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

SADiLaR
South African Centre for Digital Language Resources

Data Management Plan
SADiLaR

1. INTRODUCTION

SADiLaR funds Node and Open Call projects. For the projects that SADiLaR funds, a Data Management Plan (DMP) is a mandatory deliverable. All SADiLaR funded projects should have the agreed data available in the repository or other deliverables in agreed locations at the conclusion of the project.

For projects or deliverable not funded by SADiLaR, but for which SADiLaR would host some or all of the deliverables, it is optional to have a DMP at the start of the research phase.

2. GOALS

The goal of the DMP is to define:

Concept	**Description**
Data Production	This is the basic definition of the data to be produced
Data Definition	This is the definition of the means for organising the data and the metadata, including the standards applied
Data Rights Management	This is the definition of the data right management including protection of privacy, maintaining security, confidentiality, copyright management, intellectual property, licensing, distribution and other rights
Data Reuse	This is the definition of how the data will be accessed and distributed if the data rights management supports distribution or creation of derivative works
Data Preservation	This references the archiving and preservation of data

3. TEMPLATE

The person responsible for the project should ensure that the template is completed through sections 3.1 to 3.5 and is submitted to SADiLaR for evaluation against the accepted practices in the domain.

3.1. Data Production

Data Production
Describe the purpose of the project or research. This can be done by reference to the project proposal where available
Describe the data which the project will deliver. Is it modelling data, new data acquisition, etc? Example: The project will deliver a prototype grade chatbot in 11 official languages. The chatbot is built on IBM's Watson with integration of language models for South Africa.
Describe if there is already data in the domain of this project, the limitations of the existing data or work and how this project will deliver enhanced results Example: The current dataset for the language are under resourced and contains 70% 1-gram and 30% 2-gram data. By increasing the number of tokens in the dataset and changing the composition to 40% 1-gram, 40% 2-gram and 20% 3-gram, the accuracy of speech recognition is expect to increase by at least 10%
Describe the approximate amount or minimum amount of data generated from this research Example: The project is committed to deliver a minimum of 100 hours of speech recording per official language of South Africa The expected size is around 2GB per language The speech recogniser models are derived from 1000+ hours of speech input as training data and are expected to be around 50MB in size per language The project is expected to derive 10000 pages of digitised information with expected size of 2MB per page The database will be populated with data from a minimum of 500 people at phase 1. Each person's data will be less than 1MB.

Define if the project uses output from any other projects and if so make reference to which datasets, models, etc. Example: https://hdl.handle.net/20.500.12185/508
Who (name of person) is responsible for managing the data and compliance to the Data Management Plan? Personname person@uni.ac.za

3.2. Data Definition

Data Definition
Is the data oriented for machine processing? Example: Yes, each modeling data file is stored as a JSON object Yes, the audio files are stored with the embedded metadata relevantly populated with the transcription of the audio sample Yes, the service is available through a REST API call No, the output is available in hardcopy form in the university library
Born Analogue – Original Data (complete only for analogue data sources) Describe the original format of the Analogue Master. E.g. Hand-written paper, printed paper, 16mm movie, audio cassette, etc and key attributes such as mono- or stereo audio, colour/B&W film, etc. Example: Archive audio cassette: Types: Normal and Chrome Analogue Noise Reduction: Dolby B Mono Recording Dynamic Range (maximum): 70dB
Born Analogue – Digital Remastering Process (complete only for analogue data sources) Describe the process to convert to a Digital Master. E.g. camera and lighting setup, scanner/OCR setup, video/film/audio conversion setup

Born Digital – Digital Creation Process (complete only for born digital data)

Describe the process and tools used to acquire the digital data. E.g. Full professional Studio audio/ smartphone audio, Full HD video professional camera/ smartphone camera, etc.

Example:
Photograph the original documents with Nikon D750 (24Mpixel Full Frame sensor) and 60mm f2.8 Nikkor Macro lens with LED ring lighting on copy stand. Onboard conversion to high resolution JPEG. Augment picture with text by OCR software acting on the JPEG input with human checking and correction. Merge the picture and text layers using Adobe Acrobat product to output as a pdf.

