openspeaks voice—odia

Subhashish Panigrahi
Musamoni Panigrahi

OPENSPEAKS VOICE ODIA

SUBHASHISH PANIGRAHI

&

MUSAMANI PANIGRAHI

FIRST EDITION

https://theofdn.org/activities/before/

This report, published in September 2023, details an Odia-language speech data repository founded by Subhashish Panigrahi in 2018. All the recordings under the repository are available under a CC0 1.0 Public Domain Release. To view a copy of this license, visit https://creativecommons.org/publicdomain/zero/1.0/.

Correspondence
Subhashish Panigrahi {subhashish@theofdn.org}
ORCID: 0000-0002-3665-8184

ISBN: 978-1-7386808-2-5 (E-Book Version)

Typeset and design by Subhashish Panigrahi. Typefaces (Cormorant Garamond, Sans, Poppins, Courier New) under Open Font License (OFL).
F/LOSS software used: Scribus (typeset), Inkscape (vector design), GIMP (image editing).

First Edition: September 2023

CONTENTS

FOREWORD

HUGO LOPEZ
Wikimedian and Lingua Libre Co-Founder

In 2007, a young French software engineer traveled to Ukraine. While there, struggling to learn the local language properly, Nicolas prototyped a rapid audio recorder, soon known as "Shtooka". This open-source tool was so blatantly efficient in recording thousands of words that it progressively became popular among technophile foreign language enthusiasts and language teachers.

Acknowledging its power to document languages efficiently, Wikimedia France took over the project and its development. An online service, not requiring burdensome installation and therefore much easier to use by the general public, came online at LinguaLibre.org. With this alliance, community, and powerful online service came a new vision; beyond major language learning, we now can push to audio document all human languages.

With a collection of over 61,000 recordings, the Odia and Baleswari-Odia projects led by the O Foundation and Mr. Subhashish Panigrahi represent news milestones. Odia is the second non-Western lexicon, the first non-national language, and the first Indian language to be methodically audio-documented with such quality and breath. Published under open license, the CC0 1.0 Universal (CC0 1.0) Public Domain Dedication, the Odia collection can be freely reused to provide new services, from language learning to machine learning-based services. This pioneering work announces future efforts, with "regional" oralities of India and the world rising back to claim their place in the digital landscape. Today, the Wikimedia and Lingua Libre communities are looking with admiration toward the O Foundation and our fellow Odia contributors.

As Nicolas said 15 years ago, the O foundation now lights the way; may we be numerous to follow your path.

INTRODUCTION

Projects like Mozilla Common Voice were created to address challenges such as the unavailability of voice data or the high cost of available data for use in speech technology, such as automatic speech recognition (ASR) research and application development. The pilot detailed in here aims to create a large, freely licensed public repository of transcribed speech in the Odia language, as no such repository was known to be available. The strategy and methodology behind this process are based on the OpenSpeaks project. Licensed under a Public Domain Dedication (CC0 1.0), the repository currently includes audio recordings of pronunciations for more than 61445 unique words in Odia, including over 5,600 recordings of words in the northern Odia dialect Baleswari. The author did not find any known public listing of words in this dialect before this pilot. This repository is arguably the most extensive transcribed speech corpus in Odia, and is publicly available under any free and open license. Detailed here the strategy, approach, and process behind building both the text and speech corpus using many open source tools, such as Lingua Libre, which can be helpful in building text and speech data for different low- to medium-resource languages.

ODIA SPEECH DATA

Introduction

The Mozilla Common Voice project establishes an important standard for creating transcribed speech corpora in an open and decentralized way. The project responds to the issue of speech data's high cost or unavailability hindering speech technology-related research and development [9]. Despite the active use of audiovisual media in the Odia language, there is no openly-licensed speech data repository beyond Common Voice. This pilot strategy and methodology, based on the OpenSpeaks Project, aim to address the limited availability of voice data in most low and medium-resource languages and detail the process for building such data for speech research and development. The initiative focuses on Odia, the native language of the author, spoken by between 35-45 million people in the Indian state of Odisha, which has seen minimal research and functional application creation in phonology and natural language processing (NLP) like speech synthesis or Automatic Speech Recognition (ASR) due to the lack of publicly available voice data and diversity in data [11].

