Proceedings of the First workshop on Resources for African Indigenous Languages (RAIL), pages 31–35
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Mobilizing Metadata: Open Data Kit (ODK) for Language Resource Development in East Africa

Richard T. Griscom
Leiden University
Van Wijkplaats 4, 2311 BX Leiden, Netherlands
r.t.l.griscom@hum.leidenuniv.nl

Abstract

Linguistic fieldworkers collect and archive metadata as part of the language resources (LRs) 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 complete time-consuming and error-prone digitization tasks that limit the quantity and quality of the resources and metadata that they produce (Thieberger & Berez 2012; Margetts & Margetts 2012). This paper describes a 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. The method was incorporated into two community-based language documentation projects in Tanzania, involving twelve researchers simultaneously 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. The resulting system serves as the basis for the ongoing development of general purpose bilingual English-Swahili metadata entry tools, to be made available for use by other researchers working in East Africa.

Keywords: metadata, language resources, Africa

1. Introduction

Collecting linguistic data to support the creation of LRs for indigenous African languages often involves working with communities in areas where regular access to electricity and internet may be limited. These resource restrictions often lead data collectors to utilize paper-based methods that produce non-digital data which must then later be digitized. Digitization is time-consuming and can introduce additional errors to data, so a method that removes digitization from the workflow has distinct advantages (Thieberger & Berez 2012: 92; Margetts & Margetts 2012: 16). The methods and tools described in this paper enable data collectors, working individually or in teams, to create rich digital metadata in remote regions without the need for a computer or internet connection at the time of metadata creation.

1.1 Mobilizing Language Resource Metadata

High quality metadata are crucial for resource discovery (Good 2002), but also for answering research questions that involve extra-linguistic information (Kendall 2008; Kendall 2011). Various metadata standards exist for LRs, including Text Encoding Initiative (TEI), ISLE Meta Data Initiative (IMDI), and Component MetaData Infrastructure (CMDI), among others. There are also multiple linguistic metadata creation tools currently available, such as ProFormA2, Arbil, COMEDI, and CMDI-Maker (Fallucchi, Steffen & De Luca 2019). All metadata creation tools currently available require either a computer or a stable internet connection to function properly.

The need for a new digital metadata entry system that does not rely on a computer or stable internet connection is exacerbated when data collection is on a large scale and conducted by multiple researchers working simultaneously in different regions. These are the exact conditions of two coordinated and community-based language documentation projects in northern Tanzania, funded by the Endangered Languages Documentation Programme (ELDP): "Gorwaa, Hadza, and Ihanzu: Grammatical Inquiries in the Tanzanian Rift" (IFP0285) and "Documenting Hadza: language contact and variation" (IPF0304). Together, these two-year projects involve the participation of 10 local researchers from the Ihanzu and Hadza indigenous communities, distributed across five stations in the Lake Eyasi Basin, as well as two principle investigators (PIs). Figure 1 shows a map of Tanzania with the location of each of the five research stations marked by a dot (Google 2020a).

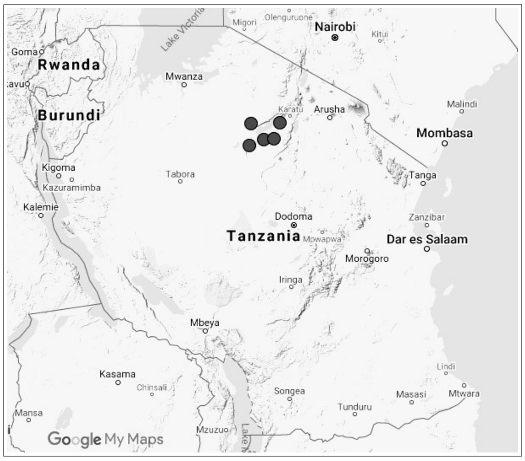


Figure 1: Map of research stations on a national scale

With each of the 12 researchers expected to produce new metadata every week for a period exceeding one calendar year, collecting and digitizing paper-based metadata was not a reasonable option. The majority of data collection for the two projects takes place in areas without electricity or internet, so typical digital metadata creation tools could not be used, either. A new method for mobile metadata entry was needed.

2. Open Data Kit

ODK is a free and open-source software platform for collecting and managing data in resource-constrained environments, and it includes three primary applications of relevance to linguistic metadata collection: ODK Build, a web application for creating custom forms for data entry based on the XForm standard, ODK Aggregate, a Java server application to store, analyze, and export form data, and ODK Collect, an Android application that allows for the entry of data directly into mobile devices. With all three of these components working together, teams of researchers can collect data quickly and simultaneously in remote areas, and all of their data can be compiled together on a single server. Figure 2 shows a schematic of the workflow during data collection: data collectors upload their data to an ODK Aggregate server from their mobile devices, and then a data reviewer compiles the data and exports it from the server for analysis.

