Incorporating W3C’s DQV and PROV in CISER’s Data Quality Review and Reproduction of Results Service

William C. Block      Florio O. Arguillas      Jeremy Williams
Cornell Institute for Social and Economic Research (CISER)

Abstract

A year and a half after its implementation, CISER’s Research Data Quality Review and Reproduction of Results Service (or R², for short), a service developed to encourage sharing of high quality data, code, documentation, and metadata associated with a study for the purpose of reproducible research, continues to evolve and improve. This poster discusses: a) the service at its current state; b) the improvements made to the service including cost-reduction and buy-in strategies to encourage researchers to use the service; c) pre- and post-reproduction services to improve data, code, documentation, and metadata quality; d) the inclusion of the Cornell’s CED2AR software for assessing and generating complete DDI metadata; and e) the utilization of the CISER Data Archive as the free and permanent home for the study and its associated files. Last, our poster will include for the first time our efforts to incorporate the W3C’s Data Quality Vocabulary (DQV) and PROV metadata into our R² service. DQV provides a means to describe the subjective measures applied to ensure the integrity of data sources and introduces a way of expressing guidelines for data quality that producers of data can abide by. The W3C PROV ontology defines entities, agents, and activities related to the origin of resources, which is particularly important in the common scenario of referencing multiple datasets for a given investigation.” By incorporating DQV and PROV into the R² data management workflow, we hope to provide a foundation of support improving the quality of research inputs and outputs, which in turn can have a chain effect for derivative data products.

Common problems: No variable and value labels

Golden Rule of CISER’s Replication Service:

Output produced by running code against the data should be identical to the publication up to the last decimal place. Slight deviation is not acceptable and must be investigated.

Common problems:
- Very long, complex codes
- Unnecessary/excess sections of codes whose outputs are not found in the paper (this delays replication because the Staff has to go through the entire code and its output, and figure out where they are on the paper)
- Code points to subdirectories for retrieving or saving data, thus Staff has to recreate the directory structure for the code to run correctly
- Some codes are not efficient, but will not be modified by the Staff. The Staff, however, will suggest ways to make it efficient.
- Codes are often multiple files with no indication of sequence. Reproducer has to determine which to run first especially if codes build on top of the other.

Common problems:
• Some results do not match article
• Not all figures printed on the article are produced by the code. Some involved other software packages such as Excel.
• Order of variables in the model in the printed table do not match the order of the variables in the output table for that model, which slows down the verification process.