37 Managing Data and Statistical Code According to the FAIR Principles

Laura A. Janda

1 Introduction: The FAIR principles

In linguistic studies that draw on empirical data and statistical code, the linguistic community needs full access not just to the scholarly article or report of the findings but also to the data and statistical code that stand behind them. Ideally researchers should post a resource that includes all of the researcher's data as well as the necessary tools for interpretation and analysis in a format that is accessible, transparent, and explicit enough to allow a second researcher to carry out the same analysis on that data or on similar data. The idea is simple, but making it work in practice requires the author of a study to take several issues into consideration.

This chapter contains recommendations for linguists working with data sets that are analyzed with statistical code to conform to the FAIR principles for data management (cf. Wilkinson et al. 2016):[1]

- Findable: Attaching metadata to a post that is indexed in an archive makes it possible for other researchers to find the data and code.
- Accessible: Using an archive that is publicly available makes it possible for other researchers to access resources that contain the data and code.
- Interoperable: Using file formats that are persistent and open-source makes it more likely that other researchers will be able to open and use the files.
- Reusable: Providing adequate description of the files, their contents, the variables, and their values, as well as annotation for the statistical code, will make it possible for other researchers to understand and reuse the data and code.[2]

Adherence to the FAIR principles with regard to data and code for statistical analysis entails additional work, but there are benefits both for the scholar and for the field as a whole. Linguistics is a relative late-comer to the scientific frontier of replicable research, and our performance thus far leaves a lot to be desired (see Gawne & Styles, chapter 2, this volume). Berez-Kroeker et al. (2017) and Gawne et al. (2017) surveyed linguistics articles, dissertations, and descriptive grammars published in 2003–2012 and found that for most journals, fewer than 50% of authors provided descriptions of methodology or citations of the sources of linguistic examples. Archiving of data was likewise found lacking among linguists: only twelve of fifty dissertations and ten of fifty grammars had archived data at the time of publication.

We can all contribute to the prestige of our field by setting ethical standards for best practices in data management and collaboration. Posting reusable data and code fosters horizontal learning across the community of researchers, facilitating the propagation of new methods. It is also a safeguard against research fraud. Under pressure to get published in prestigious journals, some researchers have fabricated or fudged data to impress reviewers, a problem that has been particularly acute in the field of medicine (cf. Fanelli 2009). Publication of data and code will not eliminate the possibility of fraud, but it will make fraud easier to detect, and, if practiced regularly, will improve the overall level of accountability and integrity in the field. Such posts can additionally help eliminate unnecessary duplication of efforts, because researchers can more easily discover whether a given study has been conducted. And there is also the possibility that a set of data and/or statistical code will have unforeseen uses, analogous to Velcro, which was developed for use by astronauts in space and is now common in clothing and widely used by earthlings.

The benefit of integrity builds the reputation of the individual researcher, because backing up arguments

with the facts they are based on contributes to the legitimacy of one's claims. Furthermore, one never knows when one might need to go back to a data set, either to squeeze another analysis out of it or to use it as a recipe for a parallel analysis of new data. If the data and code are adequately annotated and archived, this will be easy to do, even if time has erased the meanings of all the abbreviations and code that were so obvious when working on the original project from the researcher's personal memory. After several mishaps with data that were lost, corrupted, or otherwise became uninterpretable, along with discontinuance of site licenses and backward-incompatible software upgrades that permanently separated me from my data and analyses, I have come to appreciate the value of things that are open-access, open-source, and generally portable across time and platforms (see Mattern, chapter 5, this volume, and Collister, chapter 9, this volume).

Here I present two examples from resources that I have created in the spirit of adhering to the FAIR principles. I think that I have been partly successful in these examples, but I have also made my share of mistakes. I present these two resources "warts and all" and point out some places where I could have done a better job. These resources (also referred to as posts) are available in the TROLLing archive: the Tromsø Repository of Language and Linguistics (https://dataverse.no/dataverse/trolling), which I use as an example in this chapter (see also Andreassen, chapter 7, this volume). As a discipline-specific archive, TROLLing offers added value to the post (Alter & Gonzalez 2018:149) and comes with three features that are particularly desirable for archiving linguistic data and statistical code: (1) public access, (2) professional management, and (3) harvestable metadata. In addition, TROLLing adheres to the FAIR principles and assists authors in achieving these goals.