The need for the availability of openly-licensed word corpus and speech data has been felt in the NLP and ASR community [11]. Recording a vast range of natural speech data is always recommended for building the foundation of voice data. Multilingual speech technologies are most needed in humanitarian areas. Their availability is critical for technological innovations, such as building applications using text-to-speech and speech-to-text engines for emergency announcements, interactive voice response (IVR) for phone-based services,

A version of this section was presented at the Companion Proceedings of the Web Conference 2022. More at https://doi.org/10.1145/3487553.3524931.

creating assistive web technologies for people with disabilities, and ensuring speech-based accessibility at public service kiosks in physical spaces.

As described here, a speech corpus of more than 61,445 Odia-language words was created using different open-source tools, particularly Lingua Libre, a web platform for recording pronunciations. These audio files were uploaded under CC0 1.0 Universal Public Domain Dedication, making them perpetually free for NLP research and development. The "Methodology" section details the process of collecting words, compiling a wordlist, recording the pronunciation of those words, and uploading the speech data to Wikimedia Commons using Lingua Libre. The workflow was tested for Odia and Baleswari Odia, a dialect that still does not have a public listing of lemmas. 5,600 Baleswari Odia words were recorded, and the learning and recommendations for similar initiatives in different low-resource languages were published as a framework.

Although initially planned as a voice-only project, this project also benefited the creation of wordlists and Wikidata lexemes, which can be further related to NLP work, as explained in the "Result" section. Apart from elaborating on the technical strategy, also discussed here the integration of lexemes for different low-resource languages focusing on voice data.

BACKGROUND AND RELATED WORK

Odia language

The South Asian languageOdia (formerly "Oriya") originated in present-day eastern Indian state of Odisha (formerly "Orissa"). As per the 2011 Indian census [1], Odia is spoken by about 35 million people, and Ethnologue mentions another 4 million speakers who speak it as their second language (L2) [7]. Odia is the most spoken macro language (81.32% of the population) in Odisha, where the neighboring languages include over 21 other minorized languages [23], which are clubbed as the Adivasi languages. Adivasi is a heterogeneous socioeconomic group comprising many indigenous peoples. Sixty-two such groups live in Odisha [23]. Their languages have immensely influenced Odia and are also influenced by Odia itself. Odisha also shares borders with the neighboring states such as West Bengal and Jharkhand to the north, Andhra Pradesh to the south, Telangana to the southwest, and Chhattisgarh to the west [28]. Angika, Bajjika, Bengali, Bhatri, Hindi, Ho, Kui, Sadri, Santali, Urdu, and Telugu are some of the other neighboring languages whose speakers live in the aforementioned neighboring states and inside Odisha [23].

Baleswari Odia (Balesoria)

Baleswari Odia (an exonym, the endonym is often shortened as "Balesoria"), which is also used as an eponym for the people from the larger Balasore area) is the northern dialect of the Odia language. It is spoken primarily in the Balasore (Baleshwar) district, an administrative region in India, of the Indian state of Odisha, the neighboring districts of Mayurbhanj and Bhadrak, and the bordering regions of West Bengal state such as the Purba Medinipur district [13].

Voice data in Odia and Baleswari Odia

Recording and transcribing voice data have been limited for Odia. Much less data is published under public and open licenses [15]. Creating audio descriptions to accompany textual data has also led to creation of some speech data. For instance, dictionary entries for the Odia Wiktionary project sometimes carry the pronunciations of words. Similarly, adding the native pronunciation of some of the Wikipedia article titles, such as biographical articles or articles with proper nouns, is helpful for the readers as audio helps avoid ambiguity in such cases. Such needs have seen responses from the volunteer contributor community contributing in recording audio. Many articles on biographies of living personalities also occasionally contain their recorded voice. The Odia Wikipedia has existed since 2003 and over 16,000 articles. There have been sporadic contributions to enhance such articles. For instance, 464 audio recordings containing text from Wikipedia articles on medicine were uploaded by some Odia Wikipedia editors [29]. Similarly, several Odia Wikipedia editors made 50 recordings of notable personalities through "Project Parichay" [22] and under the Voice Intro project [31]. Panda and Nayak