Describe the core format of the low-level content data comprehensively for audio, video, graphic/picture/visual, text data (optional depending on project output)

For example; Audio: file format, encoding standard, audio sample rate, Anti-aliasing filter cutoff frequency, etc; **Text**: file format, character mapping format, special characters, etc.; **Visual**: file format, layer format, etc; **Video**: file format, encoding standard, video fps, audio sample rate, Anti-aliasing filter cutoff frequency, resolution.

Example:
Audio: Analogue low-pass filter with -3dB loss at 18kHz and 6dB per octave roll-off. 44.1kHz sampling with 22kHz digital anti-aliasing filter. Speech samples mastered with 24-bit integer arithmetic.

Describe the format of any models that are produced as a result of the project (optional depending on project output)

e.g. Speech recognizer training model, text-to-speech model, tree-bank etc.

Example:
The REST API returns a JSON object with the audio encoded as Base64. The local machine should convert the Base64 data to binary data for audio.

Describe any data overlays to assist machine processing with reference to standards (optional depending on project output)

Mark-up e.g. XML/HTML/DITA, TEI; Machine structure: e.g. JSON/YAML; Multimodal data format: Adobe Acrobat. Describe any specific structure by reference to such as to a .xsd file

Example
There are no data overlays. The corpus is entirely 1-gram audio data wav file with a separate metadata file in Excel which defines the language and word sampled.

There are data overlays represented in the ELAN format. The continuous audio data of each radio discussion lasting 10+ minutes has been transcribed using the ELAN tool and the transcription file is available.
There are data overlays represented in the pdf. There is a picture layer and a text layer of the digitized handwritten analogue master. The text layer is built using OCR and then human correction.
There are data overlays in TEI format. The sampled audio radio show has been automatically transcribed using speech recognition and level of confidence in the decode is included in the TEI markup

Justify if the format is not included in https://www.clarin.eu/content/standards-and-formats

Example:
The project delivers software only and so is not included in the list

Describe the data packaging (e.g. zip, tarball, blob, etc)

The developed software will be available as a tarball (.tz) format in our institutional GIT repository for ingestion to SADiLaR's repo
All the digitised text document will be distributed in a zip archive (.zip)

Describe the metadata format. If the data is to be part of repo.sadilar.org then it is mandatory that the CMDI format is used for the highest level metadata describing the package. In this case it is not necessary to describe the metadata format, just define that the metadata would comply to CMDI

Multiple metadata formats should all be described. Lower level metadata (non-package level) can be in different formats to CMDI.

Top-level metadata will be CDMI generated by SADiLaR after the submission. Detailed metadata will be (i) embedded in the .wav file as an ID3 chunk and (ii) there will be a separate list correlating the text of each audio sample and language covering all audio samples

Describe the processes in acquisition or validation to ensure that the data captured in the project is compliant to the data definition

Example:
The software is tested against the test specification derived from the requirements specification by manual testing on (link to) example dataset

Describe any dependencies to view, use and reuse the data, e.g. open or proprietary tools for which a single option exists

Example The text is available in pdf format – it can be read by multiple tools The text is marked up using the Oxygen tool using our proprietary schema. The software is developed on the Linux OS and tested on CentOS 7. The code requires the httpd, tomcat, java and etil libraries and Postgres DB to be available. The website is built on Joomla with the XYZ plugin and the demo site built in a Kubernetes instance on a bare metal server

3.3. Data Rights Management

Data Rights Management
Describe the ethical clearance for this project and the current status Example: The data collection project has been submitted to the university ethic committee and the review process is in-flight The data collection has no ethical issues and has been reviewed and approved by the Head of Department
Describe the privacy issues of the data subjects and project workers and how their privacy rights would be managed Example: All data subject are anonymised by reference to a fictitious names. Real names and personal details of data subjects and project workers are managed in accordance with the organisations PoPI policy
Define any confidentiality issues with the project execution or project output generation and how these will be managed. Example The project derives algorithms for classification of the subject's emotional state from analytics of audio data gathered through consented call centre interactions. All call centre interactions are confidential as they contain personal and private data. The confidentiality will be managed through (a) consent gathering (b) storage of all gathered data privately in accordance with best practice (c) publication of the outcomes only (no raw data published of any recorded call)
Define any copyright issues with the project execution, input materials or input data, such as computer software, photographed images, audio/video broadcast, or project output generation and how these will be managed Example

