Technical Documentation: Reproducible Harmonization of Arab Barometer Files

Mohamed Abufalgha, Zuzana Gombiková, Daniel Antal,  
Michael Robbins, MaryClare Roche

2020-09-20

# Overview

To accommodate different workflows, this document is available in numerous formats. It is available (and probably best viewed) as a password protected project intranet website on [arabbarometer.dataobservatory.eu](http://arabbarometer.dataobservatory.eu/).

While [retroharmonize](https://retroharmonize.dataobservatory.eu/) is open-source, and open for all, and we expect that we will create materials that are certainly not confidential, we tried to set up a private repo and workflow that is discrete, lean, and leaves the choice to make results visible for the Arab Barometer team.

In the Annex, you can see who has access to our private repo, and of course, how can you add any of your colleagues to have oversight of the entire process.

**Project Planning & Management**

[Project Planning & Management](#bookmark=id.30j0zll) consists of three parts.

Part [General questions & clarifications](#bookmark=id.1fob9te) currently contains questions from Daniel to the Arab Barometer team about considerations, or survey file specific questions that should be taken into account for our joint work.

Please review the 1.2 [Specific questions](#bookmark=id.2et92p0). They are simple issues that can be cleared off to deliver the first outcomes.

When the second part, [Specific questions](#bookmark=id.2et92p0), is answered and cleared, these will be moved to the [Technical Report](#bookmark=id.3whwml4).

The tird part, [Desired outputs](#bookmark=id.4d34og8), is a first proposal from Daniel that should be reviewed by the Arab Barometer team.

Desired Outcomes

[Tutorial](#bookmark=id.1ksv4uv) is the first draft of a tutorial article (‘vignette’ in R parlance) to our [retroharmonize](https://retroharmonize.dataobservatory.eu/) package, to make this work more accessible is to create and later for other researchers to create their own harmonized data. This was the original request from Daniel and Marta to be included in the package.

It is a simple tutorial, and a lot of questions came up while making it. Most probably a more meaningful or interesting tutorial should come out once we review those questions.

The Tutorial currently looks like [this](http://arabbarometer.dataobservatory.eu/static/arabbarometer.html). Once you are comfortable with it, I can place it into the [retroharmonize open repo](https://github.com/antaldaniel/retroharmonize/), and we can finalize and test it there. Anyone, can fork the repo, make pull request, or join in as a contributor to the package that is currently under review on CRAN.

[Creating Harmonized Datasets](#bookmark=id.4i7ojhp) is an outcome in general agreed with Michael.

[Comparison with Afrobarometer](#bookmark=id.2xcytpi) and [Comparison with Eurobarometer](#bookmark=id.1ci93xb) are Daniel’s suggestions, and they can be reviewed or dropped from the [Desired outputs](#bookmark=id.4d34og8), should Arab Barometer’s team disagree with them.

[Technical Report](#bookmark=id.3whwml4) will be developed version by version in this document by moving cleared [General questions & clarifications](#bookmark=id.1fob9te) and [Specific questions](#bookmark=id.2et92p0) to a clean, final documentation.

We believe that the tutorial, the larger dataset, the two optional, experimental use cases with Afrobarometer and Eurobarometer cross-program harmonization will lead to enough conclusions in data management, workflow management, and hopefully the datasets that have their own scientific value. The documentation of this process may serve as a Technical Guideline for further development of the Arab Barometer project, or, if data errors are found, to make [revisions](#bookmark=id.1hmsyys).

The **Annex** contains miscellaneous information that is not intended for any future publication.

**Practicalities** is a short overview of how this document is being created, and explains how you can edit, comment, question this document via a private github <https://github.com/dataobservatory-eu/arabbarometer> repo, only visible for invited team members, or in a shared [Google Drive](https://drive.google.com/drive/folders/11ZBKae9buRqijbeMwrrwIyWmc5s1BCp_?usp=sharing), again, only available for the same teams. *If you are experienced with github and RStudio, the description will be trivial, but you will find the practical repo links and authorizations given to team members here.*

At this stage, **Disclosures** are intended only for internal use. As we are finalizing outputs that may become public, we can review or amend them whenever a disclosure must be made to public.

# Project Management

## General questions & clarifications

This would help us understand how we can help, and how to record the problems, ideas that we are coming across while using your data and files. They are rather different from the files we see in Eurobarometer and Afrobarometer, they certainly have very specific strengths and weaknesses.

### Arab Barometer file & document revision policy

* Eurobarometer’s files are stored by GESIS, and they review small file errors based on user feedback for a couple years after first version publication (mislabelling, accidentally left out interview dates or weighting variables, etc.)

Is there such a policy with Arab Barometer? If yes, is it an aim of this small project to document small errors in the files for a new version SPSS file release? How about clarifications for the documentation of the already released files?

**Arab Barometer team** *please comment or discuss in a call*

## Specific questions

Once answered and cleared, these will be moved to the [Technical Report](#bookmark=id.3whwml4).

**Daniel**: > My general impression is that it is far more difficult to find variables than in Afrobarometer or Eurobarometer, because neither the question labels or the question numbers appear to be fully in harmony. However, the actual (SPSS) variable names (like q101) appear to be far more consistent. If there is any recorded practice here, then we could really speed up the process. With Eurobarometer and Afrobarometer we really relied on a lot of string processing with the labels. The labels seem a bit more messy in Afrobarometer, but the actual file organization is better.

I also have a question mark if we should aim at first go to include Wave I. It appears to have quite a few differences from waves II-V.

