Creating electronic resources for African languages through digitisation: a technical report

Taljard, Elsabé
University of Pretoria
elsabe.taljard@up.ac.za

Prinsloo, Danie
University of Pretoria
danie.prinsloo@up.ac.za

Goosen, Michelle
University of Pretoria
michelle.goosen@up.ac.za

Abstract

The need for electronic resources for (under-resourced) African languages is an often stated one. These resources are needed for language research in general, and more specifically for the development of Human Language Technology (HLT) applications such as machine translation, speech recognition, electronic dictionaries, spelling and grammar checkers, and optical character recognition. These technologies rely on large quantities of high-quality electronic data. Digitisation is one of the strategies that can be used to collect such data. For the purpose of this paper, digitisation is understood as the conversion of analogue text, audio and video data into digital form, as well as the provision of born digital data that is currently not available in a format that enables downstream processing. There is a general perception that the African languages are under-resourced with regard to sufficient digitisation tools to function effectively in the modern digital world.

Our paper is presented as a technical report, detailing the tools, procedures, best practices and standards that are utilised by the UP digitisation node to digitise text, audio and audio-visual material for the African languages. The digitisation effort is part of the South African Digital Languages Resources (SADiLaR) project (https://www.sadilar.org/index.php/en/), funded by the Department of Science and Innovation. Our report is based on a best practices document, developed through the course of our digitisation project and forms part of the deliverables as per contractual agreement between the UP digitisation node and the SADiLaR Hub. The workflow as explained in this document was designed with this specific project in mind; software and hardware utilised were also selected based on the constraints with regard to capacity and available technical skills in mind. We motivate our choice of Optical Character Recognition (OCR) software by referring to an earlier experiment in which we evaluated three commercially available OCR programmes. We did not attempt a full-scale evaluation of all available OCR software, but rather focused on selecting one that renders high quality outputs. We also reflect on one of the challenges specific to our project, i.e. copyright clearance. This is particularly relevant with regard to published material. In the absence of newspapers for specifically the African Languages (isiZulu being a notable exception), the biggest portion of textual material available for digitisation consists of printed material such as textbooks, novels, dramas, short stories and other literary genres. The digitisation process is driven by the availability of material for the different languages. Furthermore, obtaining copyright clearance from publishers is a prerequisite for digitisation and especially for the release of any digitised text data for further use and / or processing. Having information on a relatively small-scale digitisation workflow and best practices readily available will enable other interested parties to participate in the digitisation effort, thus contributing to the collection of electronic data for the African languages.

Keywords: digitisation, Optical Character Recognition (OCR), copyright, metadata, electronic resoources for African Languages

1 Text digitisation

For the purpose of text digitisation, we use an Epson DS-50000 portable flatbed scanner. Camera scanners are problematic in terms of quality and consistency.

Once the resources to be digitised have been identified, the software *ABBYY FineReader 14*, an OCR application, is used for the scanning process. The decision to utilise this particular software programme is informed by a small experiment, reported on in detail in Prinsloo, Taljard and Goosen (to appear). In this experiment, two commercially available OCR programmes i.e. *ABBYY FineReader 14* and *Omnipage Professional 18,* and one locally developed scanning package, *CTexTools* were compared on good quality printouts from President Cyril Ramaphosa's 2020 state of the nation (SONA) address (https://www.gov.za/ve/speeches/president-cyril-ramaphosa-2020-state-nation-address-13-feb-2020-0000) with reference to percentage of scanning errors and overall accuracy rate. Afrikaans, isiZulu, Sepedi and Tshivenḓa were used as test languages. Our results indicated that *ABBYY* would be the preferred OCR tool for languages not utilizing more than a minimum of diacritic signs, even though those languages may not be specifically supported by the software. For Sepedi, for example, the software does not recognize the frequently occurring *š*, but activating Slovenian as the proofing language does support this character. In our experiment, the average accuracy rate for the three packages are as follows:

	ABBYY	*Omnipage*	*CTexTools*
Afrikaans	99.64	99.14	99.10
Sepedi	99.72	96.30	99.52
isiZulu	99.55	95.23	96.81
Tshivenḓa	95.61	95.50	98.30
AVERAGE	**98.63**	**96.54**	**98.43**

Accuracy of OCR scanning is affected by the quality of the source text. Defects such as distorted text lines, skewed images and noise can reduce the recognition quality of a scanned document. *ABBYY FineReader*'s automatic image pre-processing editor can remove some of the defects that may occur in a scanned document. The tools for the correction of defects include (but are not limited to) the following (https://help.abbyy.com/en-us/finereader/15/user_guide/adjustimage/):