The project digitizes out of copyright content – there is no copyright issue
The project packages copyrighted content and we already have the copyright permission from the publisher

Define any Intellectual Property Right issues with the project execution and input materials including:

- Patent related to the use of software and royalty fee and how these will be managed
- Restrictive license for software, e.g. GPL, use in development and how this would be managed

Who would be the contact person in perpetuity?

personName person@uni.ac.za

Select the distribution model for the project outputs:
1) Public - Unlimited distribution without any login
2) Academic – distribution to those with Academic credentials after login
3) Restricted – distribution to those meeting specific criteria on a request basis to be approved by the data owner

Example

The data is restricted as we are trying to commericialise the outcomes

The data is restricted as it contains confidential materials. It may be shared with researchers able to demonstrate a genuine need to access the data.

Define the license for the distribution of the project outputs

Define the distribution channel of the project outputs. Justify if the project outputs are not to be distributed through SADiLaR

The software is distributed through github. The reason is that a repository is optimised for long term data storage and is sub-optimal for distribution of software code.

The data is delivered to SADiLaR for distribution in the SADiLaR repository

The data corpus is kept internally on our own server in accordance with our preservation policy as the data is confidential and cannot be anonimised

For project outcomes that are restricted for commercial reasons, describe the commercial benefit

Define any changes with time for the data right management, e.g. project outputs under embargo until commercialised, etc.

Example:

The project outcomes are anticipated to be commercialized. The project outcomes can be publicly released 5 years after the first commercial product once there is commercial momentum around the product.

3.4. Data Reuse

Data Reuse
Define how the data or other project outcome will be made accessible, if the data is not distributed through SADiLaR Example: The data will not be accessible to other people as I keep it on a portable hard drive with a personal password The software will be fully accessible on github, where it will be maintained by us for at least 3 years after initial release The service will be available to system integrator behind a paywall
Define how the data or other project outcome will be made findable, if the data is not distributed through SADiLaR Example: The outcomes are made available through the repository which is part of the university library. The metadata is exposed to Google Scholar, from which other can find it.
Define how the environment will be interoperable by reference to standard, if the data is not distributed through SADiLaR
Define the people and process for managing the data on an ongoing basis, e.g. approving and distributing restricted data, data migrations, etc. Example: SADiLaR distributes metadata for the restricted resources and others can request distribution of the resource through email to the published contact person, which is part of the metadata. The contact person receives the request through email. They contact the requestor with an enquiry form to find out key data through which a distribution decision will be made. All requests for distribution of restricted resources will be administered by the contact person, and the decision maker for distribution approval is the Head of Department
Define if or how the data will be referenced through a persistent identifier, e.g. handle, DOI, if it is distributed outside SADiLaR. All project outputs distributed through SADiLaR have a Handle as a persistent identifier

3.5. Data Preservation

Data Preservation
SADiLaR has a Data Preservation Policy at https://www.sadilar.org/index.php/en/guidelines-standards/data-preservation-policy. Project outcomes distributed through SADiLaR will all have their data managed in accordance with this policy. Describe how data will be preserved should the project outcomes not be distributed through SADiLaR, in terms of: • Backup policy and operation • Response to infrastructure changes • Response to data format changes (obsolete data formats) • Response to closure of the body holding the data (repository closure) • • Example The data will be held in the university repository. This is operated in accordance with the linked file which details backup, operational processes and response to infrastructure changes. Should the university repository close, there is agreement with our partner university to take over the dataset. The data is stored as ELAN files and if this format is no longer maintained in the next 10 years, then the university would transform the data to TEI format or other contemporary format.
Define the period for which the project outcomes be retained For example: The data will be stored in the university repository for a minimum of 20 years.