What's Not on the Plate? Rethinking Food Computing through Indigenous Indian Datasets

Pamir Gogoi,
Neha Joshi,
Ayushi Pandey,
Vivek Seshadri
Karya Inc., Bengaluru, Karnataka, India

KARYA

Deepthi Sudharsan,
Kalika Bali
Microsoft Research
Bengaluru, Karnataka, India

Microsoft Research

Saransh Kumar Gupta,
Lipika Dey,
Partha Pratim Das,
Ashoka University
Sonepat, Haryana, India

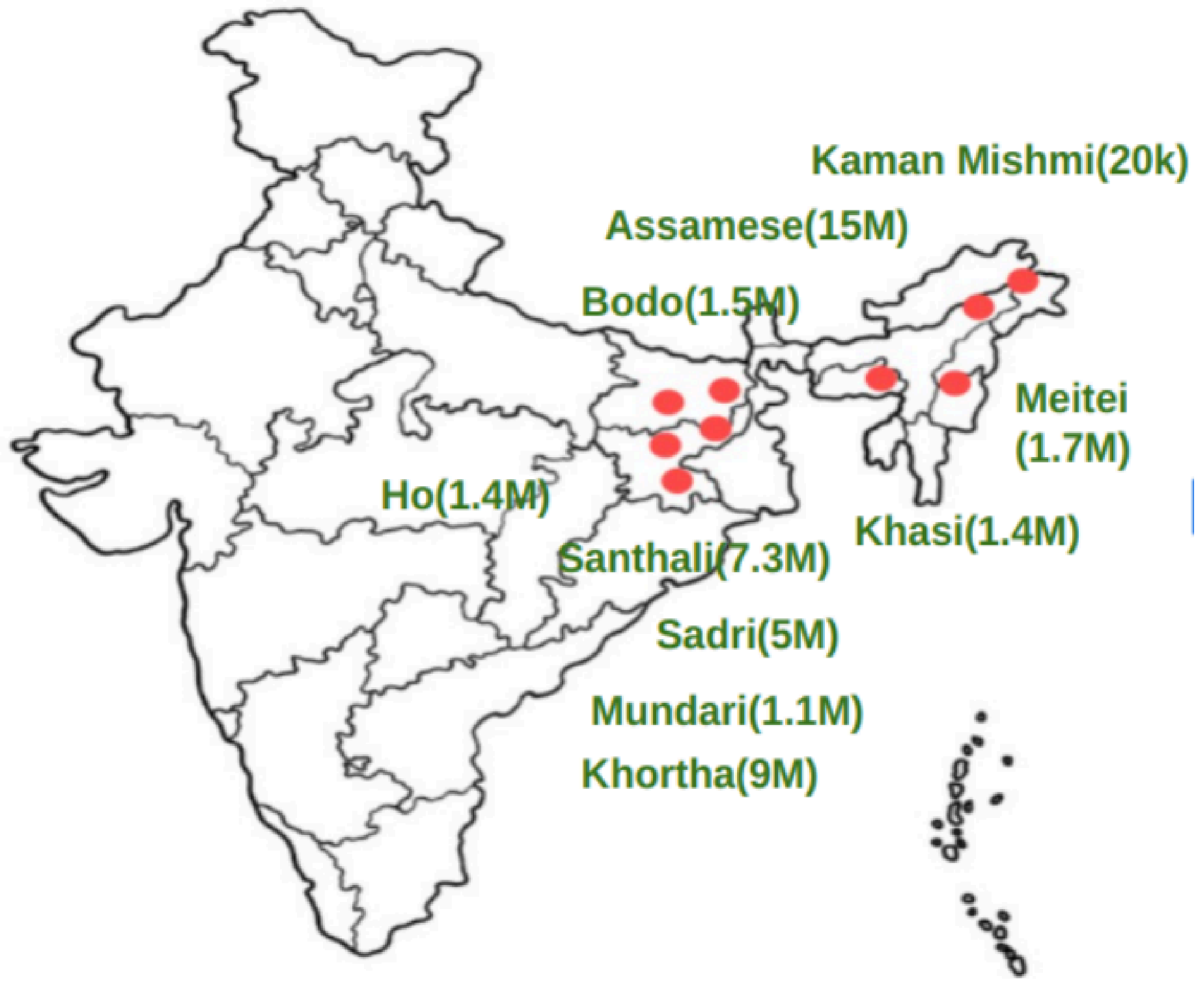
ASHOKA UNIVERSITY
Mphasis AI & Applied Tech Lab