I am wondering if you have written guidelines on using / labelling / coding various forms of missing data (don’t know, refused, question filter out.) I think that the three survey programs have very different issues. GESIS produces the most consistent missing data labelling, but when it fails, failure is costly. Afrobarometer has consistent labelling, but inconsistent coding in SPSS. (for example na\_range and na\_labels contradict in the SPSS coding.) Arab Barometer appears to use a lot more labels than GESIS or Afrobarometer. I cannot yet form an opinion on file coding level.

### Date

Almost all Eurobarometer files and all but the first Afrobarometer files contain the data of the interview, but I found these only in Arab Barometer 3 and Arab Barometer 5. Did I overlook something?

Having at least a notional date for the interview is important for many reasons. One way to overcome this problem is to use the middle date of the fieldwork period. However, currently I only find this information in the PDF files, and they vary by country. Is this information available in a different format?

### Weights & parget population

What is the exact definition of the weight variables? Are these post-stratification weights per sample, country, region? Am I correct that this variable is missing from *Arab Barometer I*?

The *Citizens aged 18 and above* is a bit puzzling target population group, but of course, easy to work with, given the large presence of non-citizen, semi-citizen and other status people in the Arab World. Particularly in Jordan this would require a footnote at least. For comparison, Eurobarometer is targeting European population aged 15 and above, regardless of citizenship, though I am very uncertain how some non-citizen classes, particularly people with not yet processed refugee application may get into the sample.

All in all, if I could have the notional population size that was used to create the weight variable, I would create a function and add to the data files a projected weight, which makes counting things much easier. GESIS produces a lot of weighting options that are well documented. I’d like to clarify how you create the weights. I would find it particularly useful if we could create an alternative projected weight programmatically (wex in Eurobarometer) showing how many respondents are notionally one respondent by a single respondent (i.e. the particular interviewed person, after post-stratification, represents 8721 people), which must total the entire target population (as far as I see, *Citizens aged 18 and above*).

Generally speaking, clarifying what the population Arab Barometer represents would give the opportunity to join other data, for example, population data for projected weights, or any further data. Our initial work with Eurobarometer started the other way around: We wanted to join data derived from Eurobarometer to various Eurostat data products. (For example, how many library uses per 10,000 people are in European regions, or library users per library per country, etc.) A correct, standardized geographic coding and a clear view on how the sample is designed could lead to added features you can find in Eurobarometer and a lot of new use cases.

### Naming and labelling consistency

Further, we select all variables that evaluate the economic situation in a country. This question was asked in a slightly different format in the *Arab Barometer Wave I* than in later versions. In this case, the question numbering is consistent, and could be selected by lowercase(var\_name\_orig) == "q101" The *Arab Barometer Wave 5* file uses uppercase names, the earlier ones lowercase names.

* Are Arab Barometer question numbers supposed to be harmonized?(Eurobarometer only uses the same variable names for some socio-demographic, “protocol” and other repeating variables.)

## Desired outputs

This document is created with the [bookdown](https://bookdown.org/) extension and package of RStudio. Bookdown allows users to *knit* together several Rmd files and publish them in an easy-to-read, intranet site or website format. Each chapter of this long-format documentation is a separate Rmd markdown document, and as we progress with the project, we will write up the later chapters.

### Tutorial for retorharmonize [vignette]

The first step to make this work more accessible is to create a tutorial article (‘vignette’ in R parlance) to our [retroharmonize](https://retroharmonize.dataobservatory.eu/) package, which is a fully open source, free software for retrospective survey harmonization written in the R language and released with a GPL-3 license. This license allows the use and the modification of our source code while it remains open source.

The vignette should enable any researcher with an intermediate level command of the R language to create harmonized data from Arab Barometer’s files. This little article after review by Arab Barometer will be released as part of the package. The article will be part of a peer-review open source program code documentation released on [CRAN - The Comprehensive R Archive Network](https://cran.r-project.org/) after automated and manual peer-review.

Apart from being approved by Arab Barometer, this article must follow the documentation standards of CRAN (i.e. take a similar format to our [Eurobarometer](https://retroharmonize.dataobservatory.eu/articles/eurobarometer.html) and [Afrobarometer](https://retroharmonize.dataobservatory.eu/articles/afrobarometer.html) articles).

The first draft of the vignette is the [Tutorial](#bookmark=id.1ksv4uv) of the book, which can be separately reviewed as 03-tutorial.Rmd. For clarity, the current chapter is stored in 02-general.Rmd.

It contains both the text and the code of the tutorial, but not in a *live* format. This means that it is not executed here, and the code cannot be tested in this technical document. Once the language of the first version is reviewed by the Arab Barometer team, and deemed ready for further testing, it will be moved to the open repo of [retroharmonize](https://github.com/antaldaniel/retroharmonize/) on a development branch. This means that general users of [retroharmonize](https://github.com/antaldaniel/retroharmonize/) will not download and use it, but, because of the nature of the GPL-3 license, anybody will be able to make suggestions to the text in the open source community.

Any members of the Arab Barometer team who contribute to the document as a text will be credited in the document. Should any member of the Arab Barometer team significantly contribute to the testing and improving of the actual code, they will get attribution as contributors to the software source (See [Current code attributions](https://retroharmonize.dataobservatory.eu/authors.html), provided that they adhere to the [Contributor Covenant](https://www.contributor-covenant.org/version/2/0/code_of_conduct/), which is an ethical standard used in the open source community to maintain an open, welcoming, diverse, inclusive, and healthy co-developer community. Should this contribution be as significant that it adds significantly new features to the entire [retroharmonize](https://github.com/antaldaniel/retroharmonize/) package, they will be attributed as co-authors of the package code.