2 A resource that adheres to FAIR principles, but is not user-friendly

The first resource that I use as an example is available at https://doi.org/10.18710/4XTXMH. It was created in connection with Janda and Antonsen (2016), a study of a language change in which a synthetic possessive construction with a possessive suffix is being replaced by an analytic construction with a reflexive pronoun in the North Saami language. We collected a data set with over two thousand examples across three generations of speakers and performed a statistical analysis to determine the influence of various factors on the ongoing change. We also made some graphs to visualize our data and results. Next I demonstrate how this resource adheres to FAIR principles.

Findable: Because the metadata can be searched, data and code become findable when posted in an archive with harvestable metadata. In the metadata section of this post I added Keywords: "North Saami, possessive suffix, Uralic, S-curve, vocative, possession, language change, complexity"; and selected Topic Classifications: "morphology, diachronic, affixes." Kind of Data is listed as "corpus." Other kinds of data could include "questionnaire" and "experiment," and in both cases one would need to make sure that no personal identifiers are present so that the data can be publicly archived. In the case of experimental data, it would be useful to include files representing the stimuli. This TROLLing post additionally lists the Geographic Coverage as: "Norway, Sweden, Finland, Sápmi, Northern Scandinavia."

Accessible: A unique identifier and URL make data and code accessible. The first thing that happens when a user initiates a TROLLing post is that the site generates a unique identifier and URL for the data set (previously this was achieved with a "handle," or hdl, as in this example, but now the system produces a digital object identifier, or DOI, for each post). The unique URL exists right from the start and persists no matter how many times the TROLLing post is updated, although it is not visible to others until it is submitted. The advantage to this is that the TROLLing post can be cited in an article, in addition to citing the article in the TROLLing post, so cross-referencing is achieved in both directions. However, it is not necessary to have a published article to create a post. Any data set can be posted, regardless of whether there is an accompanying article. This includes data representing so-called negative results, which can be challenging to publish, but can be very valuable to the greater scientific community nevertheless.

Interoperable: The use of appropriate formats makes data and code usable across platforms. A crucial issue for sharing data and code is that they be presented in formats that can be accessed by anyone and at any future time (see Mattern, chapter 5, this volume). For this reason, it is best to select persistent non-proprietary open formats such as .pdf and .txt, as required in the TROLLing repository (cf. TROLLing guidelines, https://site.uit

.no/dataverseno/deposit/prepare/#what-are-persistent-file-formats; Andreassen, chapter 7, this volume). This will hopefully safeguard data and code against a situation that would lack any device or software needed to access them. Using such formats can mean losing the direct connection to software that comes along with an item such as an R script. This can be handled by providing duplicate files in different formats.

Reusable: In addition to the information in the metadata, particularly the Description and the citation of the related article, the next most important item is a readme file. Ideally the readme file should explain what all the other files in the post are, providing details of the data files including all of the factors and what their values are. In TROLLing, it is also possible to attach a description to each file. Providing details about data and code can be a tedious process, but it is essential to make them reusable.

In the post about North Saami possessive constructions I did some things right and there were other things I could have done better. I did make a readme file in a persistent open format (Readme file for DiachronicaCART.txt), but that readme file is not the first file that a user sees—it is the seventh of ten because TROLLing lists the files in alphabetical order according to their titles. I have since learned that one should give the readme file a name like "01Readme.txt" so that it will be put at the beginning of the list. Another problem with my readme file is that it does not provide a guide to all the files in the post; instead I relied on the descriptions attached to each file. While most of these descriptions are reasonably straightforward, such as "This is the R code for the CART analysis," the fact that the files have been alphabetized means that the logical order of the files is a bit jumbled in addition to the fact that information has been scattered across the ten files. A strategy to improve this would be to give all the files names starting with "01," "02," and so on. But even so, it would be better to have all of the file descriptions in one place connected in a coherent way, namely in a "01Readme" file.