ABSTRACT

This paper presents a multimodal dataset of 1,000 indigenous recipes from remote regions of India, collected through a participatory model involving first-time digital workers from rural areas. The project covers ten endangered language communities in six states. Documented using a dedicated mobile app, the data set includes text, images, and audio, capturing traditional food practices along with their ecological and cultural contexts. This initiative addresses gaps in food computing, such as the lack of culturally inclusive, multimodal, and community-authored data. By documenting food as it is practiced rather than prescribed, this work advances inclusive, ethical, and scalable approaches to AI-driven food systems and opens new directions in cultural AI, public health, and sustainable agriculture. By integrating this rich data resource with the Indian Food Knowledge Graph (FKG.in), the project advances culturally aware AI applications, such as personalized nutrition and translation for low-resource languages."

THE DATASET

- 1,000 traditional recipes
- 10 endangered Indian languages
- Geographic coverage: Jharkhand, Bihar, Assam, Manipur, Arunachal Pradesh, Meghalaya
- 338 rural women participants(mainly aged 15-45)
- Data formats: text, audio, images



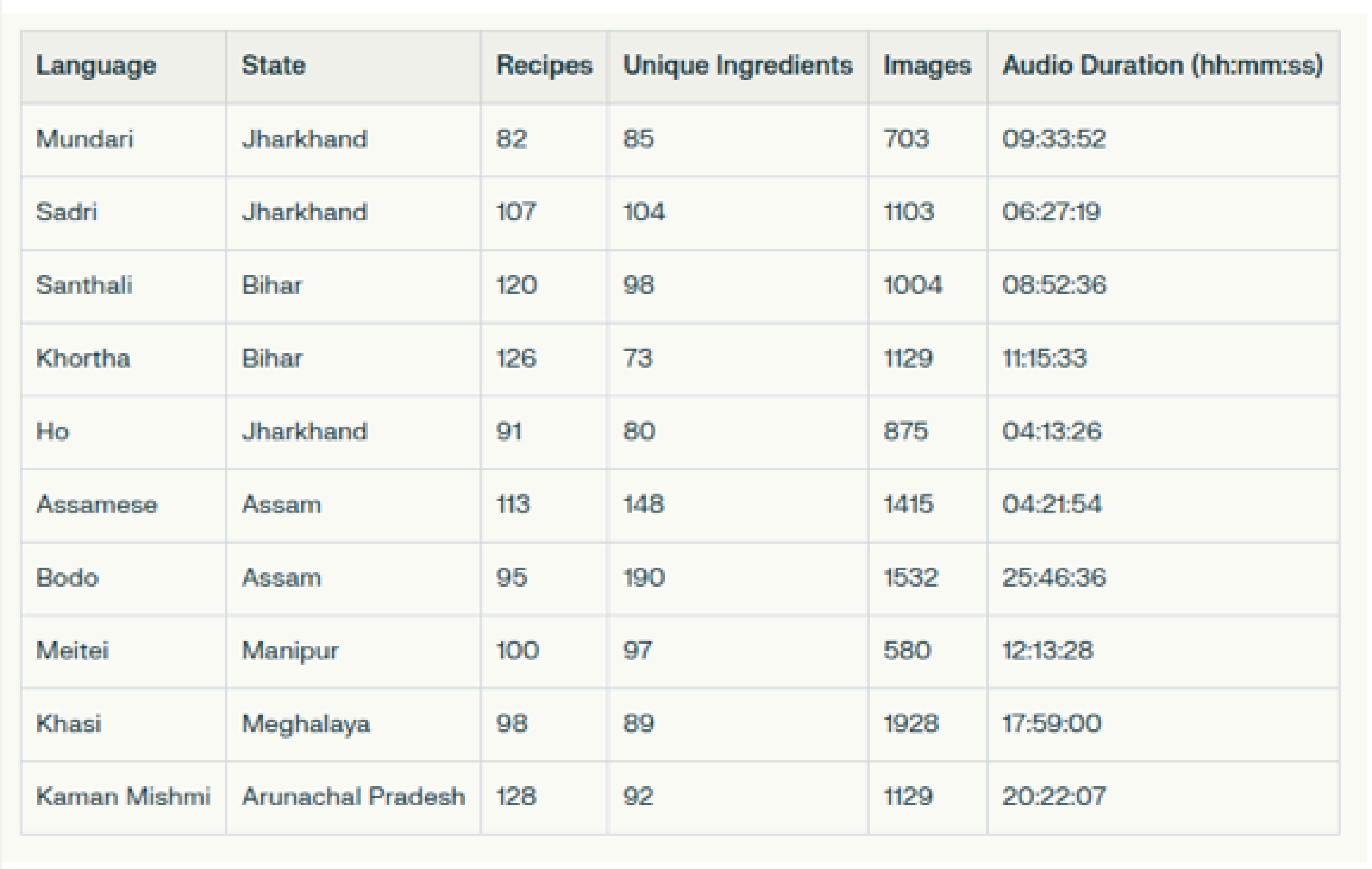
Data Samples