The open source statistical community has its own peer-review system that is related, but partly independent from the more general peer-review system of academia. Open source code published on [CRAN - The Comprehensive R Archive Network](https://cran.r-project.org/) after automated and manual peer-review is a special type of peer-reviewed publication. The practice of [CRAN - The Comprehensive R Archive Network](https://cran.r-project.org/) and the R community is that once a peer-reviewed package creates scientifically valid results as statistical software, and its creators, with potentially other authors describe the methodology of the statistical software and its scientific merits from a methodological point of view, and it is published in a peer-reviewed journal, than the peer-reviewed code and the separate peer reviewed article are both cited, and all users of the package must cite both. The creators of [retroharmonize](https://github.com/antaldaniel/retroharmonize/) do aim for such a publication, but only after there have been successful scientific uses of the package that were individually used in other peer-reviewed or audited work — i.e. after harmonized files created from Arab Barometer, Afrobarometer, Eurobarometer or other survey programs were successfully used in scientific uses independently of the package peer-review (i.e. in scientific journals, papers that were not reviewed by CRAN or publications like the Journal of Statistical Software.)

In short:

\* The Tutorial, once approved, will be part of the [retroharmonize](https://github.com/antaldaniel/retroharmonize/) package and will be sent for open source code publication review to [CRAN](https://cran.r-project.org/). This should give a relatively high level of confirmation for the usability of and datasets we create in this collaboration.

\* The aim of this Tutorial mini-article is not to go for peer-review as an article. It is aimed to help researchers to create valid datasets that they can use in their own peer-reviewed, scientific contexts, where they must give a citation to [retroharmonize](https://github.com/antaldaniel/retroharmonize/) as a software code only.

\* If there will be a few successful, peer-reviewed publications made with [retroharmonize](https://github.com/antaldaniel/retroharmonize/), the co-authors and contributors of the package will try to create a separate publication in a journal focusing on statistical software methodology or reproducible research practices, such as the R Journal, the Journal of Statistical Software or PLOS One. But this is clearly *not the aim* of this project, and it may take more than a year to reach this stage with [retroharmonize](https://github.com/antaldaniel/retroharmonize/) even if the researchers using Eurobarometer, Afrobarometer and, of course, Arab Barometer start to use it.

### Creating h=Harmonized Datasets

We will make an initial assessment with example data files, descriptions, about what we think to be *easily harmonized* with our package from Arab Barometer files.

*Easily harmonized* means that we aim to create harmonized datasets that are scientifically valid, and *do not* require a very significant amount of work.

We believe that these **Harmonized Arab Barometer files** will have their own scientific, publication and data release value, because our experience dictates that a large part of the variables are easily harmonized with our software. We will refine these datasets until Arab Barometer’s team believes that it is scientifically valid and can be released or published. We leave it to Arab Barometer’s discretion how to publish these data sets as long as we get proper attribution for our work.

We will continue to develop ideas and the documentation of dataset candidates in the subsequent versions of [Creating Harmonized Datasets](#bookmark=id.4i7ojhp) of this document.

### Comparison with Afrobarometer

We would also like to attempt to create a small, harmonized dataset that potentially connects data from Arab Barometer and Afrobarometer.

In this case, we will refine these datasets if Arab Barometer’s team believes that the dataset can be brought into a form that has scientific value, which we cannot foresee before trying to make the harmonized dataset.

Marta & Daniel got in touch with Afrobarometer, who were very enthusiastic about our work, and promised to look into ways of collaboration. We do not know what this means, but it would be desirable to review the findings in a joint call with them.

Questions from a technical point of view:

* what are the technical challenges to join data technically between Afrobarometer and Arab Barometer files;
* what can our software solve, and what it cannot; what lessons can be learned;

Questions from a content point of view:

* if we can make a (relatively) successful technical harmonization, on what scientific grounds / with what expectations can such harmonized files be exploited;
* what are the advantages of the geographical overlap.

We would like to present this small dataset with explanatory notes on our experience about the harmonization and ideas how this harmonization world could be further developed to both Arab Barometer, and if they want to be engaged, to Afrobarometer, with Arab Barometer’s leading role.

### Comparison with Eurobarometer

We would make a similar attempt to make a technical and a very high-level content comparison with Arab Barometer and Eurobarometer in a form of limited dataset, similar to the above

Daniel is in discussion with Kantar’s R user group about a possible engagement with [retroharmonize](https://retroharmonize.dataobservatory.eu/) and its connecting platform, co-created with Marta, [Eurobarometer](https://eurobarometer.dataobservatory.eu/). One of Kantar’s team members wanted to join our work, and we advised him to seek explicit approval from the management. This may or may not bring a possible external review opportunity into these files.

### Technical Report

We believe that the Tutorial, the larger **Harmonized Arab Barometer files**, two more experimental use cases to created an Arab Barometer / Afrobarometer and Arab Barometer / Eurobarometer joined, harmonized dataset will lead to enough conclusions, and hopefully datasets that have their own scientific value.

The report will be developed version by version in this document in [Technical Report](#bookmark=id.3whwml4).

