SADiLaR
South African Centre for Digital Language Resources

On SADiLaR and the open science initiatives

Professor Langa Khumalo
Executive Director: SADiLaR
27 August 2024

science & innovation
Department:
Science and Innovation
REPUBLIC OF SOUTH AFRICA

Background and context

SA Research Infrastructure Roadmap - SARIR

SADiLaR
South African Centre for Digital Language Resources

Department of Science and Innovation funded initiative

- launched in Sept 2016 (15 Years)
- to establishment and run the research infrastructures
  1) Facilities, resources, and services
  2) For the scientific community
  3) Generate, exchange, and preserve knowledge
  4) Strong support for notions of open data and open access
- 13 RIs established covering 5 scientific domains
- SADiLaR is the only one focusing primarily on the **Humanities**

SADiLaR
South African Centre for Digital Language Resources

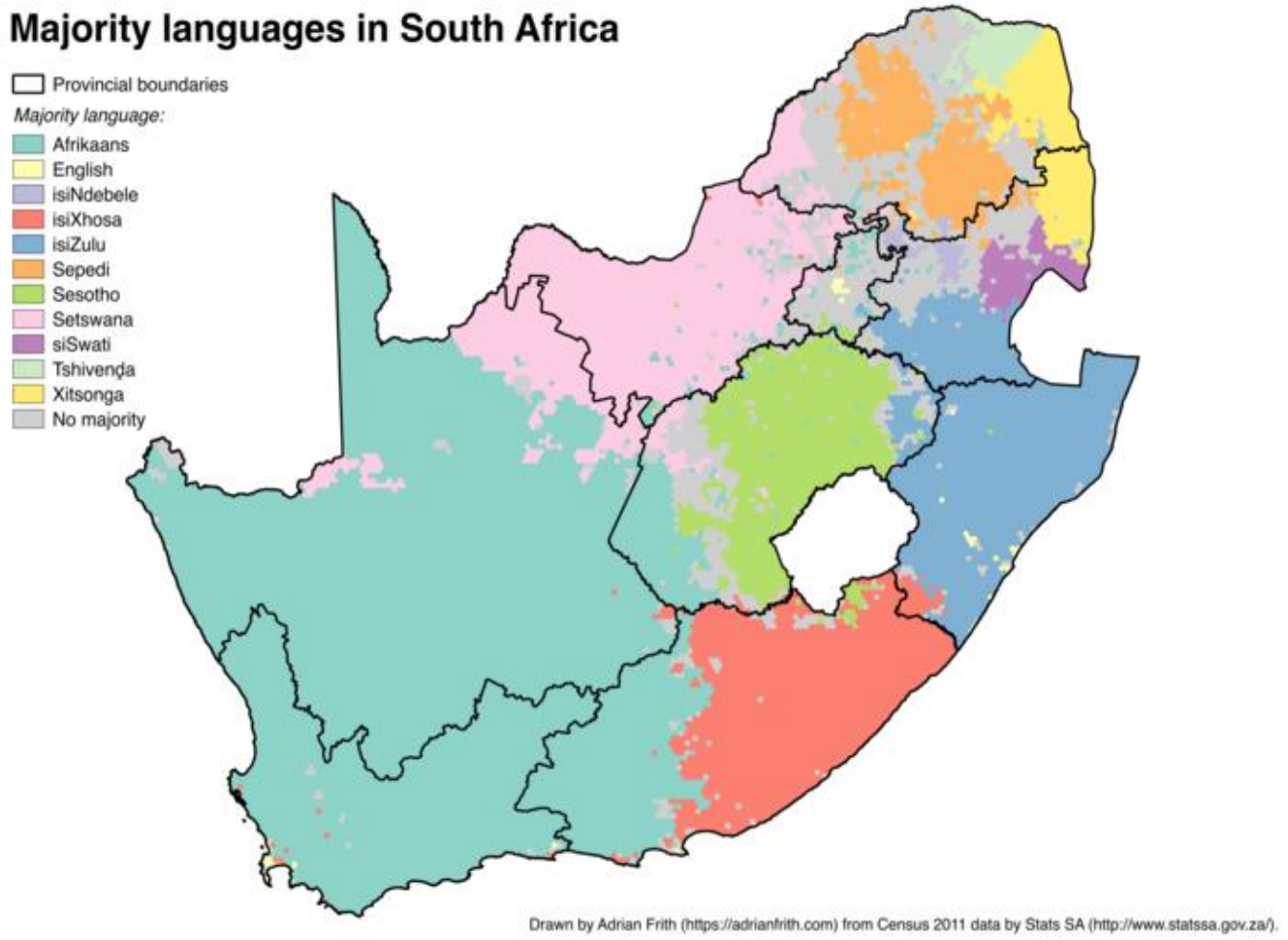
Mandate and mission

Ensuring a digital future for all official languages in South Africa.