- *Recommended pre-processing:* The software will automatically determine and apply the necessary corrections. The corrections that may be applied include noise and blur removal, colour inversion, skew correction, straightening of text lines, corrections of trapezoid distortion and cropping of image borders.
- *Split facing pages*: When scanning a book, a scanned image will usually contain two facing pages. Facing pages are split into two images.
- *Deskew images*: Corrects skewed images.
- *Straighten text lines*: Curved text lines on images are straightened.
- *Correct trapezoid distortion*: Corrects trapezoid distortions and removes image edges not containing any useful information.
- *Rotate and flip*: Images can be flipped vertically or horizontally to get them facing the right direction.
- *Crop*: The software allows the user to select a pre-set scanning area size. By cropping a document, one removes unwanted edges that do not contain any useful information.
- *Invert*: Inverts colour images. The function is useful when dealing with non-standard text colouring, such as light text on a dark background.
- *Resolution*: Changes the resolution of an image.
- *Brightness and contrast*: Changes the brightness and contrast of an image.
- *Levels*: The colour levels of the images can be adjusted by changing the intensity of shadows, light and halftones.
- *Eraser*: Erases a part or parts of images.

- *Remove colour marks*: Removes any colour stamps and marks made in pen to facilitate the OCR of the text hidden by such marks.

The default setting at which texts are scanned is 300 dots per inch (dpi). This is also the dpi recommended by *ABBYY FineReader* (https://help.abbyy.com/en-us/finereader/15/user_guide/scangeneral). In cases where the original material is poor because of age, the dpi may be increased to 600 dpi to enhance the scanning quality. As a default, scanning at 600 dpi does not seem to be feasible, since it increases the file size and the time spent on scanning, with no real improvement of the OCR scanning quality. Once a text has been scanned and the image editor used to correct defects, the scanned text is saved in an image-only PDF format. By saving the scanned document in an image-only PDF format, the document will not be searchable or contain any text layers. A second copy of the (edited) scanned document is put through the OCR function. It must be ensured that the correct language/languages are selected to enhance the OCR quality. For best results, the OCR document is first saved in UTF-8 format (a .txt file). When using ABBYY, the scanned text cannot be directly saved in Word format as the software attempts to replicate the orginal scanned document. One would also encounter scanning errors which would most likely not appear in the .txt file. Scanned documents that have been OCR'ed are then saved in PDF format. For the text cleaning process and for the running of spell checkers, the .txt files are converted to Word format.

The purpose of the cleaning process is mainly to correct scanning errors. Taking the skills level of project participants into consideration (these are mainly student assistants who have an African language as first language), we opted for a text cleaning strategy that does not need skilled computer programmers. We are aware of more sophisticated procedures such as the use of N-grams, but these require a high level of computational skill, which makes these procedures not always ideal within the context of lesser-resourced languages. We therefore rely mostly on a process of manual correction with spellchecker support. Spellcheckers are supplemented with custom dictionaries, based on word lists generated from corpora compiled by the UP digitisation node. After quality control has been carried out, the final version of the cleaned texts is stored in UTF-8 format.

In cases of born digital data, these are usually available either in .pdf or MSWord (.doc or .docx) format. In case of PDF documents, OCR scanning needs to be carried out; for Word documents, these are directly saved in UTF-8 format. Once again, it must be ensured that the PDF document should not contain a text layer. An image only PDF document allows other individuals or institutions to utilise it for the purposes of OCR research.

2 Copyright considerations

A salient aspect of text digitisation is that of copyright, especially when working with published texts such as text books, novels, dramas and other literary genres. In order to understand the complexities of copyright on digitised texts, it is important to understand the exact nature of a digitised text. In essence, digitisation is a process of converting printed texts into a machine-readable format. A digitised version involves more than a mere reproduction, as is evident from the procedure described above. As pointed out by Nicholson (2010:10), "it involves the conversion to another format, often involving modification, adaptation, or cropping, even translation, where necessary". Digitisation potentially makes information available and accessible to a wide audience and can therefore be regarded as a form of (re)publishing. Strictly speaking, the act of digitisation therefore constitutes in itself an act of infringement of copyright, unless prior clearance has been obtained from the copyright holder, which in the case of published material, is the publisher. In discussions on copyright the notion of 'fair use' is often referred to, and publishers are more

inclined to provide copyright clearance if they are convinced that digitisation amounts to fair use. 'Fair use' is determined by four factors, i.e. the purpose of the intended use, the nature of the work, the amount or substantiality used, and market impact (Besek 2003: 5; Senekal and Kotzé 2018: 267). With regard to the first factor, the distinction between commercial use and use for research purposes is relevant. Use for commercial purposes is unlikely to be viewed as fair use. Secondly, if a text is of a factual nature, rather than a creative text, the scope of fair use is generally broader. Thirdly, the smaller the portion that is digitised, the more likely it is to be regarded as fair use. It is often argued that scanning a 10% section of a text source is acceptable as fair use. However, copyright experts are quick to point out that even a single page from a text source which represent a core design can be judged as copyright infringement. Determining the effect of making a digital copy available on the potential market for the digitised text or work, constitutes the fourth factor. As Besek (2003: op cit.) points out, use that supplants the market for the original is unlikely to qualify as fair. However, deciding on whether use is 'fair' seems in many cases a subjective decision and needs to be determined on a case by case basis.