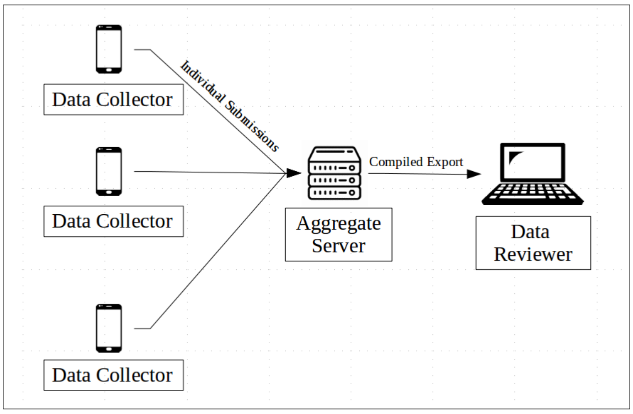


Figure 2: Schematic of the ODK data collection workflow

The following criteria were considered to be crucial for a successful metadata entry system: the removal of digitization from the metadata creation workflow, scalability so that the system could be used by teams of data collectors working independently and simultaneously, the utilization of mobile technology to enable metadata creation in areas without electricity, and an open source software platform that makes the method accessible to the researcher community.

The Open Data Kit (ODK) suite was selected as the primary platform for metadata collection because it satisfies all of the above criteria. It also has additional advantages, including an established record as a data collection platform among NGOs and non-profits working in Africa, support for multilingual data entry forms, and the collection of geo-spatial data.

3. The ODK Linguistic Metadata Method

The ODK metadata entry system created for the Hadza and Ihanzu community language documentation projects was designed and tested over a period of a few months prior to implementation. The method was designed to be as accurate and efficient as possible given the specific needs of the research projects, and later the specifics of the system were used as the basis for the development of general purpose tools.

3.1 Identifying Metadata Needs

A first step in developing a new tool or method is the identification of research values and desiderata (Good 2010). For the language documentation projects in Tanzania, our desiderata were 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.

The archive deposits for ELDP projects hosted on ELAR use a metadata profile that includes components for deposit, bundle, resource, and participant metadata (Duin et al. 2019). In total, three deposits were prepared for the two ELDP projects: one for Ihanzu and two for Hadza. A method was thus needed for specifying the appropriate deposit for each bundle. Bundles in ELAR are used to group together different types of resources and participants, so we also needed a method that would facilitate this grouping and correctly categorize resources and participants. The different types of resources include audio and video recordings, as well as text data such as transcriptions and translations. The two categories of participants include researchers and speakers.

The metadata required for studying language variation and contact include resource metadata such as speech genre, interactivity, and location of speech act, as well as participant metadata such as age, gender, education background, location and location history, and language background. Any data entry forms created for the Hadza and Ihanzu projects would therefore need to incorporate fields for entering these types of metadata, and the information would need to be processed in such a way that it can be easily retrieved.

3.2 Creation of Metadata Entry Forms

Once the desired metadata had been identified, metadata entry forms were created using ODK Build. A number of efficiencies were built into the metadata entry system through the identification of project-specific data dependencies and redundancies. These efficiencies included the use of closed vocabularies, unique sets of forms for different categories of data collectors, and the division of resource and participant forms.

3.2.1 Closed Vocabularies and Form Sets

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 constant (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 constants and closed vocabularies to increase speed and accuracy. Rather than create a single metadata entry form for all data collectors, which would include closed vocabulary sets with entries that were not relevant for some data collectors, we created three sets of forms: one set each for principle investigators, Hadza local researchers, and Ihanzu local researchers.

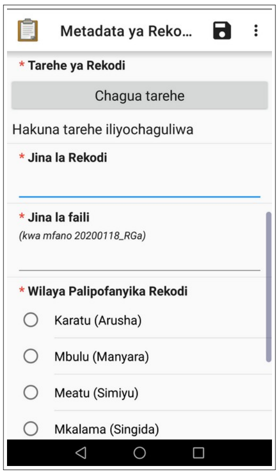
By creating three separate sets of metadata entry forms, we were able to restrict closed sets to only the values that were viable options for each category of data collector. This reduced the likelihood of categorical data entry errors and made the forms easier to navigate with fewer options to choose from.

3.2.2 Session and Speaker Metadata

Speakers often participate in the creation of multiple recordings, but participant metadata is only collected once. For this reason, two separate forms were created for resource metadata and participant metadata. This reduced data redundancy during the data collection stage, but also introduced the requirement for post-collection data processing (see Section 3.4).