# Project Outcomes

## Tutorial for retroharmonize [vignette]

We believe that the tutorial, the larger **Harmonized Arab Barometer files**, two more experimental use cases to created an Arab Barometer / Afrobarometer and Arab Barometer / Eurobarometer joined, harmonized dataset will lead to enough conclusions, and hopefully datasets that have their own scientific value. It would be very impractical to edit the actual code in this document, so we reprint it without being executed as a program code to be reviewed as a documentation only.

The development version from [GitHub](https://github.com/) can be installed with:

**require**("devtools")  
devtools**::install\_github**("antaldaniel/retroharmonize")

Below is the entire vignette, if executed, it looks like [this](http://arabbarometer.dataobservatory.eu/static/arabbarometer.html).

*If you want to contribute to the retroharmonize package or the vignette code, please for the repository.*

The goal of this case study is to explore the variation in trust in various state institutions among African societies, as well as changes in trust over time.

To do this, we use data from [Arab Barometer](about:blank), a nonpartisan research network that provides insight into the social, political, and economic attitudes and values of ordinary citizens across the Arab world.

retroharmonize is not affiliated with Arab Barometer. To fully reproduce this example, you must acquire the data files from them, which is free of charge. If you download and use the survey data, the Arab Barometer does not take any responsibility for the results of the analysis of the Arab Barometer data. All results published using the Arab Barometer data and related content are the responsibility of the respective author.

Some elements of the vignette are not “live”, because we want to avoid re-publishing the original microdata files from Arab Barometer. However you can access the data directly from the [arabbarometer.org website](https://www.arabbarometer.org/survey-data/data-downloads/).

The files are stored on the website in .zip format, which may be extracted to folders. We assume that you extracted and copied all .sav files into a single folder that we will call in this vignette the arabbarometer\_dir. Define your arabbarometer\_dir with file.path() in your own system.

|  |  |
| --- | --- |
| File name | Wave |
| ABI\_English.sav | Arab Barometer Wave I |
| ABII\_English.sav | Arab Barometer Wave II |
| ABIII\_English.sav | Arab Barometer Wave III |
| ABIV\_English.sav | Arab Barometer Wave IV |
| AB-WaveV-EN.sav | Arab Barometer Wave V |

## Importing Arab Barometer Files

We start by reading in the three rounds of the Arab Barometer.

**library**(retroharmonize)  
**library**(dplyr)  
**library**(tidyr)

*### use here your own directory*  
ab <- **dir** ( arabbarometer\_dir )  
arabbarometer\_rounds <- **file.path**(arabbarometer\_dir, ab)  
  
arab\_waves <- **read\_surveys**(arabbarometer\_rounds, .f='read\_spss')

Let’s give a bit more meaningful identifiers than the file names, which were imported in alphabetical order.

**attr**(arab\_waves[[1]], "id") <- "Arab-Barometer\_5"  
**attr**(arab\_waves[[2]], "id") <- "Arab-Barometer\_1"  
**attr**(arab\_waves[[3]], "id") <- "Arab-Barometer\_2"  
**attr**(arab\_waves[[4]], "id") <- "Arab-Barometer\_3"  
**attr**(arab\_waves[[5]], "id") <- "Arab-Barometer\_4"

We can review, if the main descriptive metadata is correctly present with document\_waves().

documented\_arab\_waves <- **document\_waves**(arab\_waves)

**print**(documented\_arab\_waves)  
**save**(documented\_arab\_waves, file = "Arabb.Rda")

Create a metadata table, or a data map that contains information about the variable names and labels, and where each row corresponds to one variable in the survey data file. We use the function metadata\_create(), which extracts the metadata from the survey data files, normalizes variable labels, identifies ranges of substantive responses and missing value codes.

arabb\_metadata <- **lapply** ( X = arab\_waves, FUN = metadata\_create )  
arabb\_metadata <- **do.call**(rbind, arabb\_metadata)

At this point, you should have a metadata table with 1587 observations and 12 variables.

### Working with metadata

From the metadata file we select only those rows that correspond to the variables that we’re interested in: the rowid being the unique case identifier, DATEINTR with the interview date, country containing information about the country where the interview was conducted, REGION with the region (sub-national unit) where the interview was conducted, and weight being the weighting factor (which is not available in the *Arab Barometer Wave I* file).

Further, we select all variables that evaluate the economic situation in a country. This question was asked in a slightly different format in the *Arab Barometer Wave I* than in later versions. In this case, the question numbering is consistent, and could be selected by lowercase(var\_name\_orig) == "q101" The *Arab Barometer Wave 5* file uses uppercase names, the earlier ones lowercase names.

For these variables, we create normalized variable names (var\_name) and labels (var\_label).

**library**(dplyr)  
to\_harmonize <- arabb\_metadata **%>%**  
 **filter** ( var\_name\_orig **%in%**   
 **c**("rowid", "country", "date", "wt") **|**   
 **grepl**("current economic|overall economic", label\_orig )  
 ) **%>%**  
 **mutate** ( var\_label = **var\_label\_normalize**(label\_orig)) **%>%**  
 **mutate** ( var\_label = **case\_when** (   
 var\_name\_orig **==** "country" **~** "Country",   
 var\_name\_orig **==** "rowid" **~** "Unique ID AB English",  
 var\_name\_orig **==** "date" **~** "Date of interview",  
 var\_name\_orig **==** "wt" **~** "Weight",  
 TRUE **~** "Evaluation economic situation")) **%>%**  
 **mutate** ( var\_name = **val\_label\_normalize**(var\_label))

Compared to our examples with [Afrobarometer](https://retroharmonize.dataobservatory.eu/articles/afrobarometer.html) and [Eurobarometer](https://retroharmonize.dataobservatory.eu/articles/eurobarometer.html), we could far better rely on the original SPSS variable names, and less on labelling. This is a safe way of harmonization, but requires a different type of reproducible workflow.