My readme file describes the factors and values used in one file that has a lot of factors and values, namely the one called DiachronicaCART.csv. That description is quite thorough, as shown in this example:

> Column A/PossCon: This column represents the possessive construction (PossCon), and the values are NPx (noun + possessive suffix) and Refl (analytic construction with reflexive genitive pronoun).

Given this information, most researchers should be able to decipher all of the abbreviations used to name both the factors and their values.

Worse, however, is the fact that the data in DiachronicaCART.csv are just the annotations for the sentences that we analyzed. The sentences themselves are presented in a file called AnnotatedSentences.txt, which has this attached description in the TROLLing post:

> This file contains the sentences that constitute our database, along with their annotations. Most of the annotations are explained in the Readme file for Diachronica CART. This file additionally cites the works of the authors that the sentences are taken from and gives some additional details concerning the semantic classes of possessums.

It is possible to identify most of the annotations from the readme file, but only because the first eleven factors appear in the same order and use all the same abbreviations. However, after that we have some additional information, the name of the literary work and the page number where the example is found, inserted before the last two factors are listed. In other words, the information is there, but it is not very user-friendly.

Another problem with DiachronicaCART.csv is that it is in .csv (comma-separated values) format, which has some quirks. This format is not fully standardized, it runs into problems with data fields that contain commas, and it can include data that use other marks of separation such as semicolons. A better alternative is tab-separated values (.tab or .tsv), because this format is more widely supported. One possibility would be to save duplicate data sets in both formats to preserve some benefits of the original file and also protect the file against future incompatibilities.

Our article about North Saami possessive constructions (Janda & Antonsen 2016) contains three figures, and our TROLLing post presents both the data and the code needed to produce these figures. Figure 1 in our article shows the longitudinal development of the language change. The relevant data for figure 1 in our article is presented in Scurve.csv, and in the TROLLing post I give details about the columns and values in the description attached to this file, because in this case there are only five columns and those contain values that are easy to describe (names of authors, year of birth). ScurveCode.txt, as shown in sample file 37.1, contains the R code used to generate the plot (labeled figure 1 in our article) from Scurve.csv.

```
Code for S-curve
> Sdat=read.csv(file=file.choose(), header=T)
#Choose Scurve.csv
> print(Sdat)
name year NPx Refl PropRefl
1 A Larsen 1870 132 11 0.07692308
2 J Turi 1895 88 24 0.21428570
3 KN Turi 1895 38 3 0.07317073
4 HA Guttorm 1907 250 9 0.03474903
5 M Bongo 1923 14 1 0.06666667
6 AO Eira 1927 31 3 0.08823529
7 JA Vest 1948 498 365 0.42294320
8 K Paltto 1947 114 50 0.30487800
9 EM Vars 1957 152 153 0.50163930
10 JM Mienna 1972 17 68 0.80000000
11 MA Sara 1983 49 63 0.56250000
> plot(Sdat$year,Sdat$PropRefl, type= "n", xlab="Year
of Birth", ylab="Proportion of ReflN")
>
text(Sdat$year,Sdat$PropRefl,as.character(Sdat$name),
cex=0.7)
> lines(lowess(Sdat$year,Sdat$PropRefl))
```

Sample file 37.1
ScurveCode.txt.

While this is an accurate representation of the statistical code needed to plot figure 1 in our article, there is not much annotation here (the only annotation is #Choose Scurve.csv, telling the user which file to use as input data) and this file has not been set up in such a way that it can be fed directly into R (one would need to delete the > characters and add # before all the rows of the table of Sdat). It would have been desirable to add annotations explaining what the last three lines of code do, namely create the plot, add the text, and add the locally weighted scatterplot smoothing line. The file Scurve.pdf shows what the result should look like (our figure 1). And because all three of these files begin with "Scurve" they are conveniently listed together.