[18] have performed an analysis by contrasting their text-to-speech system in Odia with the free software Dhvani TTS [14] that makes use of compressed audio and concatenation to produce "intelligible speech". As documented on OpenSpeaks, experimental open-source projects such as Lekatha also demonstrate natural voice data for speech synthesis. [19] Mozilla's Common Voice project emphasizes the use of high-quality and openly-licensed (Public Domain) natural voice recordings as a recommended practice for the future of speech technology, considering the fact that natural speech that inherently carries intonations and accents can be helpful in speech recognition [9]. A dataset containing more than nine hours of the audio recording of sentences in Odia was uploaded by volunteers to Common Voice by the end of 2021. Over one hour of this recording is validated by a community of contributors [12]. The 55,000 recordings elaborated on here add nearly 22 hours of voice data to such existing openly-licensed data [20].

Platforms and applications used

The OpenSpeaks project provides a wide range of resources, such as Open Educational Resources (OER) and templates that can be used in different linguistic and demographic environments to conduct audiovisual documentation of languages, particularly those with low and medium resources [9]. It is hosted on the English Wikiversity (https://en.wikiversity.org/wiki/OpenSpeaks). Lingua Libre (https://lingualibre.org/) is a browser-based online platform that allows users to record pronunciations of words in any language or dialect that can be written using a writing system with Unicode encoding. Meanwhile, Common Voice (https://commonvoice.mozilla.org/en) is a web platform by Mozilla that encourages contributors to submit sentences available under a Public Domain release in supported languages, record pronunciations of the sentences, and review recordings made by others. The recordings are made available for download under a Public Domain license. Unlike Lingua Libre, Common Voice anonymizes the recordings by stripping the usernames [9]. Moreover, the Python script created by T. Shrinivasan [27] for generating a list of unique words from any text file (with .txt, .xml, or .json extensions) can be modified to fit the needs of various languages with different writing systems.

RESULTS

This section provides an overview of the outcomes of the process outlined in the "Methodology" section. I compiled a wordlist comprising more than 530,000 distinct words, including over 200,000 lemmas and their corresponding forms. Since the words were sourced from different locations, some forms without related lemmas also exist. Between June 2018 and March 2022, I recorded and uploaded pronunciations for over 61,445 unique words. The current workflow is based on a functional model established by December 2021, with nearly 21,000 words in the list being modern-day words, including several loanwords. Out of the 55,000 words, 5,605 belong to the Baleswari Odia dialect.

The lexemes in Odia currently being created on Wikidata are expected to aid in developing forms for numerous existing lemmas using the created wordlist. Relevant pronunciations from the voice data repository can accompany lemmas and forms. Wikidata also provides options for incorporating multilingual translations to lexemes to develop a parallel corpus. Although the present text and voice data can be employed for spell-checking, dictionary with word pronunciations, and Automatic Speech Recognition (ASR) related work, the development of a parallel corpus can lead to further work on machine translation. A list of all the recorded Odia words, including the Baleswari Odia words, and a separate list containing only the recorded Baleswari Odia words are available along with this paper.

OBSERVATIONS

The current wordlist presents some primary issues and features, which include:

•Several words in the wordlist contain pre-existing spelling errors as many Wikipedia articles (or other source content) are not always edited for spelling. Currently, Odia lacks a spell-checker with a comprehensive list of all words to detect and correct such mistakes. Although some errors are detected and corrected during recording, the remaining words require cleaning up.

•Many spelling errors in the wordlist were also caused by using script encoding converters (legacy to Unicode encoding standards) or input tools that were designed to function for typefaces of legacy encoding when creating Wikipedia articles. Manual searching and replacing of such mistyped words or their removal from the entire document helped clean the list. However, as previously mentioned, this does not ensure complete accuracy, and some spelling mistakes persist.

•In this wordlist, Odia words are also appended with characters from other writing systems.

•Certain words contain insertions of Zero-width joiner (ZWJ) and Zero-width non-joiner (ZWNJ) characters that appear visually identical but are considered distinct words when sorting. Both ZWNJ and ZWJ are employed in Odia to prevent a character from joining with the preceding character when the

"halanta" (or "virama"; 0B4D in Unicode block [4]) sign is suffixed to the latter. These entries result in the same words being recorded twice when entered into Lingua Libre.

