Issues in Data Attribution

Gary Holton
Alaska Native Language Archive
University of Hawai‘i at Mānoa
**Attribution** to refer to the practice of valuing data sets created by other researchers.

- For the most part discussions of data citation focus on the practice of referencing data collected by the researcher or by a member of her research team.
- To inspire reproducible linguistic research we must also facilitate citation of data collected by other researchers.
- Attribution of others’ research is key to the process, but in general data sets are not valued as highly as other forms of linguistic publications for the purposes of peer review.
- Thus, attribution is crucially about giving credit for the work of preparing, archiving, and making available linguistic data sets.
Reproducible research

- Attribution is key to inspiring reproducible linguistic research
  - preparing, archiving, and making available linguistic data sets involves a significant amount of work
  - if we don’t give people credit for this work then they are less likely to make that data available
  - in our current model data sets are not valued as highly as other forms of linguistic publications for the purposes of peer review.
Data has scholarly merit

• “shift in practice has broadened the range of scholarly work to include ... archives of primary data, electronic databases, corpora...” (LSA Resolution Recognizing the Scholarly Merit of Language Documentation, May 2011)

• “the Australian Research Council (ARC) made it clear that the ARC accepts that curated corpora can legitimately be seen as research output” (Margetts et al., to appear)
Existing solutions

● Proxy publications
  ○ create citable publication which describes the collection

● Co-authorship
  ○ include data providers as co-authors
Proxy publication

- **Common solution for software tools**
  - MrBayes (Ronquist and Huelsenbeck 2003)

- **Collection guides**
  - Published article providing overview of archival data, becomes citable object (Woodbury 2014)

- **May not sufficiently reflect the work put into creation of the archival data set**
Co-authorship

- A blunt instrument
  - doesn't distinguish between contributing data and actually collaborating on the analysis
  - not very fine-grained.
    - someone who collects a little data on lots of languages may be overly cited
  - doesn't scale well
    - collaborative projects will require numerous co-authors
Additional solutions

● **Review article**
  ○ peer review of archival data as if it were a publication
  ○ forthcoming in *Language Documentation & Conservation*

● **Data paper**
  ○ create venues for archival publication (Haskelmath & Michaelis 2014)

● **Peer review of archival data**
  ○ establish bodies to review deposits (Margetts et al., to appear)
Archiving as publication

• Problem: it is not only well-annotated data which needs to be recognized but also raw and less annotated primary data (cf. Margetts et al., to appear)

• Publication model might encourage cherry-picking of datasets, as linguists choose only the most richly-annotated or “prettiest” data for inclusion in the archival record

• Haspelmath & Michaelis 2014 explicitly emphasize selectivity
Athabascan (Dene) astronomy

- For more than 150 years the Gwich’in [gwi] term *yahdii* was reported in published literature as denoting the ‘Big Dipper’
- Yet the archival record reveals that this constellation is actually much larger
Athabascan (Dene) astronomy

• “Yahdii cover all the sky”

• Jules Jetté (1900) also hints at a similar whole-sky constellation in Koyukon, though this reference is not found in the later edition of the dictionary (Jones & Jetté 2000)

• [Ahtna field notes.] De Laguna and McClellan (1960) report similar constellation based on body-part terms in Ahtna (AT954DM1960, Fairbanks: ANLA)
Athabascan (Dene) astronomy

- Mega-constellation *Yahdii*, with body-part asterisms (Cannon 2014)
- Could not have been recovered without archival data
Peer review of archival data

- Requires development of new infrastructure
- Allows the establishment of metrics specifically relevant to archival data
  - quality of the repository
  - accessibility of the data (cf. Callaghan 2015)
  - format of the data
  - data richness / amount of annotation
  - diversity of content / representativeness
  - uniqueness of contribution
References


Woodbury, Anthony. 2014.