3.2.3 Field types and organization

A variety of different entry widgets were integrated into the ODK forms, depending on the type of metadata to be collected. Open text widgets were used for metadata that aren not restricted to closed vocabularies, such as the names of participants and locations. Single choice widgets were used for metadata categories that constitute closed sets of mutually exclusive values, such as the gender of a participant or the name of the researcher collecting data. Multiple choice widgets were used for metadata categories consisting of closed sets of non-mutually exclusive values, such as the languages spoken by a participant. Date widgets were used for metadata such as recording date and participant birth year, a GPS widget was used to retrieve geo-spatial data for the location of recording, and a photo widget was used for creating photos of participants for identification purposes.


Metadata ya Reko…

* Tarehe ya Rekodi

Hakuna tarehe iliyochaguliwa

* Jina la Rekodi

* Jina la faili
(kwa mfano 20200118_RGa)

* Wilaya Palipofanyika Rekodi

○ Karatu (Arusha)
○ Mbulu (Manyara)
○ Meatu (Simiyu)
○ Mkalama (Singida)

Figure 3: A form for entering recording session metadata

Figure 3 shows a screenshot one of the ODK forms used by local researchers. The “Tarehe ya Rekodi” widget is a date widget used to choose the date of a recording, the “Jina la Rekodi” and “Jina la Faili” widgets are open text widgets for entering the names recordings and files, and the “Wilaya Palipoganyika Rekodi” widget is a single choice widget for choosing the district where a recording was made. The red asterisk by the name of each widget indicates that it is required, and users need to complete these widgets before continuing with the rest of the form.

3.3 Setting up the ODK System

Installation of the mobile device and server components of the ODK system was straightforward and did not require significant technical expertise. An ODK Aggregate server was prepared following the step-by-step directions on the ODK website (ODK 2017) FreeDNS was used to host the server free of cost, and Google Cloud Platform was used to create a virtual server (Google 2020b). Due to the low volume of data, a small virtual server and drive were deemed sufficient (g1-small virtual machine and 30 GB standard persistent disk).

ODK Collect was installed on five Android mobile phones, purchased locally in Tanzania. Each phone was given a unique username that identifies the team using the device to collect metadata. Access information for the ODK Aggregate server was stored in each phone's settings and the appropriate set of metadata entry forms for each phone was downloaded. An administrator password was put in place on each phone so that some options, such as deleting and downloading new forms, were made unavailable to local researchers.

ODK comes pre-installed with support for multiple interface languages. For the local researchers in Tanzania we set the interface to be displayed in Swahili, the lingua franca of East Africa. We additionally designed the forms to be visible in either English or Swahili, and set the default language of the forms as Swahili for the local researcher devices.

Figure 4: Local researchers practice using ODK (Photo credit Nadia Jassim)

During a five-day language documentation training workshop, all local researchers were given instruction on the use of ODK for metadata creation. Local researchers practiced entering data, saving forms, and uploading to the Aggregate server. During the training itself, some minor modifications were made to the forms, including the reordering and rewording of questions, based on feedback from the local researchers.

After the training, and once data collection had been initiated, a few additional modifications were made to the forms. For example, an additional question was added to the recording session metadata forms for the local researchers and PIs to create a unique filename for each recording, which includes the recording date in ISO format (YYYYMMDD), a two-letter code for the researcher who created the recording, and an alphabetic system for organizing the recordings based on the order in which they were created. These filenames are used for all of the resource files associated with a given recording. After local researchers experienced repeated difficulties with data management, it was determined that adding the filename question would make it easier for them to bundle resource files after data collection.

Follow-up visits to research stations provided opportunities to give continued feedback on metadata collection. Common issues included inconsistencies in the spelling of participant names and misunderstandings about meta-linguistic descriptions such as interactivity and speech genre,. The most frequent mistake initially was simply forgetting to enter metadata, either for a resource or a participant.

3.4 Metadata Processing

Through the method described here, metadata is entered into mobile devices and then uploaded to an ODK Aggregate server. The data from that server can then be exported as a comma-separated value (.CSV) file, or streamed live to Google Sheets. All of the submissions for each form can be exported together. In order to produce metadata files in the appropriate format for archiving with ELAR, data from the CSV files for participant and resource metadata forms need to be linked together.

A Python script was created to produce the final metadata files for each bundle to be deposited in ELAR. A bundle is a group of associated resources and participants. The script compiles information from the resource and participant metadata and identifies the types of resource files associated with each recording session to create a bundle that can be deposited in ELAR.

For example, if a given entry in the resource metadata specifies that two speaker-participants were involved in the creation of the recording, then the script uses the participant names in the resource metadata to extract additional metadata for those two speaker-participants from their corresponding entries in the participant metadata, and the script then creates a bundled metadata file using the extracted information.

4. Method Assessment

The competing goals of achieving representative data volume and data accessibility, collectively described by some as the "reproducibility crisis" (Gezelter 2015), can be addressed within the domain of linguistic fieldwork in at least two ways: the active participation of the speech community in data collection ("crowd-sourcing"), and the strategic use of computational technologies ("automation"). The ODK linguistic metadata method attempts to utilize both solutions, and has a number of notable advantages over non-digital entry methods.

