

Authors

Stephanie Wykstra
Research Manager

IASSIST Quarterly

New Curation Software:

Step-by-Step Preparation of Social Science Data and Code
for Publication and Preservation

by Limor Peer¹ and Stephanie Wykstra²

Abstract

As data-sharing becomes more prevalent throughout the natural and social sciences, the research community is working to meet the demands of managing and publishing data in ways that facilitate sharing. Despite the availability of repositories and research data management plans, fundamental concerns remain about how to best manage and curate data for long-term usability. The value of shared data is very much linked to its usability, and a big question remains: What tools support the preparation and review of research materials for replication, reproducibility, repurposing, and reuse? This paper describes key curation tasks and new data curation software designed specifically for reviewing and enhancing research data. It is being developed by two research groups, the Institution for Social and Policy Studies at Yale University and Innovations for Poverty Action, in collaboration with CodeCite. The software includes curation steps designed to improve the research materials and thus to enable users to derive greater value from the data: (1) reviewing variable-level and study-level metadata, verifying the code can reproduce published results, and ensuring that it is removed. The tool is based upon the best practices of data archives and fits into existing and research workflows. It is open-source, modifiable, and will help ensure that shared data can be used.

Keywords

Data curation, curation software,
data sharing, social sciences,
reproducibility, research data

Introduction

Over the past 10 years, many scientific communities have embarked on discussions of data-sharing and reproducibility. From Biology (Olson, 2014) to Epidemiology (Peng, 2006) to Economics (Hansmann et al., 2011) to Political Science (Ding, 1998), researchers are calling for more data sharing. Research funders and journals have been encouraging data sharing and adopting data access policies in greater numbers over the past decade. For example, in the UK, all of the Research Councils have adopted data-sharing policies (see Data Curation Center's useful summary¹ of

all of these policies). Wellcome Trust in the UK has led a joint statement² of support on data-sharing principles, which includes over 75 funders. In the US, the Office of Science and Technology Policy memorandum of 2013³ stipulated that US funders receiving \$100M or more in federal research funds adopt data-sharing policies, and the government is working to facilitate such sharing. Major foundations such as the Bill and Melinda Gates Foundation⁴ and the Conrad and John Arnold Foundations⁵ have also adopted data-sharing policies. A number of journals are instituting policies in which they require researchers to share the data and code underlying the published research results (see this list of social science journals with a data sharing policy⁶ and the journal data policy review⁷).

There is much variety across policies: funders' policies differ in their timeliness, whether data should be made openly available or simply available on request, which materials should be shared, and in many other ways. (For an overview, see Wykstra, 2015.) Likewise, journals vary on whether data should be available openly. Some journals, for example the *American Economic Review*⁸, require researchers to post the data on the journal website, whereas other journals merely ask researchers to report in the article where they shared the data or that they make it available upon request.

research will be more credible if
others can have full access to all
aspects of scholarly work

While the language and particulars may vary, a common theme running through these discussions is the desire for scientists to be able to examine each other's work. Can others digest the analysis and data, can others understand the study in enough detail to try to repeat it?

In this paper, we focus on an issue which is crucial for examining others' work: that of the usability of shared data. By "data" here, we mean not just the datasets,

New Curation Software: Step-by-Step Preparation of Social Science Data and Code for Publication and Preservation

As data-sharing becomes more prevalent throughout the natural and social sciences, the research community is working to meet the demands of managing and publishing data in ways that facilitate sharing. Despite the availability of repositories and research data management plans, fundamental concerns remain about how to best manage and curate data for long-term usability. The value of shared data is very much linked to its usability, and a big question remains: What tools support the preparation and review of research materials for replication, reproducibility, repurposing, and reuse? This paper describes key curation tasks and new data curation software designed specifically for reviewing and enhancing research data. It is being developed by two research groups, the Institution for Social and

Policy Studies at Yale University and Innovations for Poverty Action, in collaboration with Colectica. The software includes curation steps designed to improve the research materials and thus to enable users to derive greater value from the data: Checking variable-level and study-level metadata, verifying that code can reproduce published results, and ensuring that PII is removed. The tool is based upon the best practices of data archives and fits into repository and research workflows. It is open-source, extensible, and will help ensure that shared data can be used.

April 18, 2016