science & technology

Department:
Science and Technology
REPUBLIC OF SOUTH AFRICA

POLICY DOCUMENT

DATA ACCESS POLICY

<Authors: Dr ER Eiselen>

<Date: 27 February 2019>

1. INTRODUCTION

The South African Centre for Digital Language Resources (SADiLaR) is committed to the principles of open data, especially for research data and tools created by government funding, which further the various research fields supported by SADiLaR. Open data supports various scholarly endeavours, including validation, replication, re-analysis, and reinterpretation, thus supporting the notion that all research results should be reproducible. Furthermore, open access to data allows for increased impact of research data developed with public funding, as the data can be reused in both research and private sector activities, which in turn will produce the greatest social good.

This is especially relevant in the South African context where ten of the eleven official languages are considered "under-resourced" or "resource-scarce" languages. For these languages, there are limited language data and tools available that enable research in the fields of linguistics, literature, digital humanities, and computational linguistics. By making data from all eleven languages more readily available, SADiLaR believes that the various language and research communities that are involved in the development of the languages will benefit in various ways, whether through better access to information in their mother tongue, or deeper insights into the structure and use of the language. This is also underscored by the Open Access statement of the National Research Foundation of 2015 (NRF), which specifically states that all publications funded by the NRF must be made available openly, while also promoting the establishment of open repositories that allow access to data from NRF-funded projects.

With these factors in mind, SADiLaR supports the notion of freely accessible resources that enable the further development of language-related research and development, wherever this is legal and ethical.

This policy was developed through various consultations with researchers and developers, as well as from reviews of various international data policy documents, including policies and guidelines from:

- Public Library of Science (PLOS);
- New York University (NYU);
- University of Cambridge; and
- Access guidelines from the Organisation for Economic Co-operation and Development (OECD), which were adopted by South Africa.

This policy provides the principles and framework within which all SADiLaR resources in the infrastructure are collected, analysed, catalogued, and distributed.

2. POLICY

SADiLaR is committed to applying the principles described in the OECD guidelines as far as possible, with consideration of the specific context of language resources and the South African context. The primary aim of the application of these principles will always remain the support of open and free exchange of ideas, information and knowledge, which broadens the scope and impact of the language resources developed and distributed by SADiLaR. The following policies address each of the OECD principles specifically in the context of SADiLaR.

2.1. Openness

SADiLaR will make all language resources developed with SADiLaR funding available to researchers in South Africa, Africa, and internationally, at the lowest cost and under the most liberal open access licence possible. The default distribution model within SADiLaR will be the Creative Commons Attribution only licence, with no associated cost, even for commercial purposes. For resources that are distributed by

SADiLaR, but not funded by SADiLaR, appropriate licensing options will be formalised with the owner of the resource to ensure that their ownership remains intact, but making the resource as widely available as possible.

All language resources distributed by SADiLaR will be available from an online institutional repository that is available to the public without restriction, unless specifically prohibited in terms of 2.4 – 2.6 below.

Any moratorium on data access will be applicable for a maximum of six months from submission, to allow principle investigators the opportunity to report on their resources, after which the moratorium will expire and data will be made available under the appropriate licence. Only in exceptional circumstances will data be made available under restrictive licences which limit the availability of the data to the public and research community.

2.2. Flexibility

Given the ever-changing nature and unpredictability of information technology, legal practices, ethical matters, and research policy, methodologies and innovation, SADiLaR is committed to being flexible in terms of updating and integrating new knowledge, legal restrictions, and changes in regulatory environments in the implementation of this data access policy. SADiLaR will continue to review new technology, legislation, and scholarly environments to ensure that the language resources distributed by SADiLaR continue to serve the largest possible community.

2.3. Transparency

SADiLaR will provide access to language resources in a transparent fashion, specifically communicating the origin, authorship, availability, etc. of the language resources in the metadata associated with all language resources. SADiLaR will also enforce the use of both national and international standards in the development and distribution of language resources as far as possible, given the regulatory environment and financial limitations of projects and resources associated with SADiLaR.

2.4. Legal conformity

SADiLaR will adhere to all South African laws governing the distribution and use of language resources. Within the South African context, two restrictions are of particular interest, and are specifically addressed in this policy document:

2.4.1. Copyright (Copyright Act, 98 of 1978)

Since language resources are specifically based on language utterances in various formats, the restrictions placed on copyrightable material is of great significance to SADiLaR. In all cases, material that is distributed by SADiLaR will be reviewed to ensure that the material included in a language resource does not have copyright restrictions, or if there are restrictions, that the necessary copyright clearance from the rights holder has been attained. No material will be distributed where it is not clear that the copyright for the given resource has been resolved.

