Actionable Data Links: Tools for Reproducibility in Social Science and History

Albert Meroño-Peñuela[1], Ashkan Ashkpour[2], Richard Zijdeman[2], and Rinke Hoekstra[1,3]

[1] Department of Computer Science, Vrije Universiteit Amsterdam, NL
{albert.merono,rinke.hoekstra}@vu.nl
[2] International Institute of Social History, Amsterdam, NL
richard.zijdeman@iisg.nl
[3] Faculty of Law,University of Amsterdam, NL

Abstract. Reproducibility of studies in social science history is problematic. Published papers do not contain enough information for a complete replication of the study: a primary principle of the scientific method. Typically, the first problem is the absence of a *link to the actual data* on which the study is based. While most times this link is included in the form of dataset citation, a link pointing directly to the exact same data used originally is required for reliable replication. Imprecise data citation is especially problematic when data is volatile, or when the dataset comprises multiple sources, an increasingly common phenomenon in social science and history. Moreover, referring to "just the data" is not enough: it is necessary to point to the *sources* and *queries* over these sources that generated the data. Not including this information means that standard methods such as regressions, correlations, and visualizations are no longer reproducible.

In this paper, we propose `grlc`, a method that enables the curation, versioning, publishing, sharing and replication of *queries* over collections of research data. `grlc` makes the results that answer historical questions *actionable* via a single, unique web address (URL). We argue that sharing these queries, along with their provenance and the meta-data needed for their execution, enables their universal reuse, thus facilitating the reproducibility of studies. To illustrate our approach, we describe three use cases of `grlc`: integrating diverse access methods and sources of the Dutch historical censuses (1795-1971); generating data actionable links over social history data in the CLARIAH/Datalegend project; and the systematic retrieval of classification systems on the Web, concretely on reusing codes of the Linked International Classification of Religions (LICR).

Keywords: Replication, actionable historical data, Linked Data