Open Data Kit (ODK) for mobile linguistic metadata entry

Richard T. Griscom
Leiden University
r.t.l.griscom@hum.leidenuniv.nl
@rtgriscom

Presentation category: Demonstration
Theme: "Replication, evaluation and quantitative analysis in the DH era"

Within the field of linguistics, high quality metadata are crucial for resource discovery (Good 2002), but also for answering research questions that involve extra-linguistic information (Kendall 2011). Linguistic fieldworkers collect and archive metadata as part of the language resources that they create, but they often work in resource-constrained environments that prevent them from using computers for data entry. In such situations, linguists must first create handwritten data in notebooks and then complete time-consuming digitization tasks that limit the quantity and quality of the resources and metadata that they produce (Thieberger & Berez 2012; Margetts & Margetts 2012). This demonstration will introduce a new method for entering linguistic metadata into mobile devices using the Open Data Kit (ODK) platform, a suite of open source tools designed for mobile data collection. This ODK method is the first linguistic metadata creation system that allows for independent teams of researchers to collect data simultaneously in remote areas and store the compiled results on the cloud. The method was developed as part of two community-based language documentation projects in Tanzania involving, twelve researchers collecting data in four administrative regions (Griscom & Harvey 2019). Through the identification of project-specific data dependencies and redundancies, a number of efficiencies were built into the metadata entry system. These include the use of closed vocabularies, unique data entry forms for distinct data collector categories, and separate forms for entering participant and resource metadata.

One of the first steps in developing a new tool or method is the identification of research values and desiderata (Good 2010). For the language documentation projects in Tanzania, the primary desiderata are metadata that satisfy the format and content requirements of the Endangered Languages Archive (ELAR), the repository in which project data will be deposited, and metadata that allow for the analysis of language variation and contact, a focus of the research program. Although many components in the ELAR metadata profile are not restricted to a closed set of possible values, within the context of a research project the value of many components is either static (e.g. target language, project) or restricted to a closed set (e.g. researcher, equipment used). A metadata creation system tailored to a specific project can therefore incorporate these static values and closed vocabularies to increase speed and accuracy.

The ODK system is not without its limitations. First, as with any metadata entry system, open text fields contain errors that must be checked either manually or through an automated system. Additionally, submissions for updated versions of forms must be manually compiled together with submissions for previous versions, which may contribute significantly to data processing, depending on the volume and timing of updates. Finally, individual data collectors are unable to easily view compiled data, and post-collection processing is required to produce metadata in the appropriate format for archiving. The piloted system offers a number of distinct advantages when compared to non-digital data entry methods, however, and promotes the creation of large and representative datasets, which constitute a primary goal of the language documentation imperative (Himmelmann 2006; Woodbury 2003).

References:

Good, Jeff. 2002. A Gentle Introduction to Metadata. http://www.language-archives.org/documents/gentle-intro.html. Accessed May 29th, 2020.

Good, Jeff. 2010. Valuing technology: Finding the linguist's place in a new technological universe. In *Language Documentation, Practice and values*. Amsterdam: John Benjamins.

Griscom, Richard T. & Andrew Harvey. 2019. Gorwaa, Hadza, and Ihanzu: Language contact, variation, and grammatical inquiries in the Tanzanian Rift. Presented at the East Africa Day Leiden, Leiden. https://doi.org/10.5281/zenodo.3509475.

Himmelmann, Nikolaus P. 2006. Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of Language Documentation*. Berlin: Mouton de Gruyter.

Kendall, Tyler. 2011. Corpora from a sociolinguistic perspective. *Revista Brasileira de Linguística Aplicada* 11(2). 361–389. https://doi.org/10.1590/S1984-63982011000200005

Margetts, Anna & Andrew Margetts. 2012. Audio and video recording techniques for linguistic research. In *The Oxford Handbook of Linguistic Fieldwork*. Oxford University Press.

Thieberger, Nicholas & Andrea L. Berez. 2012. Linguistic Data Management. In *The Oxford Handbook of Linguistic Fieldwork*. Oxford University Press.

Woodbury, Anthony C. 2003. Defining Documentary Linguistics. In *Language Documentation and Description*, vol. 1. London: Hans Rausing Endangered Languages Project. http://www.elpublishing.org/PID/006