OpenSpeaks: Development of open resources for citizen multimedia archivists

Subhashish Panigrahi*

As many indigenous and other low-resource language speakers around the world grapple with the horror of the COVID-19 pandemic, the imminent danger of their languages going extinct is strongly felt. (Langlois 2020) While working closely with several language digital activists in Asia, particularly during the surge of the pandemic last year, we came across some of the practical challenges with audio-visual documentation of indigenous and endangered languages which helped create a framework keeping in mind the citizen-archivists. This pandemic was a caution to build both open resources and capacity for community-led and decentralized language protection initiatives.

Based on a bilingual survey (English and Santali), anecdotal narratives of several indigenous friends and colleagues, and past experience of working closely for media development in indigenous languages, we added a simple English chapter to OpenSpeaks, our openly-licensed toolkit for multimedia documentation of languages. The topic for this chapter was "Consent, Content Rights and Content Licensing" as most citizen-archivists do not have a clear sense of these critical topics that are tied to the future of the archival.

OpenSpeaks is hosted on Wikiversity (a Wikipedia sister project). It is not only open to all for contributions, but is also used by archivists. We were honoured to have three Santali-language speakers (Santali is one of the indigenous languages from India) to lead the localisation of the Chapter into Santali. This was arguably the first time these concepts were created in Santali as our friends R Ashwani Banjan Murmu, Fagu Baskey and Joy sagar Murmu juggled to define and transliterate terms like "consent", "copyright" and "open licenses" in Santali.

When COVID-19 aggravates the issues with healthcare, information access, livelihoods among others, the critical information related to the preventive ways being available in native languages is vital. (UN 2021) Initiatives such as virALLanguages are ensuring that many local communities are being able to create video resources in native languages so that they can spread the videos over social media and other popular channels. (Panigrahi 2020)

As we digged deeper to understand more about the issues that many activists face for general audiovisual documentation through a survey, we learnt that lack of financial and technical resources top the chart. But most importantly, many activists do not also have the knowhow about vital topics such as copyright, licensing and acquiring consent while documenting in their languages. As these topics are the foundation to the future of the archived materials, we focused on expanding a full-fledged chapter inside OpenSpeaks. Keeping in mind a new or intermediate multimedia archivist whose first language is not English, we tried to frame in simple terms the modality of seeking consent based in a complex and diverse environment. Figuring out the license for proper attribution and ensuring that the archived audio and videos are available to the native speaker community openly are also hard. So, we used a "community-first" approach while explaining the related legal and technical subjects. But to ensure that the content is relevant to an indigenous community, we invited some of the noted leaders from the Santali community who are active in the open knowledge movement so that they can localize and confirm that the chapter is not only relevant to their own community but is also placed well in the cultural context. They created new words based on vocabulary that is otherwise known to the community, and transliterated words that are widely popular (e.g. license) or brand names (e.g. Creative Commons). As the Santal diaspora is spread across eastern India and partly in neighbouring Bangladesh, Nepal and Bhutan, we were fortunate to have one out of the three contributors from the Indian state of Assam. All of the three contributors ensured that Santals across borders can relate to this Chapter.

Noted linguists Mandana Seyfeddinipur and Felix Rau among others have underlined the factor of the constantly evolving technical specifications for documentary linguistics (Seyfeddinipur 2020) and such professional guidelines help strengthen the citizen archival strategies that we aimed to create. We focused on the social and cultural factors in addition to the legal and technical aspects of the topics covered. Our hope is that educational resources like this toolkit is of help to communities as they continue to document in their languages.

Going forward, we will be releasing the parallel corpus (English-Santali) under a Creative Commons Zero License (CC0 1.0 -- equivalent to Public Domain) so that it adds to slowly evolving community efforts for machine translation. We are also constantly challenged with disability and our next goal is to expand a full-fledged chapter on accessibility in the context of citizen language documentation.

---

*Author bio: Subhashish Panigrahi is the founder of OpenSpeaks. A documentary filmmaker and a National Geographic Explorer, he has been involved in development of frameworks for Open Educational Resources (OER) for indigenous languages and audiovisual documentation of low-resource languages. Panigrahi contributed to the Linguapax MOOC "Language Diversity: What For?"