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

SADiLaR
South African Centre for Digital Language Resources

Policy: Data Acquisition and Distribution

1. INTRODUCTION

SADiLaR is a national centre supported by the Department of Science and Innovation (DSI) as part of the new South African Research Infrastructure Roadmap (SARIR). "SARIR is a high-level strategic and systemic intervention to provide research infrastructure across the entire public research system, building on existing capabilities and strengths, and drawing on future needs." SADiLaR has an enabling function, with a focus on all official languages of South Africa, supporting research and development in the domains of language technologies and language related studies in the humanities and social sciences. SADiLaR supports the creation, management and distribution of digital language resources, as well as applicable software, which are freely available for research purposes. (See https://www.sadilar.org)

SADiLaR clients include academic scholars and professionals in all domains of Humanities and Social Sciences, Language Technologies, Natural Language Processing, Computer Science, as well as potential end-users in education, business and industry.

The South African Centre for Digital language resources has as one of its primary aims, the creation and distribution of digital versions of language data and resources, which will be integrated into collections that allow for digital scholarship. This includes works in all of the South African languages.

Data and resource creators have copyright over their work. SADiLaR functions as a distribution channel for data and resource creators work and does not hold copyright over any of the resources distributed.

Data and resource creators must license how their copyrighted data or resource may be accessed, copied and used to make derivative works by consumers of the data or resource. Resource creators must choose if they allow commercial usage and if they allow creation of derivative works. Resource creators must be aware of the ethical issues, including confidentiality, associated with distribution of their resource. Resource creators must be aware of the licenses under which any input resource was distributed and consider the inheritance of and license condition. These decisions determine the level of copyright restrictions applied to the distribution of the resource.

This document provides an outline of the licenses under which SADiLaR proposes to distribute the language resources. These licenses specifically determine how the data may be accessed and used for various purposes, such as research or development. Furthermore, the document outlines different models through which the resources can be distributed. There are several different options for distributing, or limiting the distribution, of data through the licences and technologies under which the various resources are distributed. In order to accommodate as wide a variety of resource providers as possible, this document provides a tiered approach, from the most lenient distribution agreements, to those that restrict access to and the way in which these resources can be used.

Software distribution is more complex as there are generally multiple source packages which can be under multiple license types. Source software packages that are re-used can inherit the license of the original distribution. Contributors can only select the license type of the new packages. Software distribution by SADiLaR must follow the license agreements of the original packages in the case of source code distribution and must follow an allowable license when distributed in target form.

2. RESOURCE LICENSING

In all environments that distribute, use, or make data available, the main tool for establishing how the resources can be used is the license that accompanies the data. Although there are a plethora of different licenses, the most commonly used licenses are the Creative Commons licenses, where a relatively straightforward set of questions can provide a user with a licence that has the appropriate restrictions. All of the licenses are available from https://creativecommons.org/ with the specific selection of a license available from https://creativecommons.org/share-your-work/. Many of the licensing options discussed here relate either directly, or are derived from these licenses.

2.1. Open access licensing

The most lenient of the licence options is open licences that allow for the use of the data in any format for any purpose, whether commercial, non-commercial, or research. The most common of these licence types is the Creative Commons Attribution license, where the only restriction on the use of the data is that attribution must be provided to the creators/copyright holders of the resource. The resources may be incorporated into other resources, changed in any way and used for any purpose. Furthermore, the resource may be distributed further by any party that has downloaded and accepted the licensing conditions.

A second open access option is a license that specifies that there may be no derivatives of the resource, i.e. no changes or derivations of the resource may be made, and the resource must be used and distributed in the exact format distributed by SADiLaR. This is typically distributed under the CC Attribution-NoDerivatives license.

A third open access option is a license that specifies that there may be derivatives of the resource, but all derivatives must also be distributed under the same or a compatible license.

This type of resource will be stored in the SADiLaR repository without requiring any authentication control to download the resource.

2.2. Academic / Non-commercial licensing

A more restrictive license is an non-commercial license, which allows users to use and distribute the resource only for non-commercial purposes. The same options with regard to derivatives and sharing are available which would restrict or allow the licensee from creating derived resources, and optionally require licensees to distribute any derivatives under the same license.

This type of resource will be stored in the SADiLaR repository requiring authentication with academic credentials to download the resource. People without academic credentials can request a credential via CLARIN (https://www.clarin.eu/).

2.3. Restricted and commercial licensing

An even more restrictive license makes provision that only the licensee may use the resource, where the use in commercial or non-commercial settings are specified with the license. The specific application of the resource use can also be specified in the specific license and can include restrictions such as:

- Limiting the type activity that the resource may be used for, e.g. creating a word list, inclusion in a corpus, evaluation of technologies, creating statistical models. This type of restriction is typically used to prevent the licensee from infringing on the copyright holders business model.
- Specifically limiting to activities that may not be performed with the data, e.g. explicitly stating that the resource may not be published in its original or derivative format.
- Limiting the distribution of confidential or ethically sensitive information.

This type of resource can be stored in the repository and distributed through a channel requiring authentication and specifically approved authorisation to download. This type of resource can also be held in the SADiLaR repository as a metadata only record as specified in section 3.3.

3. DISTRIBUTION OPTIONS

In addition to the restrictions that can be placed on the use of particular resources, there are also different distribution models for resources distributed by SADiLaR. These distribution options can typically be used to protect certain aspects of copyright material or confidential material, while still making resources available in as open access a format as possible.

3.1. Original format

The most open distribution option is making the resource available in its original format, including ordering, layout, etc. where it would typically be possible to completely recreate the original work in its entirety from the resource. This original format may also be adjusted to reflect standard distribution models, such as converting the original work to an XML format with metadata that allows more structured access to the data, but in principle, the original work could still be recreated without much effort.

3.2. Full access, restricted format

A second distribution model is making all of the data available, but in a format that makes it impossible to recreate the original copyrighted work from the restricted format. Examples of such restricted formats include:

- Randomising data on a particular layout or syntactic level such that there are never units, e.g. sentences or paragraphs, that follow one another as it did in the original copyrighted resource.
- Restricting access to specific lexical or syntactic elements, with limited contextual access to the resource, e.g. Key word in context searches.
- Restricting access to the resource by integrating the resource into a larger collection of resources where the origin of the specific instance may be obfuscated.
- Restricting access to only derivatives of the resource, where the derivation has no information that would allow reconstruction of the original resource, such as:
    - Word and/or frequency lists; and
    - Statistical models, such as n-gram or other language models.

Lastly, SADiLaR can also place restrictions on the time periods for which a resource will be made available. Two examples would be that the resource is only available for a restricted time, or that a resource only becomes available after some mutually agreed upon date.

3.3. Restricted access

A third possible distribution option is to make the metadata of a resource available, but not the resource itself. The metadata will then contain a reference to a contact person who can provide potential users with access to the resource on a case-by-case basis, and where necessary set up a costing structure that allows the resource to be used for commercial purposes. In this case, SADiLaR will not distribute the resource, but only maintain and make metadata associated with the resource available via its online repository.