Chambai,
Kaman Mishmi dish

Amaltas flower
(Cassia fistula)

Red ant eggs

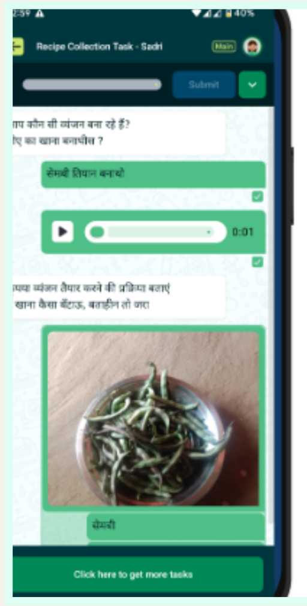
Kachnar flower *dish*
(Bauhinia variegata)

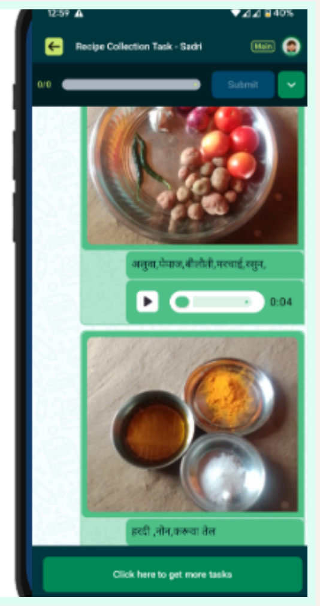


Language	State	Recipes	Unique Ingredients	Images	Audio Duration (hh:mm:ss)
Mundari	Jharkhand	82	85	703	09:33:52
Sadri	Jharkhand	107	104	1103	06:27:19
Santhali	Bihar	120	98	1004	08:52:36
Khortha	Bihar	126	73	1129	11:15:33
Ho	Jharkhand	91	80	875	04:13:26
Assamese	Assam	113	148	1415	04:21:54
Bodo	Assam	95	190	1532	25:46:36
Meitei	Manipur	100	97	580	12:13:28
Khasi	Meghalaya	98	89	1928	17:59:00
Kaman Mishmi	Arunachal Pradesh	128	92	1129	20:22:07

Data Collection Methodology

- Participatory approach with first-time digital workers
- Use of a dedicated mobile app with low text, offline support, and audiovisual cues
- Local coordinators recruited participants and managed data quality
- Contributors paid INR 750 per recipe to ensure fair compensation





APPLICATION & RESEARCH DIRECTIONS

- AI-driven procedural soundness checks for recipes using LLMs
- Culturally aware machine translation for endangered Indian languages
- Creation of culturally rich datasets for benchmarking AI models
- Multimodal understanding and generation (text, audio, image)
- Contextual reasoning cooking assistants sensitive to culture and allergens