The representation of the R code for the classification and regression tree (CART) analysis is somewhat more successful in the sense that it can be directly fed into R and is given in both .R and .txt file formats (DiachronicaCode.R, DiachronicaRcode.txt). But again, there should have been more annotation, particularly to identify the parts of the code that produce the plots that are in figures 2 and 3 in our article. Furthermore, the names of the files could have been more informative, something like "CARTanalysis.txt." "Diachronica" here refers to the journal that the article was published in, a piece of information that was useful to me on my personal computer, but is not so useful for another user.

3 A resource that is somewhat more successful

The second example is of the TROLLing post at https://doi.org/10.18710/VDWPZS for data and statistical code from Janda and Tyers (2018), an article about Russian paradigms. This post both conforms to the FAIR principles and does a better job in terms of user-friendliness. Sample file 37.2 shows an example of an R script in this TROLLing post for which I provided better annotation:

```
#This R script is supplementary to the article
#"Less is More: Why All Paradigms are Defective, and
Why that is a Good Thing"
#by Laura A. Janda and Francis M. Tyers
#This script shows how to create the plot in Figure 1:
#Correspondence Analysis for Masculine Animate Lexemes
#The same code can be used (just changing the names of
the files)
#to make similar plots for the other groups of nouns.
#This code also shows how to get the data for Tables
4a–b and 5.
#First load the languageR package.
#Note that if you do not have that package, you will
need to install it first.
library(languageR)
#Then load the data that we need:
mascanim<-read.csv(file.choose(), header=T)
#We name this data mascanim because it shows the
grammatical profiles
#of masculine animate nouns.
#Choose this file: procent-I-m.aa.csv
#Notice that this same code can be used to load any of
the other datasets,
#but of course each one should get a corresponding
name.
#Check to make sure that this file has loaded
correctly:
head(mascanim)
#There is one item that has accidentally been included
in this file,
#кто-то 'someone' is actually a pronoun and needs to
be removed.
```

```
#You can see it here:
mascanim[mascanim$lemma=="кто-то",]
#Notice that this same code can be used to look at the
grammatical profile
#of any noun in the dataset.
#This tells us that кто-то is on line 20, so that is
the line to remove.
#Remove it with this code:
mascanim<-mascanim[-20,]
#Now we need to subset the data to take only the
columns that we need
#for the correspondence analysis, namely the lemma
(column 2) and
#the grammatical profile (the percentages listed in
columns 5 through 16).
#Here is how we do that:
mascanimdata <- mascanim[, c(2, 5:16)]
#Now we need to make this into a dataframe:
mascanimdataframe <- data.frame(mascanimdata)
#Now we run the correspondence analysis:
mascanimdataframe.ca <-
corres.fnc(mascanimdataframe[2:13])
#And then we plot the result:
plot(mascanimdataframe.ca,
rlabels=mascanimdataframe$lemma, rcex=0.75)
#This is the plot found in Figure 1.
#Note that you can use the following code to get the
values for
#Factor 1 and Factor 2 from the correspondence
analysis:
mascanimCoor=attr(mascanimdataframe.ca,
"data")$origOut$rproj
#You can see what this looks like here:
head(mascanimCoor)
#And you can add these into the dataframe so that they
are aligned
#with the lemmas:
mascanimdataframe$Factor1 <- mascanimCoor[,1]
mascanimdataframe$Factor2 <- mascanimCoor[,2]
#Now they are in your dataframe, as you see here:
head(mascanimdataframe)
#It is also possible to order lemmas according to
their
#Factor 1 values, for example:
mascanimdataframe[order(mascanimdataframe$Factor1),]
```

Sample file 37.2
Paradigms.R (also presented as Paradigms.R.pdf).

This annotation is far more explicit, even giving directions that would enable the reader to engage further with the data and analysis. However, this post is not perfect either. The name of the file is not very helpful; "Paradigms" merely tells the user that this is a file related to paradigms, which is what the whole post is about. It would have been better to give this file a more descriptive name, like "RScriptForCorrespondenceAnalysisOfMasculineAnimateParadigms.pdf."

