Community-friendly portals to language data

Jonathan D. Amith
Research Fellow, Department of Anthropology, Gettysburg College
Research Associate, Department of Botany, National Museum of Natural History, Smithsonian Institution
jonamith@gmail.com

Abstract

The proposed development of community-friendly portals responds to the realization that "preservation and access," the cornerstones of language and cultural documentation, are complex and often somewhat contrary goals. Whereas *preservation* prioritizes long-term security of deposited resources, *access* is a more nuanced term, whose meaning may vary according to the type of resource in question and the particular needs of distinct end users. At a basic level, "open-access" developed out of a scholarly initiative to make the data and results of research freely available, while at the same time imposing some control (e.g., via Creative Commons licenses) over credit and use. Over time, however, it became apparent that both the resources that emerge from documentation projects and the identity and interests of potential end users are quite divergent, requiring different approaches to resource and metadata management and dissemination. For example, a focus on primary data resources led to the development of the FAIR (findable, accessible, interoperable, reusable) standard, whereas a concern with native rights to knowledge led to the development of TK (Traditional Knowledge) labels as well as CARE (Collective Benefit, Authority to Control, Responsibility, Ethics) principles. A distinct model is provided by collective *open source* efforts such as those represented by GitHub and other collaborative repositories such as iNaturalist that promote citizen science efforts.

The discussion that follows builds upon the preceding reflections, along with personal experience and collaboration across a range of disciplines, each with their own "culture" of resource development and distribution. Preservation of physical objects (e.g., material culture, biological specimens) is more in the realm of museum curation, whereas literary and audiovisual resources are more in the realm of archives and libraries. But even among the latter, there is a disjunction: essays, time-coded transcriptions, databases, and dictionaries may all be "textual" at a basic level. And they may be archived and preserved as such. But providing meaningful, operational access to each involves distinct mechanisms. Essays may simply be made available in pdf or text formats. Time-coded transcriptions involve a two-part delivery: (1) audio (usually wav or mp3) and (2) time-coded annotations (e.g., .eaf or .trs files). It is then up to the end user to put the parts together, inevitably through open-source software such as ELAN. In addition to the basic challenge this might create, it is inefficient: each end-user must repeat this process for each individual resource. Dictionaries present the problem of discovery and retrieval, not a trivial task. Continual improvements in information retrieval (e.g., using finite state transducers to facilitate searches across linguistic and orthographic variation in native languages; using word vectors in national languages to facilitate look-up in minority or endangered languages) provide one innovative approach. Other resources (ethnobiological information) and potential end users (educational institutions, biologists, communities) present distinct challenges to data management and distribution, including continual changes in metadata.

In sum, the amalgamation of metadata standards, resources, stakeholders, and needs requires a nuanced approach to language and cultural documentation, an approach that should perhaps focus to a great extent on operational access by native speakers and communities to the materials that emerge from collaboration with non-native academic researchers.