•In terms of the depth of topics covered in this wordlist, Odia Wikipedia articles cover a broad range of subjects, but the content depth is not always uniform [16]. For instance, articles related to specific topics are adequately covered, whereas other areas may be insufficiently covered. Such instances are primarily due to the lack of a sizable, active, and diverse editor community that encompasses a balanced representation of people from different age groups, genders, and educational/work backgrounds.

•The wordlist contains a considerable proportion of loanwords, including a long list of medicine-related terms primarily transliterated in the Odia alphabet.

Issues and Features Found in the Recording Process

•Lingua Libre allows a user to re-record a word, even if they had previously recorded the same word using a different platform or process. While this is a known issue, it is also useful for ASR training, as it allows users to record multiple variations of pronunciations of the same word.

•However, Lingua Libre also allows a user to record the same word multiple times in a row, which rewrites the file on Wikimedia Commons. This is not useful for recording intonations and accents spoken by the same user. To avoid this issue, users can create a separate profile on Lingua Libre.

•While one of the demographic fields on Lingua Libre is "Place of residence," from a linguistic standpoint, "Place of language learning" could be more relevant. "Place of residence" may not have any impact if a user lives in a location with a different language environment.

•Wikimedia Commons does not currently detect characters from a range of writing systems, including Chakma, Sharada, Grantha, Warang Citi, among others. This was identified during a workshop I conducted for the Ho language of India. [10] While this issue does not affect the Odia alphabet, it does create a barrier for contributors of other languages.

RECOMMENDATIONS

This section includes a set of broad recommendations that require further consideration when framing strategies and plans for developing speech corpora in different low-medium resource languages and dialects. While these recommendations are not definitive or exhaustive, they can serve as a guide or template to help create a workflow for low-limited-resource languages. Native speakers can create a more specific strategy relevant to their constraints and advantages.

1.The core layer of any speech corpus development is creating a wordlist of unique words. Collecting and compiling words in a target language or dialect is paramount. The lack of such wordlists could hinder NLP research and development. It is highly recommended that the creator of such data publicly release their corpus under open licenses as long as they can compensate for the labor, keeping in mind that volunteer labor is not universal in all parts of the world.

2.Wordlists created from a monolithic source such as literary works of a single author or literature of a similar genre might lack phonetic diversity. Hence, growing the wordlist gradually and incorporating contemporary words (including known loanwords) is highly recommended. Contemporary vocabulary could also ease any entry-level barrier of everyday users who might not have the fluency that users familiar with the historical use of language through the study of ancient literature might have.

3.Social media can be a great starting point for community engagement, and such interactions, when done in the target language, can help generate more content. This content can, in turn, be used to generate a wordlist for future speech projects.

4.Speech corpus development is a slow and gradual process. Hence, incremental and iterative growth in data quality is a known feature of speech data development. Contributors often need to have a long-term strategy for such a project to address the issues of not having a contributor community with diversity in gender, geographical, and socioeconomic background.

5.While it is essential to have the diversity of speech while building voice data, the initial process might be driven by a handful of people, and the initial data would lack diversity. Setting up an initial workflow can follow a contributor community-building exercise.

6.The socioeconomic mobility of individual contributors (including systemic oppression of certain marginalized groups such as the Dalits and Adivasis in India) and the resulting inequitable access to technical know-how can hinder the NLP research in many languages. It is vital to put early effort into generating financial compensation for paid labor (as opposed to expecting free labor from those who lack affordability) while developing speech corpus.

7.It is recommended to evaluate the dialects and accents of a target language during speech data creation. These linguistic nuances could help identify individuals within the speaker community and widen the speech data diversity.

CONCLUSION

Elucidated here an end-to-end process of creating speech data for both Odia, a formal and standardized form of a language, and Baleswari Odia, a dialect containing words with different intonations and accents. The process detailed here relies primarily on Lingua Libre, an open-source web platform that allows recording pronunciation of words, phrases, and sentences in a language or dialect and is integrated into the Wikimedia projects. Over 21 hours of voice data of a male contributor is published under a Creative Commons CC0 1.0 Universal Public Domain Dedication. The wordlist used for the pronunciation recording is also published under Public Domain. This license makes the data perpetually free of all copyright restrictions for Natural Language Processing (NLP) research, such as Automatic Speech Recognition, speech-to-text, and other possible applications.

