Linguistics Data Interest Group (LDIG)

RDA Plenary 10
Sept 19, 2017

Session Notes Document:
http://bit.ly/LDIG-plen10-sessionnotes

LDIG Aims (from the LDIG Charter)

- Development and adoption of common principles and guidelines for data citation and attribution
  – researchers, professional organizations, academic publishers, archives
- Education and outreach efforts
  – practical training and awareness of principles/sociological change
- Greater attribution of linguistic data set preparation within the linguistics profession
  – value “data work” as scholarly output at all career stages

We're new! How did we get here?

These aims for LDIG grew out of 3-year NSF project *Developing Standards for Data Citation and Attribution in Linguistics* (NSF SMA-1447886)
3 workshops, 40+ international participants, 2 dozen presentations

Project Outcomes:

- Survey of data citation practices across 9 journals over 10 years (tl;dr almost nobody does it)
- Survey of data citation practices descriptive linguistics over 10 years (ditto)
- Position paper on data citation, to appear in *Linguistics* in early 2018
- Draft of Austin Principles of Data Citation and Attribution in Linguistics
- Formation of LDIG
    - Austin Principles is our first task, to facilitate sociological change in the field

What is linguistic data?

- All levels of language:
  - from recordings of individual words to hours of audio-video
  - from a single sentence to hours of conversation
  - From an individual speaker to a whole community
  - from any of the world's 7000 languages (many with no written standard)
- audio recordings (stories, conversations, interviews, elicitation), video recordings, xml annotations, transcripts, 'glossed' text, dictionaries, experimental data (eye tracking data, reaction time data etc.), introspection, tagged corpora, spectrograms, sonograms, GPS data...
- Archived with DELAMAN, institutional repository, on personal hard drives, in shoeboxes under the bed

Austin Principles of Data Citation: Background

- Goals
  - Encourage and improve visibility and retrievability of research data
  - Guidelines for formatting data citations (for creators and users of data, humans and machines)

- Based on the FORCE 11 Joint Declaration of Data Citation Principles

- Format
  - Standard, up-to-date requirements on data citation formatting
  - Guidelines for citation of elements considered important to (but probably not unique to) linguistics

Austin Principles of Data Citation

1. **Importance:** Data should be considered legitimate, citable products of research.

In linguistics, data might also form a record of cultural heritage, societal evolution, and human potential.

2. **Credit and attribution:** Data citations should facilitate giving scholarly credit and normative and legal attribution.

In linguistics, this might apply to any individual participating in data collection/creation, including native speakers, interviewees, and transcribers.

3. **Evidence:** In scholarly literature, when a claim relies upon data, the data should be cited.

In linguistics, the method of data collection should also be made apparent in the text.

Austin Principles of Data Citation

4. **Unique identification:** A data citation should include a persistent, unique, widely used type of identification.

In linguistics, many data repositories specializing in linguistic data offer such identification in the form of a Persistent Identifier (PID).

5. **Access:** Data citations should facilitate access, human and machine readable, to the data and related information.

Data should be as open as possible and as closed as necessary based on relevant ethical, legal and speaker community constraints.

Austin Principles of Data Citation

6. **Persistence:** PIDs, metadata and the data should persist.

Linguists should store data in archives with written policies stating the persistence of data and metadata.

7. **Specificity and Verifiability:** Data citations should facilitate identification of, access to, and verification of the specific data 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.

8. **Interoperability and flexibility:** Data citation methods should be sufficiently flexible, but no more than necessary, to accommodate the variant practices among communities

Austin principles of data citation: Current feedback

- Need to be applicable to all types of research in the various subfields of linguistics
  - Historical archival data, experimental sound recordings, syntactic acceptability judgments
  - Are specific examples the way to go?

- Need to provide guidelines about coauthorship and attribution practices
  - Who to cite? The researcher, the field worker, the research assistant, the informant?
  - Is PID and link to a well constructed metadata template the way to go?
  - To keep in mind: 1) The necessity and desire to credit various participants will vary from project to project. 2) Granularity in reference vs. specified in-text citations, 2) Criteria for evaluating potential archives.

Austin principles of data citation: Current feedback

- Need to define audience for this document
  - Data producers AND data users
  - Other things to define (inclusively) > ‘data’, ‘metadata’, ‘citation’, ‘reference’ - what else?

- Citation practices: take into account current practices but also see what is required “out there”
  - Listen to creators’ guidelines for citing their data
  - Encourage/require reference elements required e.g. by DataCite
  - Encourage bi-directional referencing

Thanks to our sponsors

Travel support:

Helene N. Andreassen, UiT The Arctic University of Norway

Andrea Berez-Kroeker, National Science Foundation and University of Hawai‘i

Lauren Gawne, La Trobe University, Australian National Data Service (ANDS)