SADiLaR
South African Centre for Digital Language Resources

South African context

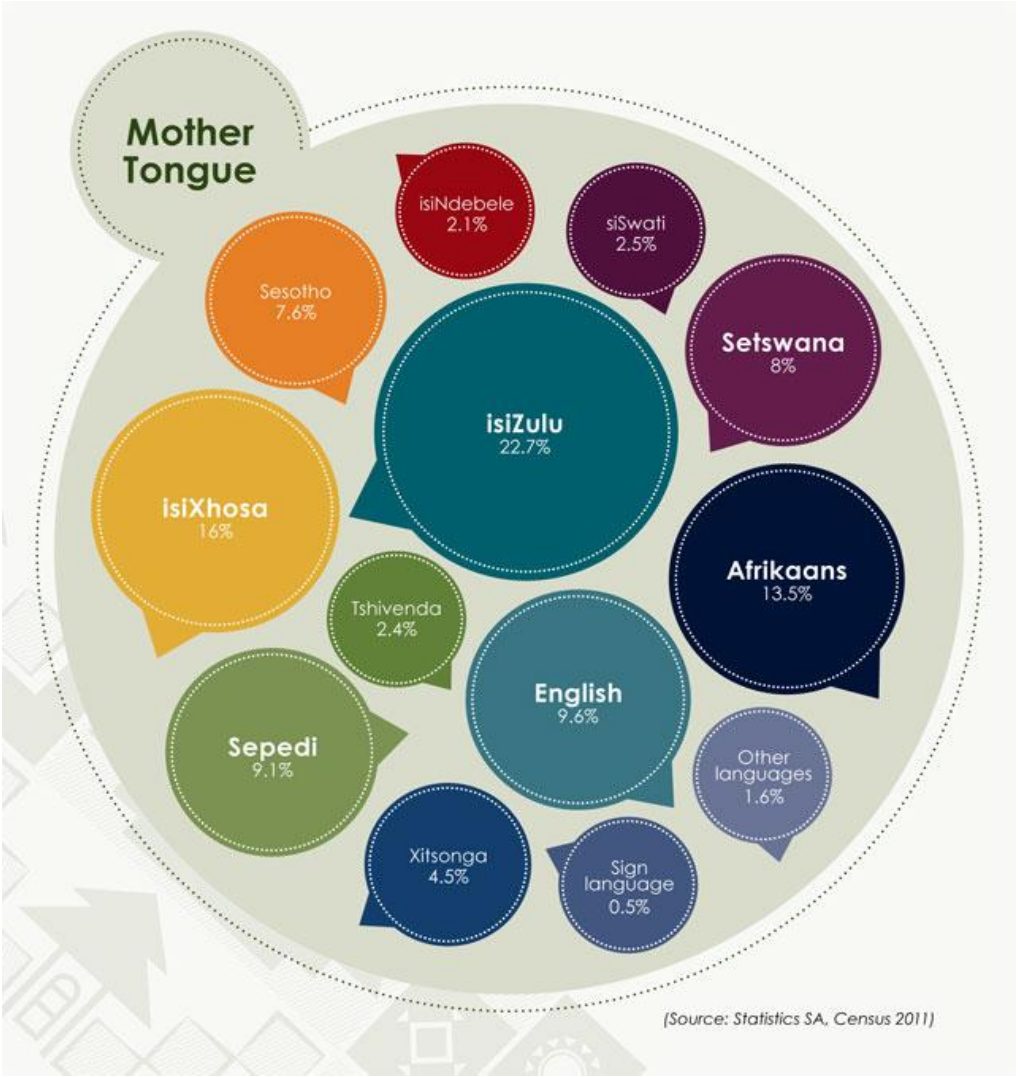
- 12 official languages
  - Conjunctively written Nguni languages (isiNdebele, siSwati, isiXhosa, isiZulu)
  - Disjunctively written Sotho languages (Setswana, Sesotho, Sepedi)
  - Disjunctive Xitsonga & Tshivenḓa
  - Afrikaans and English
  - South African Sign Language – official since 2023
- Constitutionally mandated that all languages are treated equally
- In practice, mostly English.



