The Journal of Open Source Education

Datos: datasets translations to promote diversity, equity, and inclusion

Riva Quiroga[1], **Mauricio Vargas**[1], **Mauro Lepore**[2], **Rayna Harris**[3], **and Daniela Vasquez**[4]

1 Pontifical Catholic University of Chile **2** 2 Degrees Investing Initiative **3** University of California, Davis **4** R-Ladies Montevideo

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Summary

'R for Data Science' (Wickham & Grolemund, 2016) is a hands-on book used by many to learn the fundamental of the R language. However, many Spanish speakers struggle to use this book as a resources because of the English language barrier. To address this gap in accessibility, Quiroga et al translated the book to Spanish ('R para Ciencia de Datos' [FIXME instert reference]). Additionally, we created the R package `datos::` to automatically translate datasets from English to Spanish using computational tools already existing in both R Core Team (2020) and Wickham et al. (2019). Together, the book 'R Para Ciencia de Datos' and the `datos::` package allows Spanish speakers to spend their energy not in understanding English but in learning data science in R.

Diversity, equity and inclusion

'R for Data Science (R4DS)' (Wickham & Grolemund, 2016) and the 'R Cookbook' (Teetor, 2011) both provide context and detailed examples for different R functions. However, understanding this context, written in English, incrases the cognitive load required to learn the R language. Both in Latin America and Spain, the lack of a key English proficiency in large groups of population, constitutes a large learning barrier that has its roots in economic inequality and access to education. Some peple may ask 'why don't they learn English?'. That is a possibility for the few, but not for all. Just as an example, in the case of Chile, an elementary level English course costs around 500 USD/month while the minimum wage is 350 USD/month. Rather than placing the burden of learning English on the learner, we, the community leaders and educators can take action to reduce the language barries with social and technological solutions.

Implementation

The `datos::`package makes use of YAML specifications to automatically translate data sets originally available in other R packages. The translated data can be used together with R4DS book or independently as a source of practice data in Spanish. The YAML specification for each dataset that provides the dataset name, how you want to translate the variables, and the description for the documentation. This process not only gets the dataset translated, but also the help page for the dataset, which is very useful for people who are learning. `datos::` translates the datasets on the fly, thanks to `delayedAssign()` from base R, so the datasets are not in the package, as it just contains YAML files

with translation specifications and functions that translate the datasets called from other packages.

As an example, let's inspect the first rows of the `airlines` table from `nycflights13::`. This dataset has two columns `carrier` and `name`, which provide a two-letter abbreviation and the full name of the airline.

```
head(nycflights13::airlines)
```

```
## # A tibble: 6 x 2
##   carrier name
##   <chr>   <chr>
## 1 9E      Endeavor Air Inc.
## 2 AA      American Airlines Inc.
## 3 AS      Alaska Airlines Inc.
## 4 B6      JetBlue Airways
## 5 DL      Delta Air Lines Inc.
## 6 EV      ExpressJet Airlines Inc.
```

This is the specification for the `airlines` table from `nycflights13::`: Here, we provide both a translation (trans:) and description (desc:) in Spanish as well as additional helpful information.

```
df:
  source: nycflights13::airlines
  name: aerolineas
variables:
  carrier:
    trans: aerolinea
    desc: "abreviaci\u00f3n de dos caracteres del nombre de la
     aerol\u00EDnea"
  name:
    trans: nombre
    desc: "nombre completo de la aerol\u00EDnea"
help:
  name: aerolineas
  alias: aerolineas
  title: "Nombres de aerol\u00EDneas"
  description: "Nombres de aerol\u00EDneas y su respectivo c\u00f3digo
   carrier de dos d\u00EDgitos."
  usage: aerolineas
  format: Un data.frame con 16 filas y 2 columnas
```

Conclusion

We the people of the R Community, in order to form a stronger and more integrated community, ease learning, promote diversity, and secure the development and usage of the R Programming language, do need to acknowledge that a large language gap exists and it prevents a large number of users from accesing the existing good quality materials created by and for ourselves. The solution to close the gap resides in the R Community itself, not in software. Our software can be used to start translating well-known R datasets and R4DS itself into other languages without reinventing the wheel. What we did is not merely translating a book and creating a package. We created the human and technical infrastrucure to shorten the language gap. Our process, which resulted in brand new errors, which leaded us to find ways to make our community stronger.

The Journal of Open Source Education

Acknowlegments

We are grateful to our colleagues in R-Ladies, R Users Groups, rOpenSci, and The Carpentries for their perspectives and support.

References

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Teetor, P. (2011). *R cookbook: Proven recipes for data analysis, statistics, and graphics*. O'Reilly Media, Inc.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi:10.21105/joss.01686

Wickham, H., & Grolemund, G. (2016). *R for data science: Import, tidy, transform, visualize, and model data*. O'Reilly Media, Inc.