Although every effort will be made to ensure that there is no copyright infringement in the language resources, the responsibility of infringement will remain with the party providing the language resources. In any case where there is uncertainty, or enquiries about a language resource, SADiLaR will cease to distribute the resource until such time as the matter has been resolved. All costs and legal services required to resolve the infringement will be the responsibility of the data provider, not SADiLaR.

SADiLaR will also make the licensing of all resources clearly available through their distribution channels to ensure that data users are fully aware of any restrictions on the data. The user of the data will be liable for any infringement of the copyright arising from the failure of the user to abide by the terms of the licence(s) associated with each resource.

SADiLaR will not take copyright ownership of any language resource, unless specifically agreed to with the copyright owner, and will merely act as a distribution agent for language resources of which it is not the copyright owner.

2.4.2. Protection of Personal Information (Act 4 of 2013)

SADiLaR will attempt, as far as it is possible, to ensure that any data developed or distributed by SADiLaR does not disclose any personal identifiable information of individuals, where the creators of the data have not received explicit permission to disclose such information, or where ethical clearance has been granted to do so.

2.5. Protection of intellectual property

All intellectual property (IP) vested in a particular language resource will remain the property of the creator/author of the resource, and in as far as SADiLaR has influence over the IP in a particular resource, SADiLaR will not divulge any IP not owned by SADiLaR.

2.6. Formal responsibility

In order to ensure the legal distribution of language resources, SADiLaR will engage all parties that provide data to SADiLaR to enter into formal agreements that stipulate the specific conditions and time period under which the resources may be distributed by SADiLaR. These agreements will formalise the procedures for the submission, distribution, and ending of such agreements. A template for such a distribution agreement is available from SADiLaR on request.

Furthermore, all access arrangements will be communicated and approved by SADiLaR's steering committee to ensure the validity of these agreements and that all aspects of submission, access, and distribution are adequately covered by the agreements.

2.7. Interoperability

One of the key attributes of data access is ensuring that research data is re-usable in formats that are up to date and follow best practices in the respective fields of study. Although applying standards to existing data sets is near impossible, SADiLaR undertakes to ensure that data in formats that are not widely supported, are updated to be compatible with current standards and technologies. For all data that is newly acquired by SADiLaR, data authors will be encouraged to make use of current ISO and SABS standards, specifically emanating from ISO technical committee 37, which is tasked with setting standards related to the development, management and distribution of language resources. SADiLaR further undertakes to implement such standards as are applicable to the management and distribution of resources as specified by these ISO standards.

2.8. Quality

As with all data initiatives, the research and development that are applicable to any given data source is fully reliant on the quality of the data. SADiLaR will endeavour to ensure that resources developed for and distributed by SADiLaR are reviewed for quality based on industry standard quality metrics, which may include Accuracy, Precision, Recall, F-scores, Word error rates, and so forth. Wherever possible, the specific quality attributes to which a resource adheres will be communicated as part of the metadata associated with the data resource. SADiLaR will also work with language resource developers to ensure that the data standards that apply are well communicated and established prior to development efforts, and where possible, the evaluation procedures will also be made available to the public.

SADiLaR will develop documentation on the techniques, methods and procedures that ensure high quality language resources and metadata, and will distribute such documentation publicly.

2.9. Security

SADiLaR undertakes to ensure that all language resources and resource metadata are securely stored, with adequate backup, recovery and disaster management procedures for all data resources. As such, SADiLaR will distribute language resources across various virtual and physical domains to ensure that data is recoverable in a reasonable amount of time. SADiLaR also undertakes to put into place measures to ensure limited risk of unauthorised access or destruction of language resources under its stewardship.

2.10. Accountability

The data access policies of SADiLaR, as well as the procedures, standards, and technologies implemented at SADiLaR will be reviewed by each of the respective management structures within SADiLaR to evaluate the effectiveness and relevance of the data access and management procedures. Each of the following committees will evaluate and approve the procedures on a biannual basis:

- Management Committee;
- Steering Committee; and
- Scientific Advisory Committee.

2.11. Sustainability

In order to ensure that the research data managed by SADiLaR remains available in a sustainable fashion, SADiLaR will coordinate with various national and international data and resource agencies to further distribute the availability of the data. Specifically, SADiLaR will work with the Data Intensive Research Initiative of South Africa (DIRISA) to mirror existing research data and ensure long-term sustainability. SADiLaR will also work on a wind-down plan, which will specifically plan for the continued availability of language resources beyond the duration of SADiLaR. This plan will be approved by the Department of Science and Technology and communicated to the various management committees of SADiLaR.