SADiLaR
South African Centre for Digital Language Resources

South African context

- Various HLT/NLP development efforts since turn of the century
  - Universities of Stellenbosch, Pretoria, South Africa, Cape Town (UCT), Johannesburg (WITS)
  - Council for Scientific and Industrial Research (CSIR)
  - Centre for Text Technology (CTexT®)
  - Multilingual Speech Technologies (MUST) @ North-West University
  - African Speech Technology Project (AST) 2000-2004
- Significant progress but
  - Still under-resourced - limited knowledge base



SADiLaR
South African Centre for Digital Language Resources

On the structure

The Hub and the Nodes

SADiLaR: Structure

❖Hub and spoke model, hosted at the North-West University, Potchefstroom

❖Currently 6 participating nodes in different areas of specialisation

- University of Pretoria – Digitisation
- University of South Africa – Semantics and terminology
- ICELDA – Language development and teaching
- CSIR – Speech resources and technologies
- CTexT, NWU – Text resources and technologies
- University of Stellenbosch – Child Language Development

❖International representatives in the Management structure

- CLARIN & ELRA (Europe) | Accepted as a full member of CLARIN-ERIC (01.01.2024)
- Linguistic Data Consortium (USA)

CLARIN ERIC
Common Language Resources and Technology Infrastructure

SADiLaR
South African Centre for Digital Language Resources

publications

MDPI

Article

Making Open Scholarship More Equitable and Inclusive

Paul Longley Arthur [1,*], Lydia Hearn [1], John C. Ryan [2], Nirmala Menon [3] and Langa Khumalo [4]

1 School of Arts and Humanities, Edith Cowan University, Mt Lawley, Perth, WA 6050, Australia; l.hearn@ecu.edu.au
2 School of Arts and Social Sciences, Southern Cross University, East Lismore, NSW 2480, Australia; john.c.ryan@scu.edu.au
3 School of Humanities and Social Sciences, Indian Institute of Technology Indore, Indore 453552, India; nmenon@iiti.ac.in
4 South African Centre for Digital Language Resources (SADiLaR), North-West University, Potchefstroom 2520, South Africa; langa.khumalo@nwu.ac.za
* Correspondence: paul.arthur@ecu.edu.au

Digitisation | DH| ESCALATOR | SWiP | Data Stewardship

SADiLaR
South African Centre for Digital Language Resources

Digitisation Programme

- Pro-active effort to extend language resources in official SA languages
- Node projects to develop language resources
  - Digital text and speech corpora from existing, non-digital resources (UP)
  - Wordnets and terminology development (UNISA)
  - Multilingual L2 learner corpus of academic writing and language specific academic literacy testing (ICELDA)
  - Transcribed speech corpora, automatic speech corpus collection and computational grammars (CSIR)
  - Annotated text corpus creation for conjunctive languages (CTexT)
  - Communicative inventories for South African Languages (SUN)
- Digitisation support to external institutions
- Provide assistance in best practices and digitisation efforts and software development

Digital Humanities Programme

- The Digital Humanities programme focuses on facilitating the building of research capacity by promoting and supporting the use of digital data and innovative methodological approaches within the Humanities and Social Sciences.

- The ESCALATOR programme stands central to all human capacity developments taking place under the DH programme

- More integrated programme – less working / training in silos
  - Work towards an inclusive and active South African CoP in Digital Humanities and Computational Social Science
  - Positive effects through broader engagement: Establishment of new DH Centers at RU, NMU and NWU

The ESCALATOR Programme

- ESCALATOR programme has three main activities
  - Multiple training/community tracks part of the **DIGITAL TRAINING & CHAMPIONS** initiative
  - **DH-IGNITEs** events continue to function as an important vehicle to bring together regions
  - **STAKEHOLDER MAP** and dedicated slack channel – support and facilitate access to expert practitioners in the field.

- Further highlights include
  - Multiple collaborations that saw mentorship training taking place; general computation training through CODATA-RDA School of Research Data Science and four conferences of the Digital Humanities of Southern Africa Association.
  - The establishment of a Journal for DHASA
  - Some activities saw also broader recognition such as the the N|uu dictionary project that recently won an award.

Promoting open science

- Central to the ESACALATOR programme is contribution towards general Open Science awareness and capacity building. To foster a science aware and science literate society.

- Core to this stands SADiLaR's SWiP project - "Preserving Languages: Open, Free and Accessible Knowledge for All."

- SADiLaR has been involved in Open Science and Data awareness as part of its contractual milestones in the past.

- To transition this to a more focused initiative through existing networks and platform by starting a regular, at least yearly Data Stewardship Summer School to strengthen these national efforts.

Conclusion

- **Long-term digital preservation and legal distribution of resources** still central to SADiLaR longer term goals

- In terms of **capacity building and awareness** workshop and events SADiLaR has been involved in nearly **200 such activities reaching more than 5300 participants** across activities

- Since the inception of SADiLaR - the addition of more than **150 new language resources as part of its repository** (previously RMA)

- SADiLaR's role as a non-competing entity has the **potential to enable long term impact in South Africa** to help ensure that isolated initiatives are linked to broader (national) activities

science & innovation

Department:
Science and Technology
REPUBLIC OF SOUTH AFRICA

SADiLaR
South African Centre for Digital Language Resources

Ngiyabonga
Enkosi
Kea leboga
Dankie
Thank you