The contributions from only a male speaker with higher social and educational access, introducing monotony into this speech data, is a known issue. However, the process and the strategy shared here can be replicated for different low-medium resource languages. There is vast potential to grow the database by incorporating data from diverse speakers of different genders, socioeconomic mobility, and educational access. The strategy and process summarized in the "Recommendations" section can also be replicated in other languages and dialects with a writing system with Unicode encoding.

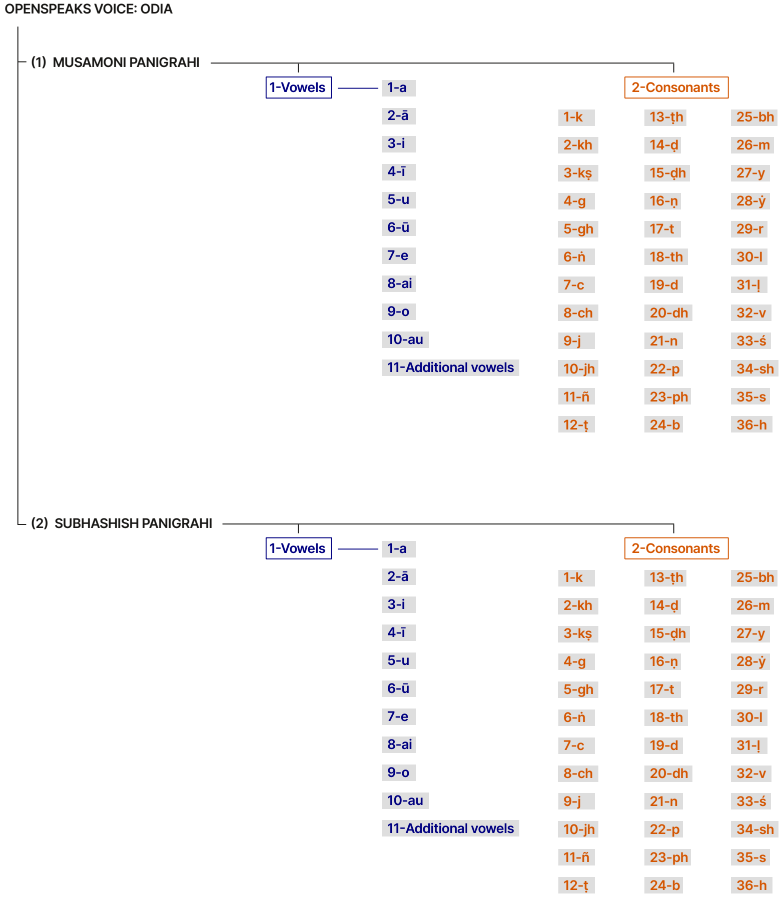
FOLDERS & AUDIO FILES

OpenSpeaks Voice: Odia is an ongoing project with periodic releases. OpenSpeaks Voice: Odia Volume I and II, and Balesoria-Odia Volume I and II were released during February–March 2023 as DVD-ROMs. This report focuses on the overall workstream and outcomes of those releases, containing 61,449 audio recordings in Waveform Audio File Format (WAVE). Later, additional recordings were added, and the repository grew into over 72,000 recordings. At the time of reporting, this is the largest speech data repository in the Odia language under a Public Domain release, with over 25 hours of recording.

The repository is divided into two main dialects of Odia—Balesoria (Northern) and Central—and contains over 72,000 audio recordings by the time of reporting. Of these, nearly 65,000 are in Central Odia, and the remaining 7,000 are in Balesoria. All these recordings were made by Panigrahi, in his own voice and 600 recordings, in Balesoria, are in the voice of the Musamoni Panigrahi, taken from the 2022 documentary Nani Ma. This repository is a part of the OpenSpeaks project, founded by Subhashish Panigrahi in 2017. A web-based open-source tool called Lingua Libre was used for most of his recordings, whereas Audacity was used for extracting audio from the recordings of the late Musamoni Panigrahi.