In our opinion there are no safe generic copyright rules for text scanning except for explicit permission of the copyright owner. Obtaining copyright can be simplified by negotiating the exact intended use of the data. So, for example, publishers might not agree to the digitisation of full texts as they fear that such data could be resold and consequently will lead to loss of income for the owner. Publishers might be more inclined to give permission to the use of data for research purposes or if the data will only be stored in scrambled format. Texts can, for instance, be scrambled on paragraph or sentence level which simply means that sentences and paragraphs no longer appear in the same order as the source texts. In order to safeguard the person(s) and / or institution(s) responsible for digitisation, a written contract stating the exact sources and the permitted utilization of the texts is the only option.

3 Digitisation of audio material / cassettes

The hardware used for converting audio material is USB Cassette Capture (tape to MP3 converter). Determining the quality of audio cassettes is the first step in the digitisation of audio material. Incorrect storage, deterioration because of age and physical damage to cassettes can all affect the quality of the digitised version. In cases where the quality of the original recording is less than perfect, a decision as to the usefulness of digitisation should be taken, based on the inherent value of the resource.

Audacity (https://www.audacityteam.org/about/), released 20 years ago, is open source software and is regarded as being as effective as many premium paid-for applications (https://www.techradar.com/reviews/audacity). For the purpose of the digitisation of audio cassettes, the software allows the user to digitise recordings from other media, edit the digitised file, i.e. cut, copy, paste and delete and export the digitised file to the desired format. It also has a noise reduction function that can reduce constant background sounds.

Prior to digitisation it must be ascertained whether any analogue noise reduction technique was applied to the tape when it was encoded, and the corresponding decoding filter, either in the analogue or digital domain, must be applied in the digitised copy of the analogue cassette. Examples of noise reduction systems include Dolby B and Dolby C.

The default quality settings at which an audio cassette is being digitised through Audacity, is set at 44100 Hertz (Hz) and 32-bit format in stereo. The final format in which the digitised audio cassettes is stored in is Waveform Audio File Format (WAV). Before the digitised files (.aup files) are stored in said format, the .aup files must be checked to ensure that the default settings were used for digitisation. For any .aup files not digitised according to the default settings, the

process must be repeated. Any static at the beginning and/or end of a digitised file must be removed.

4 Digitisation of video material/ cassettes

The first challenge posed by the digitisation of video material is finding VHS (Video Home Systems) video players needed for playing of video cassettes. Since this is old technology, these players are not readily available. Spare parts, such as drive belts are also only available outside of South Africa. The software used is Elgato video capture (https://www.elgato.com/en/video-capture) and is regarded as one of the best video capture devices (https://www.msn.com/en-us/Lifestyle/rf-buying-guides/best-video-capture-devices-reviews). The resolution at which videos are digitised, is 720x576p (720 pixels across and 576 pixels tall). The recommended video bitrate is 5 Megabits per second (Mbps) and the frame rate 25 frames per second (fps). The recommended audio bitrate is 224 Kilobits per second (Kbps) and the audio sample rate is 44 100 Hertz (Hz). The colour mode is Red, Green and Blue (RGB) colour space. The recording format is Digital Video Disc or Digital Versatile Disc (DVD), the recording video type is Phase Alternating Line (PAL) and the quality is set at best. The digitised version is stored in .mpg (MPEG2) format (codec: MPEG-2 video (mpgv) / codec: MPEG audio layer 2 (mpga)). MPEG, which stands for *Moving Pictures Expert Group,* is a standard audio and video coding compression. The On Screen Display (OSD) messages must be turned off and may not appear in any digitised video. It also needs to be verified that the digitisation process accounts for the analogue Dolby B encoding applied in the VHS recording standard.