4 Closing recommendations

Solid data management principles are fairly new to most linguists, something that we are just beginning to wrap our heads around. In this situation, we are better off helping each other out and learning from each other. Gone are the days of the solo linguist like Mr. Higgins in his cozy library. For many scholarly works today, no single individual can command all of the relevant areas of expertise. Making linguistic data and code accessible according to FAIR principles is one way to promote collaboration and raise competence within the field. And there is evidence that this actually works: a couple of years ago while serving on a dissertation committee, I discovered that the candidate had downloaded one of my TROLLing posts and used it as a model for the analysis of his own data.

Because we all have limited time and energy, making data and code reusable can seem like one more burden. And even when one takes on this extra burden, like so many other things, every time one goes back to a post, it is possible to identify ways in which it could have been improved. Hopefully this set of guidelines will convince other researchers that it is worth trying and will help to streamline the process and avoid pitfalls.

To make these tasks more manageable, one could begin working on documentation early in the research process, with a data management plan (see Kung, chapter 8, this volume), and make sure that readme files and annotated R scripts are created early on and updated periodically, rather than tackling the whole job after a publication is accepted. Giving files informative names and ordering them in a logical fashion is helpful. All files should be presented in persistent open-source formats and should be archived in a public discipline-specific archive that collects harvestable metadata in conformity with scholarly standards.

Notes

1. Brief statements of the FAIR principles are also available at https://www.force11.org/group/fairgroup/fairprinciples and https://www.go-fair.org/fair-principles/.

2. I have opted to use *reusable* instead of "reproducible" or "replicable" as it is the term adopted in the FAIR principles statements.

References

Alter, G., and R. Gonzalez. 2018. Responsible practices for data sharing. *American Psychologist* 73 (2): 146–156. http://dx.doi.org/10.1037/amp0000258.

Berez-Kroeker, Andrea L., Lauren Gawne, Barbara F. Kelly, and Tyler Heston. 2017. *A Survey of Current Reproducibility Practices in Linguistics Journals, 2003–2012*. https://sites.google.com/a/hawaii.edu/data-citation/survey.

Fanelli, Daniele. 2009. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS One* 4 (5): e5738. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0005738.

Gawne, Lauren, Barbara F. Kelly, Andrea L. Berez-Kroeker, and Tyler Heston. 2017. Putting practice into words: The state of data and methods transparency in grammatical descriptions. *Language Documentation and Conservation* 11:157–189. http://hdl.handle.net/10125/24731.

Janda, Laura A., and Lene Antonsen. 2016. The ongoing eclipse of possessive suffixes in North Saami: A case study in reduction of morphological complexity. *Diachronica* 33 (3): 330–366. http://dx.doi.org/10.1075/dia.33.3.02jan.

Janda, Laura A., and Francis M. Tyers. 2018. Less is more: Why all paradigms are defective, and why that is a good thing. *Corpus Linguistics and Linguistic Theory* 14 (2). doi.org/10.1515/cllt-2018-0031.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3:160018. https://doi.org/10.1038/sdata.2016.18.

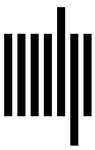
The Open Handbook of Linguistic Data Management

Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

Citation:

The Open Handbook of Linguistic Data Management
Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister
DOI: 10.7551/mitpress/12200.001.0001
ISBN (electronic): 9780262366076
Publisher: The MIT Press
Published: 2022

The open access edition of this book was made possible by generous funding and support from the authors



The MIT Press

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Berez-Kroeker, Andrea L., editor. | McDonnell, Bradley James, editor. | Koller, Eve, editor. | Collister, Lauren B., editor.
Title: The open handbook of linguistic data management / edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller and Lauren B. Collister.
Description: Cambridge, Massachusetts : The MIT Press, [2021] | Series: Open handbooks in linguistics series | Includes bibliographical references and index.
Identifiers: LCCN 2020044363 | ISBN 9780262045261 (hardcover)
Subjects: LCSH: Computational linguistics. | Natural language processing (Computer science) | Data mining.
Classification: LCC P98 .O64 2021 | DDC 410.285—dc23
LC record available at https://lccn.loc.gov/2020044363