Daniel’s comment: if the original variable names are largely harmonized, the best would be to create an explicit variable name and label metadata table, and use it, maybe even release an updated version of the AB SPSS files. Compared to the other survey programs, at least at first sight, the labelling seems to be far less coherent, and the naming far more coherent. In the other cases, we overcome this by trying to match variables by labels.

The resulting table with information about the variables selected for harmonization looks as follows:

**head**(to\_harmonize **%>%**  
 **select** ( **all\_of**(**c**("id", "var\_name", "var\_label"))), 10)

The merge\_waves() function harmonizes the variable names, the variable labels and survey identifiers and returns a list of surveys (of class survey().) The parameter var\_harmonization must be a list or a data frame that contains at least the original file name (filename), original variable names (var\_name\_orig), the new variable names (var\_name) and their labels (var\_label), so that the program knows which variables to take from what files and how to call and label them after transformation.

merged\_ab <- **merge\_waves** ( waves = arab\_waves,   
 var\_harmonization = to\_harmonize )  
  
*# country will be a character variable, and doesn't need a label*  
merged\_ab <- **lapply** ( merged\_ab,   
 FUN = **function**(x) x **%>%**  
 **mutate**( country = **as\_character**(country)))

Review the most important metadata with document\_waves():

documenteded\_merged\_ab <- **document\_waves**(merged\_ab)

**print**(documenteded\_merged\_ab)

The files we want to harmonize must contain the same variables. If we want to weight the answers, we should either drop Arab-Barometer\_1, or for simplicity, add weight = 1 to this survey.

*## not evaluated, this is how to drop the Arab-Barometer\_1*  
merged\_ab[[2]] <- NULL

merged\_ab[[2]]**$**weight <- 1

Similarly, we can either drop the date of the interview or fill it up with, for example, the middle day of the surveying period.

merged\_arabb <- **lapply** ( merged\_ab, **function**(x){  
 **if** ( "date\_of\_interview" **%in%** **names**(x) ) {  
 **subset** ( x, select = **-c**(date\_of\_interview))  
 } **else** {  
 *## subset works with survey class, if omitted, will return*   
 *## data frame without metadata*  
 **subset** ( x )  
 }  
})

documenteded\_merged\_arabb <- **document\_waves**(merged\_arabb)

**print**(documenteded\_merged\_arabb)

Now we have the same four variables in each wave.

## Harmonization

In the Arab Barometer, the trust is measured with four-point ordinal rating scales. Such data are best analyzed with ordinal models, which do not assume that the points are equidistant. However, to get a quick idea of how the data look like, we will assign numbers 0-3, with 0 corresponding to the least trust, and 3 corresponding to the most trust, and for the time-being analyze the data as if they were metric.

To review the harmonization on a single survey use pull\_survey(). Here we select Arab Barometer Round 6.

R3 <- **pull\_survey** ( merged\_ab, id = "Arab-Barometer\_3")

**attributes**(R3**$**evaluation\_economic\_situation[1**:**20])

The document\_survey\_item() function shows the metadata of a single variable.

**document\_survey\_item**(R3**$**evaluation\_economic\_situation)

Arab Barometer’s SPSS files do not mark the missing values, so we have to be careful.

The set of valid category labels and missing value labels are as follows:

**collect\_val\_labels** (to\_harmonize **%>%**  
 **filter** ( **grepl**( "evaluation\_economic\_situation", var\_name) ))

Apparently, there is a number of missing value labels here, and it is not even always clear if they refer to the same type of “missingness” with the same word.

In the examples with [Afrobarometer](https://retroharmonize.dataobservatory.eu/articles/afrobarometer.html) and [Eurobarometer](https://retroharmonize.dataobservatory.eu/articles/eurobarometer.html), we saw different problems with missing variables.

* In Eurobarometer, the type of missingness is relatively clear, and in the SPSS files, the missing labels are usually, but not always identified.
* In Afrobarometer, the missing classes are clear, but the SPSS missing labels are not used, on the other hand, the SPSS missing range (na\_range) is used, but sometimes inconsistent with the missing labels (na\_labels).

In some statistical analysis, it does not matter, why a value is missing (for example, when creating averages from numeric values or numeric representations), and in others it doesmatter sometimes (working with categories.)

We create a harmonization function from the harmonize\_values() prototype function. In fact, this is just a re-setting the default values of the original function. It makes future reference in pipelines easier, or it can be used for a question block only, in this case to variables with starts\_with("trust").

harmonize\_arabb\_trust <- **function**(x) {  
 label\_list <- **list**(  
 from = **c**("very\\sgood", "(?<!very\\s)good",  
 "(?<!very\\s)bad", "very\\sbad",   
 "t\\sknow", "refuse", "decline", "missing"),   
 to = **c**("very\_good", "good", "bad",   
 "very\_bad", "do\_not\_know",   
 "declined", "declined", "missing"),   
 numeric\_values = **c**(3,2,1,0, 99997, 99998, 99998,99999)  
 )  
   
 **harmonize\_values**(  
 x,   
 harmonize\_labels = label\_list,   
 na\_values = **c**("do\_not\_know"=99997,  
 "declined"=99998,  
 "missing"=99999),   
 perl = TRUE  
 )  
}

