The Austin Principles of Data Citation in Linguistics
Version 1.0

Linguistics Data Interest Group (LDIG)
May 3, 2018

Preamble

Data is central to empirical linguistic research. Linguistic data comes in many different forms, and is collected and processed with a wide range of methods. Data citation recognizes the centrality of data to research. Furthermore, it facilitates verification of claims and repurposing of data for other studies.

The FORCE11 Joint Declaration of Data Citation Principles* state that "[s]ound, reproducible scholarship rests upon a foundation of robust, accessible data. For this to be so in practice as well as theory, data must be accorded due importance in the practice of scholarship and in the enduring scholarly record. In other words, data should be considered legitimate, citable products of research. Data citation, like the citation of other evidence and sources, is good research practice and is part of the scholarly ecosystem supporting data reuse."

The FORCE11 Joint Declaration of Data Citation Principles is intentionally broad so to be as inclusive of data from as many scientific disciplines as possible. This document, the Austin Principles of Data Citation for Linguistics, interprets the FORCE11 document to address linguistic data specifically. These guiding principles have been created to enable linguists to make decisions about their data that ensure it is as accessible and transparent as possible. Some subfields of linguistics may already have specific guidelines for data citation; in these cases the Austin Principles can supplement extant guidelines to ensure that data citation conforms with current best practices.

Principles

The Austin Principles of Data Citation cover purpose, function and attributes of citations. These principles recognize the dual necessity of creating citation practices that are both human understandable and machine-actionable. They are not comprehensive recommendations for data stewardship. And, as practices vary across communities and technologies will evolve over time, we do not include recommendations for specific implementations, but encourage communities to develop practices and tools that embody these principles.

The principles are grouped so as to facilitate understanding, rather than according to any perceived criteria of importance. Text in italics is taken from the FORCE11 document.

1. Importance

Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.

Linguistic data form not only a record of scholarship, but also of cultural heritage, societal evolution, and human potential. Because of this, the data on which linguistic analyses are based are of fundamental importance to the field and should be treated as such. Linguistic data should be citable and cited, and these citations should be accorded the same importance as citations of other, more recognizable products of linguistic research like publications.

2. Credit and Attribution

Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.

In linguistics, citations should facilitate readers retrieving information about who contributed to the data, and how they contributed, when it is appropriate to do so. One way to do this is through citations that list individual contributors and their roles. Another way is by using citations that link to metadata about contributors and their roles.

3. Evidence

In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.

Linguists should cite the data upon which scholarly claims are based. In order for data to be citable, it should be stored in an accessible location, preferably a data archive or other trusted repository. Authors should ensure that data collection and processing methods are transparent, either through links to metadata or a direct statement in the text, to make clear the relationship between the data and the scholarly claims based on it.

4. Unique Identification

A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.

When selecting a data repository or other resources for housing and providing access to linguistic data, linguists should look for services that provide the means for identification in the form of a Persistent Identifier (PID). For digital data, examples of these include Digital Object Identifiers (DOI) and Handles.

5. Access

Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.

Linguistic data should be as open as possible, in order to facilitate reproducibility; and as closed as necessary**, to honor relevant ethical, legal and speaker community constraints.

6. Persistence

Unique identifiers, and metadata describing the data, and its disposition, should persist -- even beyond the lifespan of the data they describe.

Linguists should confirm that the archives or repositories where they are storing their data have written policies pertaining to persistence of data, metadata, and identifiers.

7. Specificity and Verifiability

Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verifying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.

Data citations should make it easy for a curious reader to find the specific datum or subset of data within the larger dataset that support a claim. For data uses that require a fine-grained citation for clarity, a systematic method of identification for the data should be used.
Many data sets are not static; rather researchers add to them all the time. Citations should specify which version of the data is being referenced.

8. Interoperability and Flexibility

Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities.

Linguists work with a wide range of data, addressing a variety of questions. Citation standards developed for linguistics need to meet the needs of the research community, while also meeting the principles described above. We encourage linguistics publishers to make data citation easier for their authors by developing data citation formats and to develop clear data policies based on this document.

* Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 https://doi.org/10.25490/a97f-egyk
** *As open as possible, as closed as necessary* is the approach taken by the European Commission in their Horizon 2020 Programme.

How to use the Austin principles

How to cite the Austin principles

Berez-Kroeker, A. L., Andreassen, H. N., Gawne, L., Holton, G., Kung, S. S., Pulsifer, P., Collister, L. B., The Data Citation and Attribution in Linguistics Group, & the Linguistics Data Interest Group. (2018). *The Austin Principles of Data Citation in Linguistics*. Version 1.0). http://site.uit.no/linguisticsdatacitation/austinprinciples/ Accessed [date]

BibTeX citation:

```
@Misc{AustinPrinciples:2018,
author = “Andrea L. Berez-Kroeker and Helene N. Andreassen and
          Lauren Gawne and Gary Holton and Susan Smythe Kung
          and Peter Pulsifer and Lauren B. Collister and {the
          Data Citation and Attribution in Linguistics Group}
          and {the Linguistics Data Interest Group}”,
title = “The {Austin} Principles of Data Citation in Linguistics”,
year = 2018,
note = “Version 1.0”,
url = “http://site.uit.no/linguisticsdatacitation/austinprinciples“}
```

How to comment on the Austin principles

All comments, questions and reflections related to the Austin Principles are highly welcome. We encourage you do to this openly in the comments field on the LDIG webpage: https://www.rd-alliance.org/groups/linguistics-data-ig. If you want to contact us directly, please send an email to lingdata@hawaii.edu.