysis and results before data publication. The Archive's strength is its position as an "on-the ground" repository for a small research community, which enable it to work with researchers in a sort of lab environment. By re-using data and code as part of the curation process, the ISPS Data Archive adds value not only to the data, but also to its researchers.