The method has the potential to increase the quality, quantity, and consistency of linguistic data and metadata deposited in language archives. It does so not just by reducing processing bottlenecks, which enables linguists to spend more time analyzing or collecting data when they would otherwise be manually digitizing data, but also by opening the door to increased involvement of speech communities in the language documentation process, which has been shown to benefit research outputs by producing linguistic data sets that are more diverse and representative (Czaykowska-Higgins 2009).

The system is not without its limitations, however. As with any metadata entry system, open text fields will still contain errors that must be checked either manually or through an automated system of some kind. Additional training and feedback may reduce the error rate, but it is not reasonable to expect error-free metadata with any system that utilizes open text fields.

The submissions for updated versions of forms need to be manually compiled together with the submissions for previous versions, at least in the current version of the ODK Aggregate software. The significance of this task depends on the volume and timing of updates made to forms. If forms are submitted through ODK Collect using a previous version that has since been deleted from the device, then those forms can no longer be viewed locally on the device. Again, the significance of this limitation depends on the volume and timing of updates.

Perhaps the biggest limitation, however, is that the output of the ODK suite must be formatted according to the metadata profile of the corresponding language archive. This requires some coding and therefore restricts the pool of researchers capable of designing a project-specific implementation to those with coding knowledge or access to someone with that knowledge.

5. Towards a Standardized System

One way to decrease the learning curve for the ODK metadata system is to develop a set of standardized general purpose forms, based on one or more common metadata profiles, and an accompanying processing script. The Hadza and Ihanzu community language documentation projects in Tanzania are now serving as the foundation for the creation of such a set of tools. Initially, these tools will be based on the ELAR metadata profile and restricted to English and Swahili interfaces, which should be useful for researchers working with endangered language communities in East Africa. In the future, it is planned to expand the tools to include a French interface and metadata profiles for other common language archives and data repositories.

6. Conclusion

The piloted ODK linguistic metadata system offers a number of advantages when compared to manual data entry methods. The removal of digitization and the use of closed-vocabularies increase the accuracy and speed of metadata entry. This is significant because it allows for the creation of large and representative datasets, which are a primary goal of language documentation (Himmelmann 1998; Himmelmann 2006; Woodbury 2003). The scalability of the ODK system also allows teams of data collectors to work together, which can allow for increased community engagement and collaboration.

The system specifically developed for the Hadza and Ihanzu community language documentation projects relies on project-specific and repository-specific closed vocabularies and constant values, but these specificities inform the design of general purpose metadata entry tools. It is hoped that these tools will make the possibility of digital metadata creation a reality for researchers working throughout remote regions of Africa.

Acknowledgements

The research reported in this paper is supported by the Endangered Languages Documentation Programme (ELDP, IPF0304) and the Leiden University Centre for Digital Humanities (LUCDH).

7. Bibliographical References

Czaykowska-Higgins, Ewa. 2009. Research Models, Community Engagement, and Linguistic Fieldwork: Reflections on Working within Canadian Indigenous Communities. Language Documentation and Conservation 3(1). 15–50.

Duin, Patrick, Twan Goosen, Mitchell Seaton, Olha Shkaravska, George Georgovassilis & Jean-Charles Ferrieres. 2019. CMDI Component Registry. CLARIN. https://catalog.clarin.eu/ds/ComponentRegistry/#/.

Fallucchi, Francesca, Hennicke Steffen & Ernesto William De Luca. 2019. Creating CMDI-Profiles for Textbook Resources. In Emmanouel Garoufallou, Fabio Sartori, Rania Siatri & Marios Zervas (eds.), Metadata and Semantic Research, vol. 846, 302–314. Cham: Springer International Publishing. doi:10.1007/978-3-030-14401-2_28. http://link.springer.com/10.1007/978-3-030-14401-2_28 (13 February, 2020).

Gezelter, Daniel J. 2015. Open Source and Open Data Should Be Standard Practices. The journal of physical chemistry letters 6(7). 1168–1169.

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

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

Google. 2020. Google Cloud Platform. https://cloud.google.com.

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. 1998. Documentary and descriptive linguistics. Linguistics 36. 161–195.

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. 2008. On the History and Future of Sociolinguistic Data. Language and Linguistics Compass 2(2). 332–351. doi:10.1111/j.1749-818X.2008.00051.x.

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

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

ODK. 2017. Getting Started With ODK. https://docs.opendatakit.org/getting-started/#install-aggregate-optional.

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

Woodbury, Anthony C. 2003. Defining Documentary Linguistics. Language Documentation and Description, vol. 1. London: Hans Rausing Endangered Languages Project.