**Software paper for submission to the Journal of Open Research Software**

To complete this template, please replace the blue text with your own. The paper has three main sections: (1) Overview; (2) Availability; (3) Reuse potential.

Please submit the completed paper to: editor.jors@ubiquitypress.com

**(1) Overview**

Title

The title of the software paper should focus on the software, e.g. “Text mining software from the X project”. If the software is closely linked to a specific research paper, then “Software from Paper Title” is appropriate. The title should be factual, relating to the functionality of the software and the area it relates to rather than making claims about the software, e.g. “Easy-to-use”.

Paper Authors

1. Antal, Dániel; *(Lead/corresponding author first)*

2. Last name, first name; *etc.*

Paper Author Roles and Affiliations

1. First author role and affiliation

2. Second author role and affiliation *etc.*

Abstract

A short (ca. 100 word) summary of the software being described: what problem the software addresses, how it was implemented and architected, where it is stored, and its reuse potential.

Survey data harmonization refers to procedures that improve the data comparability or the inferential capacity of multiple surveys **conducted in different periods of time, or in different geographical locations, using different languages.** The retroharmonize package support various data processing, documentation, file/type conversion aspects of various survey harmonization workflows.

Keywords

keyword 1; keyword 2; *etc.*

Keywords should make it easy to identify who and what the software will be useful for.

Introduction

An overview of the software, how it was produced, and the research for which it has been used, including references to relevant research articles. A short comparison with software which implements similar functionality should be included in this section.

Surveys, i.e., systematic primary observation and data collections are important data sources of both social and natural sciences. Survey data harmonization refers to procedures that improve the data comparability or the inferential capacity of multiple surveys conducted in different periods of time, or in different geographical locations, using different languages. The retroharmonize R package support various data processing, documentation, file/type conversion aspects of various survey harmonization workflows. The software was developed and tested with social sciences surveys using questionnaires, but various other surveying modes (inflation surveys with price scanning, […]) could be harmonized with our approach.

Various R packages support aspects of the workflow of survey harmonization.

Importing data from survey files (containing numeric codes and value labels), harmonizing concepts, then variables, labels, and coding, and eventually bringing them to the same variable types for binding or joining into a single, tidy dataset. The haven package in the tidyverse imports single SPSS and STATA files (which are almost always used for social sciences surveys.) The haven package is in turn builds on the labelled package for using variables that have both numeric and labelled representations. Whilst haven and label work perfectly with single surveys, they do not crosscheck the potential conflicts of conflicting labels, particularly conflicting special and missing value labels across files. We created inherited classes from these packages that create truly unique identifiers across several files and contain methods that prevent type conversion logical or syntax errors with inconsistent labelling.

“Data is only potential information without metadata”. This statement is painfully clear when you work with several surveys, which may contain the measurement of the same concept in differently named variables, held in different file types, using different numerical codes, value labels, and conflicting special characters. Retroharmonize extents the tidyverse packages for consistently mapping the imbedded metadata of various SPSS, STATA or even CSV files, and use this information for systematically change the names of variables, xxxx,xxxx.

There are several R packages that do a similar job, but have a less ambitious aim. Xxxxx., xxxxxx, xxxxx.

**Implementation and architecture**

How the software was implemented, with details of the architecture where relevant. Use of relevant diagrams is appropriate. Please also describe any variants and associated implementation differences.

Retroharmonize was developed over several years with implementing more and more harmonization tasks in various ex ante harmonized surveys, starting with the European Eurobarometer series, then adding Afrobarometer, Lationbarometro and private surveys. These international survey research programs provide access to their harmonized surveys in “waves”. Usually, they call a way a set of ex ante harmonized surveys (containing the same questionnaire in several languages) in one data collection period. Our added value has been that we further harmonize data among waves (when data is not fully ex ante harmonized and requires ex post or retrospective harmonization.)

**Quality control**

Detail the level of testing that has been carried out on the code (e.g. unit, functional, load etc.), and in which environments. If not already included in the software documentation, provide details of how a user could quickly understand if the software is working (e.g. providing examples of running the software with sample input and output data).

Retroharmonize was extensively tested on privately conducted surveys, and three large, international, ex ante harmonized survey programs (questionnaire-based social science research aimed for ex post or retrospective harmonization across countries and years): Eurobarometer, Afrobarometer and Latinobarometro.

For unit testing, we included in the R package three subsets of published Eurobarometer surveys. The package’s unit testing consists of about 130-unit tests made with this real-life survey excepts.

**(2) Availability**

***Operating system***

Please include minimum version compatibility.

***Programming language***

Please include minimum version compatibility.

***Additional system requirements***

E.g. memory, disk space, processor, input devices, output devices.

***Dependencies***

E.g. libraries, frameworks, incl. minimum version compatibility.