Let’s see if the output makes sense:

*# We only print a random sample from the results.*  
**set.seed**(2020)  
**harmonize\_arabb\_trust**(  
 **sample**(R3**$**evaluation\_economic\_situation, 12)  
)

We apply this function to the trust variables. The harmonize\_waves() function binds all variables that are present in all surveys.

harmonized\_arabb\_waves <- **harmonize\_waves** (   
 waves = merged\_arabb,   
 .f = harmonize\_arabb\_trust )

Let’s look at the attributes of harmonized\_arabb\_waves.

h\_arabb\_structure <- **attributes**(harmonized\_arabb\_waves)

h\_arabb\_structure**$**row.names <- NULL *# We have over 70K row names*  
h\_arabb\_structure

To keep our example manageable, we subset the datasets to include only five countries.

harmonized\_arabb\_waves <- harmonized\_arabb\_waves **%>%**   
 **mutate** ( id = **strsplit**(unique\_id\_ab\_english, "[\_]")[[1]][1])

### Analyzing the harmonized data

The harmonized data can be exported and analyzed in another statistical program. The labelled survey data is stored in labelled\_spss\_survey() vectors, which is a complex class that retains metadata for reproducibility. Most statistical R packages do not know it. The data should be presented either as numeric data with as\_numeric() or as categorical with as\_factor(). (See more why you should not fall back on the more generic as.factor() or as.numeric() methods in [The labelled\_spss\_survey class vignette.](https://retroharmonize.dataobservatory.eu/articles/labelled_spss_survey.html))

Please note that the numeric form of these trust variables is not directly comparable with the numeric averages of the Eurobarometer trust variables, because the middle of the range is at 1.5 and not 0.5.

harmonized\_arabb\_waves **%>%**  
 **mutate\_at** ( **vars**(**starts\_with**("evaluation")),   
 **~as\_numeric**(.)**\***weight) **%>%**  
 **select** ( **-all\_of**("weight") ) **%>%**  
 **group\_by** ( country, id ) **%>%**  
 **summarize\_if** ( is.numeric, mean, na.rm=TRUE )

And the factor representation, without weighting:

**library**(tidyr) *## tidyr::pivot\_longer()*  
harmonized\_arabb\_waves **%>%**  
 **select** ( **-all\_of**(**c**("weight", "unique\_id\_ab\_english")) ) **%>%**  
 **mutate** (   
 evaluation\_economic\_situation =   
 **as\_factor**(evaluation\_economic\_situation)) **%>%**  
 **pivot\_longer** ( **starts\_with**("evaluation"),   
 names\_to = "indicator",   
 values\_to = "valuation") **%>%**  
 **group\_by** ( country, id, valuation ) **%>%**  
 **summarize** ( n = **n**() )

# Planned Outcomes

## Creating Harmonized Datasets

We have not started working on this. Before we can start working on this the[Desired outputs](#bookmark=id.4d34og8) have to be finalized.**Comparison with Afrobarometer**

See @rep (outcome-afrobarometer) [Comparison with Afrobarometer](#bookmark=id.3rdcrjn)

## Comparison with Eurobarometer

See @rep (outcome-eurobarometer) [Comparison with Eurobarometer](#bookmark=id.26in1rg)

## Technical Report

We believe that the tutorial, the larger **Harmonized Arab Barometer files**, two more experimental use cases to created an Arab Barometer / Afrobarometer and Arab Barometer / Eurobarometer joined, harmonized dataset will lead to enough conclusions, and hopefully datasets that have their own scientific value.

Because Daniel’s validation deadline with the TU Delft Validation Lab (see [disclosures](#bookmark=id.49x2ik5) below), provided his team gets there, is around 15 November 2020, we would like to finalize the Technical Reports by early November.

# Annex

## Practicalities

This document is created in [RMarkdown](https://rmarkdown.rstudio.com/).

Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents, or even PowerPoint, iWork Keynote, Google Docs, html, epub and MOBI books, or various PDF formats conforming journal or book publisher guidelines (with further conversion steps.) (See [The Markdown Guide](https://www.markdownguide.org/)) For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

RMarkdown is special, because it allows the insertion of little code snippets in the R, Python or C++ language, and the insertion of Latex tables or html elements. If you do not write code, you can just ignore these so-called chunks in the text. The markdown text is just a simple text file where formatting is made with explicit text symbols.

The easiest to work with markdown or RMarkdown documents is [Rstudio](https://rstudio.com/). It is an integrated development environment for programming, and if you do not write code, then you will simply ignore most of its functions. Its markdown editor window is simply a markdown word processor that can convert your simple markdown text to various files formats, e.g. Word, html, epub, PDF or via various converters into many-many file formats in a consistent manner. There are many markdown editors available for free, but RStudio is one of the best, with a good spell checker, and many quick editing functions. If you don't use them, you can simply minimize the window panes that are only used for programming.

However, if you do write code, Rmarkdown is probably the IDE for data science, because it allows the simultaneous use of little code snippets in the R, Python, C++ and SQL languages, and the insertion of Latex tables or html elements. True, it is mainly an R language IDE, but by enabling the use of lower-level, compiled code, and SQL (which is anyway very well harmonized with R’s tidyverse) you can facilitate teamwork among colleagues who use different languages, or want to use particular open source libraries that are only available in one of these languages.