The main directory of this repository presently has subfolders for each speaker, namely "Musamoni Panigrahi" and "Subhashish Panigrahi". Each speaker subfolder has subfolders under it for vowels, consonants and numerals. Note that the titles for the lowest level of subfolders include romanized/transliterated following the ALA-LC (American Library Association – Library of Congress) standard for Odia/Oriya (2013), whereas the filenames are in Hex NCR, with a slight modification, where "&#" (ampersand followed by a pound sign) is replaced by "("(parenthesis), and ";" by")". Three open-source text replacement converter was built by Subhashish Panigrahi based on input from the US Library of Congress to handle the filename conversation, all available online at https://or.wikipedia.org/s/1qwd.

File Structure
(September 2023 release)



REFERENCES

[1] 2011. Abstract of Speakers' Strength of Languages and Mother Tongues - 2011. Technical Report. Registrar General and Census Commissioner of India, New Delhi. 6 pages. https://www.censusindia.gov.in/2011Census/Language-2011/Statement-1.pdf

[2] 2021. Help:RecordWizard Manual. https://lingualibre.org/wiki/Help:RecordWizard_manual

[3] 2021. Tutoriel: Comment contribuer à Lingua Libre? https://commons.wikimedia.org/wiki/File:Tutoriel_Lingua_Libre.webm

[4] 2021. The Unicode Standard, Version 14.0: Oriya. Technical Report. Unicode, Inc. 4 pages. https://www.unicode.org/charts/PDF/U0B00.pdf

[5] 2022. Bigyan Diganta. http://odishabigyanacademy.in/bigyan-diganta/

[6] 2022. Index of /datasets: Q322719-mis-Baleswari Oriya.zip. https://lingualibre.org/datasets/Q322719-mis-Baleswari%20Oriya.zip

[7] 2022. Odia: Ethnologue. https://www.ethnologue.com/language/ory

[8] 2022. orwiki dump progress on 20220301. https://dumps.wikimedia.org/orwiki/20220301/

[9] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. arXiv:1912.06670 [cs] (March 2020). http://arxiv.org/abs/1912.06670 arXiv: 1912.06670.

[10] Biswajeet3. 2021. T297351 Warang Citi (Ho-language writing system) characters not detected on Wikimedia Commons. https://phabricator.wikimedia.org/T297351

[11] Britone Mwasaru. 2022. Why Voice is Important. https://foundation.mozilla.org/en/blog/why-voice-is-important/ Section: Common Voice.

[12] Common Voice Contributors. 2022. Common Voice by Mozilla. https://commonvoice.mozilla.org/

[13] Artatrana Gochhayat. 2016. Odisha as a multicultural state: from multiculturalism to politics of sub-regionalism. Afro Asian Journal of Social Sciences 7, 2 (2016), 28. http://mail.onlineresearchjournals.com/aajoss/art/197.pdf

[14] Ramesh Hariharan, Ravi Masalthi, Rileen Sinha, and Santhosh Thottingal. 2021. Dhvani TTS. https://github.com/dhvani-tts/dhvani-tts original-date: 2013-12-

17T05:16:31Z.

[15] Josh Meyer. 2022. Open Speech Corpora. https://github.com/coqui-ai/openspeech-corpora original-date: 2019-01-31T14:57:39Z.

[16] Marc Miquel-Ribé and David Laniado. 2018. Wikipedia Culture Gap: Quantifying Content Imbalances Across 40 Language Editions. Frontiers in Physics 6 (2018), 54. https://doi.org/10.3389/fphy.2018.00054 12 citations (Semantic Scholar/DOI) [2022-03-08] Publisher: Frontiers.

[17] Mahir Morshed. 2022. tool-twofivesixlex. https://phabricator.wikimedia.org/source/tool-twofivesixlex/

[18] Soumya Priyadarsini Panda and Ajit Kumar Nayak. 2015. An efficient model for text-to-speech synthesis in Indian languages. International Journal of Speech Technology 18, 3 (Sept. 2015), 305–315. https://doi.org/10.1007/s10772-015-9271-y 13 citations (Semantic Scholar/DOI) [2022-03-08].

