# SADiLaR's repository as a cyber-infrastructure: Improving data accessibility for South African languages

Menno van Zaanen and Jessica Mabaso
South African Centre for Digital Language Resources
North-West University
Potchefstroom
South Africa
{menno.vanzaanen, jessica.mabaso}@nwu.ac.za

Computing processes typically require input data to perform actions that generate output data. While input data can sometimes be generated computationally, it often originates from external sources. In Natural Language Processing and Digital Humanities, this input is typically sourced from human activities, including spoken or written language and music.

In the current era of Large Language Models (LLMs) that provide practically usable tools, access to appropriate training data is essential. These models generally perform better when larger data collections are available for training, making data accessibility crucial.

For most South African languages, only limited amounts of digitally accessible data are available. Many existing data collections are sourced from government websites, providing texts in highly specific genres. Texts from diverse genres — including newspaper articles, literary works, and social media data — are not openly accessible in digital formats.

The SADiLaR (South African Centre for Digital Language Resources) repository hosts data collections that are as openly accessible as possible. The repository currently contains 357 directly downloadable items and 56 metadata-only items (indicating the existence of data collections).

The underlying principle of SADiLaR's repository is that providing a centralized space for data collections makes them more easily findable and accessible. (Additionally, submitted data collections are requested to be as interoperable and reusable as possible to ensure adherence to FAIR principles.)

This abstract serves as a call for action with two main objectives. First, we encourage researchers to submit their digital language data to the SADiLaR repository. Contributing to the repository increases the availability of South African language data, making it more easily findable and accessible. This data can then be used for training models, ultimately benefiting language users, for

example, through the development of LLMs for these languages. (Note: copyright remains with the original copyright owner; contributing to the repository does not transfer copyright ownership.)

Each contribution to SADiLaR's repository receives a persistent identifier, enabling consistent referencing of data collections. These identifiers can be used as citations in publications, ultimately benefiting researchers associated with the data collections.

Second, we encourage researchers to utilize the data collections available in SADiLaR's repository. The repository contains a wealth of useful data collections, and searching there first can streamline research processes. SADiLaR's repository exists to facilitate work in Natural Language Processing and Digital Humanities; collectively, we can leverage this cyber-infrastructure to advance our fields.