When you click the **Knit** button in RStudio, a document will be generated that includes both content and the output of any embedded R code chunks within the document. **Knit** offers a dropdown menu of available file formats such as PDF or Html or MS Word. You can embed an R code chunk like this:

**summary**(cars)

Again, if you do not write code (yet), you can just simply ignore the chunks. Or, by pressing the play button, you can see what they do.

You can also embed plots, for example:

### Cooperation

If you use [github](https://github.com/), then you can directly edit this document in the original markdown format (and if you write R code, then you can add, modify or comment the code), which will then be converted to the Word Doc, Google Doc, PDF, epub formats. Our experience with technical documentation shows that they are the most usable in html mini website or intranet formats, because that makes the structure and hypertext links more transparent.

* Working in the shared Google Doc: This is in fact an MS Word version of this document, which can be converted in Google Doc to Google Doc format. The advantage of working in Google Doc is that several people can edit, comment, question the files. The disadvantage is that the agreed changes and insertions / deletions must be re-created manually in the RMarkdown .Rmd file.
* Working directly in the .Rmd file: The advantage of editing directly the .Rmd file is that simultaneous changes can be made by several contributors, which can be automatically merged and added to all html, epub, doc, formats. Of course, editing, commenting the R code is only possible in this way.

The documentation is present in the <https://github.com/dataobservatory-eu/arabbarometer> private repo. Its private status means that it is available for contributors who were granted access by Daniel, or indirectly, via Michael’s instructions. It is a safe, private working environment that is not visible for the rest of the world. If you contribute, please fork it and send pull requests instead of directly working in this repo.

Daniel as [antaldaniel](https://github.com/antaldaniel) gave access to:

**Arab Barometer team**:

**On the Arab Barometer side — Mohamed, Michael did not find you on github, let me know if you use github and wish to get access.**

* [Michael Robbins, PhD](https://www.arabbarometer.org/experts/michael-robbins-phd/). Michael Robbins is Director of the Arab Barometer.
* [Mohamed Abufalgha](https://www.arabbarometer.org/experts/mohamed-abufalgha/), Research Data Specialist, Arab Barometer.
* [MaryClare Roche, Phd](https://www.arabbarometer.org/experts/maryclare-rochephd/), Senior Research Specialist, Arab Barometer, on github [MCRoche](https://github.com/MCRoche).

**Reprex team**:

* Reprex’s co-founder Réka Szentirmay [szentrek](https://github.com/szentrek) generally has access to all our private repos. She works with data journalism partners, and as a curator, to bring our data narratives and visualizations into socially engaged arts scenes. She is not expected to participate other than review deliveries on our side, unless later some communication of results makes her involvement practical.
* Our project manager, Zuzana Gombiková [zuzanagombikova](https://github.com/zuzanagombikova), works on the better automation of our reproducible workflows. She makes sure that our contributors and other R fluent team members will on time sufficiently test any new code or addition made to the [retroharmonize](https://retroharmonize.dataobservatory.eu/) package or specific deliverables. She is expected to join in some organizational calls, and advise on any research workflow recommendations of the planned outcome [Technical Report](#bookmark=id.3whwml4), but she is not doing any programming or data work.

It is also present in the Google Shared folder [Arab Barometer](https://drive.google.com/drive/folders/11ZBKae9buRqijbeMwrrwIyWmc5s1BCp_?usp=sharing) with the same people.

Invitations were sent to reka.szentirmay@dataobservatory.eu, zuzana.gombikova@dataobservatory.eu mdr7@princeton.edu, mcroche@princeton.edu, mema@princeton.edu from antaldaniel@gmail.com.

*We do not know what is your preferred workflow / cloud storage, and Google Drive certainly is not our first choice. Our team usually uses* [*keybase.io*](https://keybase.io/team/reprexcommunity)*, which has a much more secure and private shared file system, but we can stay only on github or any format that you prefer.*

If you want to raise a quick question or avoid email queues, you can find the Reprex team on [keybase](https://keybase.io/team/reprexcommunity), the discrete open-source, crypo chat / messaging / social media platform.

### Long-form documentation

This document is created with the [bookdown](https://bookdown.org/) extension and package of RStudio. Bookdown allows users to *knit* together several Rmd files and publish them in an easy-to-read, intranet site or website format. Each chapter of this long-format documentation is a separate Rmd markdown document, and as we progress with the project, we will write up the later chapters.

## Disclosures

These disclosures are intended only for internal use at this stage. As we are finalizing outputs that may become public, we can review or amend them whenever disclosures must be made to the public.

### Disclosures by Daniel & Reprex

Daniel Antal, CFA and Réka Szentirmay are co-founders of [Reprex B.V.](https://dataobservatory.eu/), a start-up (founded on 1 September 2020) that is seeking validation for its open source, open data, open collaboration based approach to reproducible finance. We are finalists in the 2020 intake to the AI+Blockchain Validation Lab of Yes!Delft, which is considered to be the most successful university-backed (non-profit) incubator program in Europe. While Reprex is technically a for-profit entity, and currently it has no paid employees, the aim of the TU Delft’s validation is to find out if the company is viable in the Dutch, Benelux or broader European scientific and business research ecosystem. This validation period lasts up till 15 November 2020, till Reprex must find out on what conditions can it engage with various entities for research projects on a paid basis.