[19] Subhashish Panigrahi. 2015. OpenSpeaks/toolkit/Lekatha - Wikimedia Commons. https://commons.wikimedia.org/w/index.php?title=OpenSpeaks/toolkit/Lekatha&oldid=246217807

[20] Subhashish Panigrahi. 2022. Lingua Libre pronunciation by Psubhashish. https://petscan.wmflabs.org/?psid=19878687

[21] Subhashish Panigrahi, Mahir Morshed, and Lucas Werkmeister. 2022. Wikidata Lexeme Forms/Odia. https://www.wikidata.org/wiki/Wikidata:Wikidata_Lexeme_Forms/Odia

[22] Project Parichay Contributors. 2022. Project Parichay. https://commons.wikimedia.org/wiki/Category:Audio_files_uploaded_under_project_Parichaya

[23] Minati Singha. 2021. Odisha to impart primary education in 21 tribal languages in schools run by SC/ST dept. The Times of India (Sept. 2021). https://timesofindia.indiatimes.com/city/bhubaneswar/state-to-impart-primary-eduin-21-tribal-languages-in-schools-run-by-sc/st-dept/articleshow/86463363.cms

[24] Subhashish Panigrahi. 2021. Before AI*. https://github.com/ofdn/Before-AI/blob/91b8d20fa3c44305cbbfe0da6ce8579a2ba0380e/data/odia-or-wordlist.txt originaldate: 2021-10-20T23:57:09Z.

[25] Subhashish Panigrahi. 2021. Help: Recording pronunciations. https://or.wiktionary.org/s/16h9 Page Version ID: 177311.

[26] Subhashish Panigrahi. 2022. Before AI: nort2660-wordlist.txt. https://github.com/ofdn/Before-AI/blob/45b5dc31c7b2fb376a68767573b472f7cf7861ca/data/nort2660-wordlist.txt original-date: 2021-10-20T23:57:09Z.

[27] T. Shrinivasan. 2021. Odia Wordlist from Wikimedia Dump. https://github.com/ofdn/odia-wordlist-from-wikimedia-dump original-date: 2021-11-29T15:00:15Z.

[28] The Editors of Encyclopaedia. 2020. Odia language: Region, History, & Basics. https://www.britannica.com/topic/Odia-language

[29] Videowiki Contributors. 2022. Videowiki project - or - Wikimedia Commons. https://commons.wikimedia.org/wiki/Category:Videowiki_project_-_or

[30] Wikimedia Contributors. 2022. Category:Lingua Libre pronunciation by Psubhashish. https://petscan.wmflabs.org/?psid=21604903

[31] Wikimedia Contributors. 2022. Category:Voice intro project - Wikimedia Commons. https://commons.wikimedia.org/wiki/Category:Voice_intro_project

[32] Wikimedia Contributors. 2022. orwikisource dump progress on 20220301. https://dumps.wikimedia.org/orwikisource/20220301/

[33] Wikimedia Contributors. 2022. orwiktionary dump progress on 20220301. https://dumps.wikimedia.org/orwiktionary/20220301/

[34] Wikimedia Contrubutors. 2022. Index of /datasets: Q336-ori-Odia.zip. https://lingualibre.org/datasets/Q336-ori-Odia.zip

[35] Wikimedia Foundation. 2022. Wikimedia Downloads. https://dumps.wikimedia.org/

MUSAMONI PANIGRAHI (1920s—2017) was a storyteller and community elder from Balasore district, Odisha, India. The only archival recordings of her stories and songs were made into the short documentary Nani Ma (2022), directed by her grandson, Subhashish Panigrahi. Recordings of over 600 audio files in the 1920s register of the Balesoria (Northern) dialect of Odia, which are featured in OpenSpeaks Voice: Odia, are extracted from the same archival media.

SUBHASHISH PANIGRAHI is a public interest archivist, researcher, civil society leader, and non-fiction filmmaker from India, known for directing documentaries such Gyani Maiya (2019), MarginalizedAadhaar (2021), Nani Ma (2022), Remosam (2019), Mage Porob (2019) and The Volunteer Archivists (2022). He founded OpenSpeaks, co-founded the nonprofit O Foundation, and led community and programmes at organisations such as Ashoka, Wikimedia, Internet Society and Mozilla.

Subhashish Panigrahi
Musamoni Panigrahi

I SBN 9781738680825

9 781738 680825