The Life of a File from Creation to CoRSAL:

An Interdisciplinary Approach to Managing Language Documentation Materials

A linguist cultivates relationships with dozens of people and entities throughout a language documentation project. Unfortunately, one of the most important relationships linguists incidentally neglect to develop is with the language repository that will facilitate the secure archival of their data. The preservation of language documentation materials is an essential goal for any language documentation project (Bird & Simons 2003, Himmelmann 2006). As such, a clear-cut plan for the journey each data file should take on its way to the archive must be made well before the project begins, and consistent communication with the repository is crucial to realizing that plan. In fact, the NSF has recently made it a requirement for NSF funded projects to outline data management and archiving plans, underscoring its importance.

And yet, despite these requirements and the literature advocating for pre-fieldwork planning, workflow processes, and metadata standards (Johnson 2005, Sullivant 2020, Thieberger & Berez 2011), problems invariably arise when linguists deliver files to the digital archive (e.g., project file naming conventions, metadata standards, or the encoding of genre types may not conform to the specific depository's guidelines).

We suggest that the crux of data mismanagement often lies in a failure to utilize relationships with a repository in order to create a unified data management strategy. Because repositories are well acquainted with the management/archival of digital resources, prioritizing this relationship helps linguists to avoid several challenges involved in managing their data. This paper addresses common data management issues that language documenters experience, asserting that a relationship with the repository is the cornerstone to successful data management plans. We address several details that linguists must work out in concert with a repository in order to avoid these issues, including data preparation guidelines, best-practices for data storage and editing, and ways to ensure digital language collections will be findable, accessible, and protected. Ultimately, we suggest a workflow that leverages the expertise of the repository every step of the way.

We use the NSF funded project documenting Mankiyali, a hitherto undocumented language of Pakistan, as a test case for effective approaches to data management. This project has, since its inception, worked closely with CoRSAL to formulate data management plans, and we draw upon this relationship to discuss strategies for choosing a repository and collaborating with that repository to prepare data for archival.

References

Bird, Steven & Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79(3). 57–582.

Computational Resource for South Asian Languages. September 2020. Retrieved from https://corsal.unt.edu/.

Henke, Ryan & Andrea Berez. 2016. A Brief History of Archiving in Language Documentation, with an Annotated Bibliography. *Language Documentation & Conservation* 10. 411–457.

Himmelmann, Nikolaus. 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* (Trends in Linguistics Studies and Monographs 178), 1–30. Berlin: Mouton de Gruyter.

Johnson, Heidi. 2005. Corpus Management 101: Creating archive-ready language documentation. Presented at The Linguistics Association of the Southwest 2005, Lubbock, TX.

Sullivant, Ryan. 2020. Archival description for language documentation collections. *Language Documentation & Conservation* 14. 520-578.

Thieberger, Nicholas & Andrea L. Berez. 2011. Linguistic Data Management. In N. Thieberger (ed.), *The Oxford Handbook of Linguistic Fieldwork*, 90-118. Oxford: Oxford University Press.