Marta Kołczyńska, together with Daniel, are co-authors of the GPL-3 open source [retroharmonize](https://retroharmonize.dataobservatory.eu/) R package which is currently under peer review, furthermore, co-author of the [eurobarometer](https://eurobarometer.dataobservatory.eu/) and which aims to may harmonizations more reproducible in Eurobarometer’s contexts. Marta originally intended to join Reprex as a team member, but currently has no affiliation, contractual obligations or shares in Reprex, and this situation will not change till 15 November 2020. As an academic researcher and grantee of the Polish National Science Centre, Marta must abide by the Code of the National Science Centre on Research Integrity and Applying for Research Financing, which regulates issues around research, professional conduct, attribution of research output, and communication of research to diverse audiences.

Because of her significant intellectual input into this project, Daniel wants to leave open for Marta to join this collaboration, even though Marta currently does not see her role in it. She will be given the option to review this document, and pre-publication outcomes, and give her contribution.

#### Nature of cooperation & thical standards

In the Netherlands, not only a statutory minimum wage, but education / experience / sector specific fair wage rules apply. That means that Reprex will not be viable if it cannot pay its internal / external staff fair wages in 2021. However, if it is found viable, it will be admitted to a non-profit incubation program financed by TU Delft and Dutch public grants.

Daniel and Reprex are not contractors neithersubcontractors of Arab Barometer, and they declare that the work foreseen in this project until 15 November 2020 is *not significant and is not made against remuneration*, with the exception that they will seek a non-significant, nominal remuneration in the context of the [TU Delft Validation Lab](#bookmark=id.3o7alnk) program. Getting this nominal remuneration is not a condition of their work, and if they do not get it, it will have no effect on the quality of their work.

Daniel as a CFA charterholder must abide by the CFA Institute Code of Ethics and Standards of Professional Conduct promoting the highest standards of ethics, education, and professional excellence for the ultimate benefit of society. herefore, our company is bound by these standards even when our services are not directly supporting investment decisions. It also regulates research financing, attribution of research output and communication to diverse audiences similarly to scientific integrity codes. The CFA ethical code and professional standards are aiming to be global standards, and dictate that whenever national regulations require higher ethical or professional standards, the higher standards must be followed.

In their cooperation, and in case of doubt, the Arab Barometer team and Daniel must decide if the Arab Barometers ethical guidelines or the CFA ethical guidelines dictate more strict restrictions, and agree to follow the rules that are less permissive. It is not foreseen that such an ethical conflict would apply here. The CFA standards on objectivity, attribution, and other values are practically identical to the standards applied in ethical and high quality academic research, and the nature of the data used in this project will not be applicable directly in an economic or investment advisory context. Even in that case, the requirements for proper disclosure on the validity and limitations of findings, fair data presentation and other similar topics are very similar to the presentation of ethical academic research.

#### Connection to TU Delft

Reprex B.V. is currently participating in the prestigious AI+Blockchain Validation Lab of Yes!Delft, which is considered to be the best university-backed incubators in the world. It is a non-profit incubator program of the Technical University of Delft, the municipality of the Hague and the charitable foundation of the Dutch royal family. It helps startups to commercialize scientific achievements. An important consideration of the validity of Reprex as a startup using scientific results is that it can find business, academic or journalistic users that are willing to pay for its work.

If the delivered results on behalf of Daniel & the Reprex team have scientific or operational merit, we would also value any feedback on the commercial viability of Reprex’s work in scientific research, particularly as data experts in academic research projects, which we would like to use in our Validation Assessment before 10 November 2020. This would be invaluable to us, because we already received valuable (and positive) feedback in Europe, but we have no feedback from outside the European Union.

TU Delft will encourage us to aim to get paid for this work, which is optional and can take form as the smallest amount that can be paid by Arab Barometer for not budgeted but valuable work. Our evaluator’s believe that start-ups often find it easy to get letters of recommendations, but processing however small invoiced sum is a firmer feedback on the viability of their enterprise.

#### Disclosure of other interests

We do not foresee any conflict of interest, and this disclosure is aimed at highlighting any issues that may *potentially* lead to ethical issues that must be resolved applying the highest standards.

Reprex B.V. is a for-profit entity in its very early, validation phase, with no employees on the payroll or substantial income. According to Dutch law, in this early phase, Daniel is technically on a payroll but is paid only after the company makes any income. Reprex does not have any income or work that may question the objectivity of the work made with Arab Barometer.

It has a forming agreement as a subcontractor of the University of Amsterdam IViR research institute, handling their data and surveying work, which is not directly or indirectly related to the Arab Barometer program.

Reprex was contacted by a researcher at Kantar, the producer of Eurobarometer, and is currently waiting for managerial approval to start a small project with Kantar’s internal R user group. We believe that this created no conflict of interest, on the contrary, it may have some extra benefits if Arab Barometer’s (otherwise public) survey files and Eurobarometer’s (otherwise public) survey files are professionally compared.

They are seeking similar engagements with several European universities (and commercial policy / business research entities, and a J++, a data journalism organization) using our technology which are usually not survey-focused and are even less likely to cause any conflict of interest with Arab Barometer.

Reprex B.V. is solely financed by its two natural person co-founders, Daniel Antal and Réka Szentirmay, Hungarian citizens residing in the Netherlands, and they are seeking to participate in the AI+Blockchain Validation Lab of Yes!Delft with the intention of validating the feasibility of their business plan. Their business plan relies on successfully acquiring paid engagements in the business research, academic research and journalistic research communities. They are in the procedure and passed the first phase successfully.

Templ, Matthias, and Valentin Todorov. 2016. “The Software Environment R for Official Statistics and Survey Methodology.” *Austrian Journal of Statistics* 45 (March): 97–124. <https://doi.org/doi:10.17713/ajs.v45i1.100>.