Integration with FKG.in Knowledge Graph

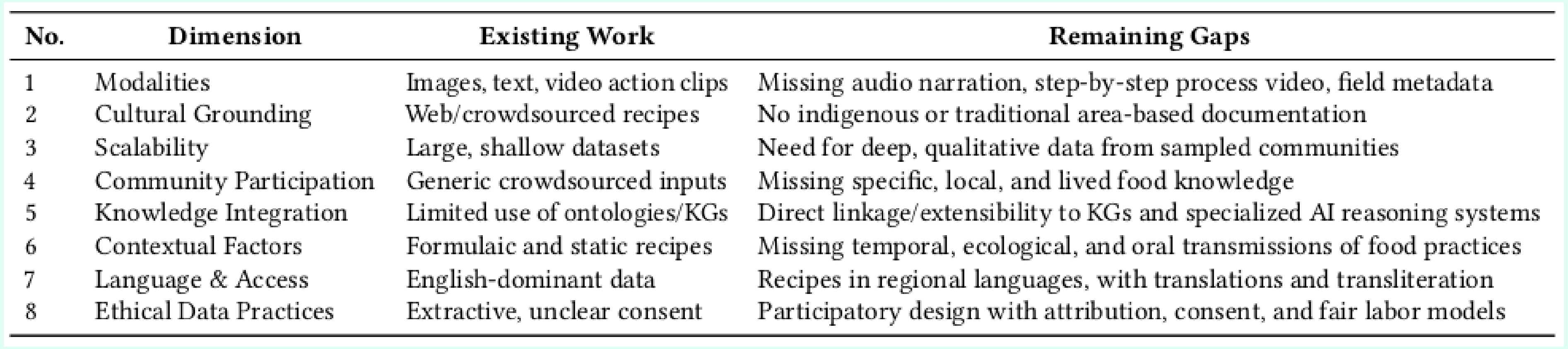
- Indian Food Knowledge Graph (FKG.in) connects data on ingredients, recipes, nutrition, claims, culinary practices.
- Integration enriches AI capabilities with cultural, ecological, and health contexts.
- Enables advanced applications like personalized health interventions and culturally sensitive cooking assistants.

Challenges Encountered

- Skepticism and trust-building with participants about digital work and payments
- Seasonality leading to limited recipe availability
- Infrastructure challenges like limited network - handled by offline mode of the app
- Balancing fair compensation for complex recipes
- Ensuring consistent data quality across diverse languages and literacy levels

This work addresses several gaps identified in existing research, particularly by integrating multimodal, culturally grounded, and community-driven data collection with ethical, participatory design practices. It contributes new, localized, and richly contextual datasets that enhance inclusivity, scalability, and knowledge integration across languages and regions.

Overview of Existing Work and Remaining Gaps in Food Knowledge Documentation



No.	Dimension	Existing Work	Remaining Gaps
1	Modalities	Images, text, video action clips	Missing audio narration, step-by-step process video, field metadata
2	Cultural Grounding	Web/crowdsourced recipes	No indigenous or traditional area-based documentation
3	Scalability	Large, shallow datasets	Need for deep, qualitative data from sampled communities
4	Community Participation	Generic crowdsourced inputs	Missing specific, local, and lived food knowledge
5	Knowledge Integration	Limited use of ontologies/KGs	Direct linkage/extensibility to KGs and specialized AI reasoning systems
6	Contextual Factors	Formulaic and static recipes	Missing temporal, ecological, and oral transmissions of food practices
7	Language & Access	English-dominant data	Recipes in regional languages, with translations and transliteration
8	Ethical Data Practices	Extractive, unclear consent	Participatory design with attribution, consent, and fair labor models

ACKNOWLEDGEMENTS

We gratefully acknowledge the local coordinators and contributors whose time, knowledge, and effort made this work possible. We are thankful to the Mphasis AI & Applied Tech Lab at Ashoka - a collaboration between Ashoka University and Mphasis Limited - for their support. We also thank Microsoft Research for their support.