The retroharmonize R package is practically a very thorough extension of the R tidyverse packages: it depends on haven (and labelled) for working with coded survey files. It uses dplyr, tidyr (and their common, deep level rlang, vctrs) dependencies for variable manipulation within a single survey (preparation for harmonization) and purrr for functional programming task with several surveys.

***List of contributors***

Please list anyone who helped to create the software (who may also not be an author of this paper), including their roles and affiliations.

Marta Kolcynska () as a survey harmonization expert contributed to the conceptual development of the first documentation, building the first use cases and exploring various survey harmonization workflows that may need a reproducible and computational support.

***Software location:***

***Archive*** (e.g. institutional repository, general repository) (required – please see instructions on journal website for depositing archive copy of software in a suitable repository)

***Name:*** The name of the archive

***Persistent identifier:*** e.g. DOI, handle, PURL, etc.

***Licence:*** Open license under which the software is licensed

***Publisher:*** Name of the person who deposited the software

***Version published:***The version number of the software archived

***Date published:*** dd/mm/yy

**Code repository** (e.g. SourceForge, GitHub etc.) (required)

***Name:*** The name of the code repository

***Identifier:*** The identifier (or URI) used by the repository

***Licence:*** Open license under which the software is licensed

***Date published:*** dd/mm/yy

**Emulation environment** (if appropriate)

***Name:*** The name of the emulation environment

***Identifier:*** The identifier (or URI) used by the emulator

***Licence:*** Open license under which the software is licensed here

***Date published:*** dd/mm/yy

***Language***

Language of repository, software and supporting files

**(3) Reuse potential**

Please describe in as much detail as possible the ways in which the software could be reused by other researchers both within and outside of your field. This should include the use cases for the software, and also details of how the software might be modified or extended (including how contributors should contact you) if appropriate. Also you must include details of what support mechanisms are in place for this software (even if there is no support).

The retroharmonize R package aims to provide a versatile support for various survey harmonization workflows. Because surveys are so fundamental to quantitative social science research and play an important role in many natural science fields, not to mention commercial applications of market research or pharmaceutical research, the package’s main reuse potential is to be a foundation of further reproducible research software aimed to automate research and harmonization aspects of specific survey programs.

The authors of this package started the development work to be able to harmonize surveys from harmonized data collections of the European Union: namely the Eurobarometer and AES surveys programs. After working with various surveys (also outside these programs) it became clear that retroharmoinze should aim to be a common demoninator to a family of similar software that solves more specific problems. The world’s

largest and oldest international public policy survey series, Eurobarometer. This program alone has conducted already thousands of surveys in more than 20 natural languages over more than 40 years, following various documentation, data management, coding practices that were not independent of the software tools available over this long period of time. The first verion of retroharmonize was separated to the retroharmonize and the eurobarometer R packages – retroharmonize providing a more general framework that has been able to serve Eurobarometer’s, Afrobarometer’s and the Arab Barometer’s different needs.

In our view, the retroharmonize package has the potential to become a general supporting software for more specific codes aimed at harmonizing surveys based first on questionnaires, later on different data inputs, such as price scanning, laboratory tests, and other standardized, discrete inputs that are carried out in different locations, with different recording tools, and with different coding (for example, because of natural languages differences, as it is the case in the social science surveys used for the testing of our software.)

**Acknowledgements**

Please add any relevant acknowledgements to anyone else who supported the project in which the software was created, but did not work directly on the software itself.

**Funding statement**

There was no funding available for the development of this software.

**Competing interests**

If any of the authors have any competing interests then these must be declared. The authors’ initials should be used to denote differing competing interests. For example: “BH has minority shares in [company name], which part funded the research grant for this project. All other authors have no competing interests."

If there are no competing interests, please add the statement:

“The authors declare that they have no competing interests.”

**References**

Please enter references in the Harvard style and include a DOI where available, citing them in the text with a number in square brackets, e.g.

[1] Piwowar, H A 2011 Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data. *PLoS ONE* 6(7): e18657. DOI: http://dx.doi.org/10.1371/journal.pone.0018657.

**Copyright Notice**

Authors who publish with this journal agree to the following terms:

Authors retain copyright and grant the journal right of first publication with the work simultaneously licensed under a [Creative Commons Attribution License](http://creativecommons.org/licenses/by/3.0/) that allows others to share the work with an acknowledgement of the work's authorship and initial publication in this journal.

Authors are able to enter into separate, additional contractual arrangements for the non-exclusive distribution of the journal's published version of the work (e.g., post it to an institutional repository or publish it in a book), with an acknowledgement of its initial publication in this journal.

By submitting this paper you agree to the terms of this Copyright Notice, which will apply to this submission if and when it is published by this journal.