Contribution ID: **2** Type: **Talk**

Building a Privacy-Preserving Conversational Data Infrastructure to Advance African Language Technologies and Data Sovereignty

South African and broader African languages are scarcely present in open training data for natural language processing despite the growing quest toward localised tools and solutions and AI sovereignty. This necessitates a need to grow low resource language datasets at an exponential rate. Cisha presents a privacy-preserving research data infrastructure that ethically sources, anonymises, and distributes consented conversational corpora in targeted languages. We run sponsor calls capped at three languages to balance coverage and review throughput; each submission passes two complementary anonymisation layers: deterministic regular-expression scrubbing of personally identifiable information, followed by named entity recognition with an artificial intelligence platform to capture additional entities and identifiers. A human-in-the-loop review flow records auditable edits, while a row-level security backend enforces least-privilege access and dataset partitioning. A credit-based marketplace links contributors, sponsors, and researchers with explicit, revocable consent, native-language dataset naming, and time-bound dataset leases for traceability and reproducibility. The result is consented, anonymised conversational datasets ready for large language model fine-tuning, automatic speech recognition, machine translation, and product localisation, directly serving approximately three hundred million African language speakers. Our contributions strengthen data infrastructure, security, and cyber resilience through multi-layer anonymisation, audit logging, and least-privilege controls; advance data literacy and data sovereignty through a transparent sponsor–researcher marketplace and credit economy that embed informed consent and culturally grounded stewardship; and enable public–private data sharing through call-based sponsorship that connects funders, communities, and researchers under auditable, policy-bound access. Cisha aligns with the principles of findable, accessible, interoperable, and reusable data, and with collective benefit, authority to control, responsibility, and ethics. This innovation shows how a purpose-built repository for low-resource conversational data can close the African language gap in artificial intelligence and machine learning and accelerate inclusive, locally relevant products.

Primary author: Mr BONZA (PC), Majozi (University of Pretoria)

Presenter: Mr BONZA (PC), Majozi (University of Pretoria)

Track Classification: DIRISA