5 Provision of metadata

The provision of metadata for any digitised resource is an indispensable part of the digitisation process. Burnard (2004) describes metadata as "the kind of data that is needed to describe a digital resource in sufficient detail and with sufficient accuracy for some agent to determine whether or not that digital resource is of relevance to a particular enquiry". With regard to digitised texts, metadata should ideally be presented in an integrated form, together with the text file, using the same encoding principles or markup language used in the text file itself. According to the Federal Agencies Digital Guidelines Initiative (FAGDI) (https://www.digitizationguidelines.gov/), presenting the metadata in this format facilitates the identification, management, access, use and preservation of a digital resource. It helps to ensure that the text and the metadata are kept together and can be distributed as a single unit. The TEI (Text Encoding Initiative) has been a major influence in this regard, publishing an extensive set of Guidelines for the Encoding of Machine Readable Data (TEI P1) (https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html). One of the recommendations of the TEI was the definition of a specific metadata component, called the *TEI Header*. This header functions as a kind of electronic title page, providing information such as *inter alia* a file description, indication of text derivation and a bibliographic description. Once again, presenting the metadata for each digitised text in this format requires considerable computational expertise and we opted for a simpler, albeit it perhaps old-fashioned approach. We provide mostly standard bibliographic description in a separate document, providing the following information: title, name of author(s), date of publication, ISBN, publisher, genre, description of the genre, language (using ISO language codes), status of copyright, number of pages (PDF document), tokens, media type, encoding, format / file extension and name of document. The name of a text document must contain the following information: language(s), title of document, author's surname and genre, for example *zul_Zibukhipha zibuthela_Shabangu_novel*. In cases where a text document's source is a newspaper or a magazine, the title of such documents in a dataset must contain the following information:

language(s), title of document, publisher and the date, for example *sot_Boleng ba popeho tse fapaneng tsa mebu_Pula Imvula_OCT 2012*. In cases where a document contains more than one language, it should be indicated in the languages field as well as in the name of the document. The ISO abbreviations for the given languages must be written consecutively, without any spaces or delimiters, for example: *nsoengafr_Re bala Sesotho 2_Britz_reader*. The symbol used to delimit the data fields is an underscore (_).

Information included in the metadata lists of audios and videos is: title, presenter, date of publication, publisher, genre, description of genre, language (see table 1 below for ISO language codes to be used), status of copyright, length, media type, encoding, format / file extension and the name of the document. The names of audio and video files should contain the following information: language(s), title of audio/video file, presenter's surname and date, for example *afr_AFR 102 verstegniek onderrigkasset 3 kant A_Marais_19990315*. In cases where a file contains more than one language, it should be indicated in the languages field as well as in the name of the document. The ISO abbreviations for the given languages must be written consecutively, without any spaces or delimiters, for example: *zuleng_Lesson 4_presenter unknown_date unknown*. The symbol used to delimit the data fields is an underscore (_).

6 In conclusion

From the discussion above it should be clear that digitisation of especially textual material is much more than scanning a text and saving it in PDF format. In order to ensure the maximum (re-)usability of data it is extremely important that (a) data are stored in the correct format, and (b) that the correct protocols, procedures and technical guidelines are followed during the digitisation process. Apart from making digitised data available for further HLT processing and application, digitisation has an additional function, i.e. preservation of material that is invaluable and / or irreplaceable, in which case the quality of the data may be of lesser quality. As a general rule, quality of digitised material should only be compromised in the case of text, audio and video when the data is invaluable to such an extent that the user will be willing to tolerate low(er) quality for the sake of the importance of the data. This is for instance the case in very old but valuable data on audio reels or VHS video tapes damaged by moisture. In such cases a notice must be posted warning potential users of lesser quality so that users are informed that bad quality is not the result of substandard digitisation processes or equipment. Potential users can then take an informed decision as to whether they are willing to work through the data.

References

Besek, J.M. 2003. *Copyright issues relevant to the creation of a digital archive: A preliminary assessment body*. Washington, D.C.: Council on Library and Information Resources.

Burnard, L. 2004. *Metadata for corpus work*. https://www.academia.edu/3234836/Metadata_for_corpus_work. Last accessed: 25-08-2022.

Liebetrau, P. (ed.). 2010. *Managing digital collections: A collaborative initiative on the South African Framework*. Pretoria: National Research Foundation.

https://help.abbyy.com/en-us/finereader/15/user_guide/adjustimage/. Last accessed: 25-10-2022. *If your document image has defects and OCR accuracy is low.*

https://www.sadilar.org/index.php/en/. Last accessed: 12-10-2022. *South African Centre for Digital Language Resources.*

https://www.gov.za/ve/speeches/president-cyril-ramaphosa-2020-state-nation-address-13-feb-2020-0000. Last accessed: 25-10-2022. *President Cyril Ramaphosa: 2020 State of the Nation Address.*

Nicholson, D. 2010. *Copyright and related matters*. In Liebetrau (red.) 2010.

Prinsloo, D.J., Taljard, E. and Goosen, M. *Optical Character Recognition and text cleaning in the indigenous*

South African languages. To appear: SpilPlusSenekal, B. A. and Kotzé, E. 2018. *Die ontwikkeling van 'n koste-effektiewe en byderwetse multimedia digitale argief by EPOG in Orania*. LitNET Akademies 15(3), 239 – 275.