11-15-2023

Copyright, Data Mining and Developing Models for South African Natural Language Processing

Chijioke Okorie
chijioke@penguideng.com

COPYRIGHT, DATA MINING AND DEVELOPING MODELS FOR SOUTH AFRICAN NATURAL LANGUAGE PROCESSING

Chijioke Okorie[*]

ABSTRACT

This paper sets out the issues of copyright ownership and risk of copyright infringement liability raised by data science research use of data held by public bodies (in particular, public service broadcasters) in South Africa. Considering both the fair dealing exception in South Africa's Copyright Act of 1978 and the proposed fair use provision in its Copyright Amendment Bill B13F-2017, the paper discusses these issues elaborating on the reasons why data science researchers in public research institutions should not require a copyright licence or be considered to be infringing copyright when they use copyright-protected materials held by public bodies for data science and artificial intelligence or machine learning research (henceforth, data science research). The paper also suggests that even where the outcomes/outputs of data science research are copyright-protected, they should be made available in an open and accessible manner with reasonable safeguards.

[*] PhD (UCT); LLM (Strathclyde), LLB (NAU). Lecturer and Founder/Leader Data Science Law Lab, Department of Private Law, University of Pretoria, South Africa. chijioke.okorie@up.ac.za. ORCID: https://orcid.org/0000-0002-1794-4396. This contribution was supported in part by the PIJIP's Project on the Right to Research in International Copyright funded by Arcadia, a Charitable Fund of Lisbet Rausing and Peter Baldwin. Special thanks to Professors Sean Flynn and Vukosi Marivate for their kind review and very helpful comments. All the views expressed herein, as well as any errors, are solely attributable to the author. Email: chijioke@penguideng.com.

INTRODUCTION

This paper explores the copyright issues raised by data science researchers' access to, and use of copyright-protected data held by public bodies, especially public service broadcasters. It considers, from a South Africa perspective, the question of whether data science researchers in a public research institution,[1] using copyright-protected broadcast news content to train and/or develop natural language processing (NLP) models infringe on copyright in those materials. Given the possibility that technological tools and materials including annotated datasets may be created as a result of licensed or unlicensed use of copyright-protected materials for data science research, this paper also explores whether copyright (or other intellectual property rights) subsist in such materials as to enable the creators exercise proprietary rights thereon. The paper also takes a normative approach to examine whether and how such proprietary rights should be exercised.

Data scientists and researchers seeking to train and/or develop NLP

[1] For purposes of this paper, the definition of institution (i.e., any higher education institution contemplated in the definition of ''higher education institution'' contained in section 1 of the Higher Education Act 101 of 1997 (South Africa); any statutory institution listed in Schedule 1 of the Intellectual Property Rights from Publicly-financed Research and Development Act 51 of 2008 (IPR Act); and any institution identified as such by the Minister under section 3(2) of the IPR Act) under the IPR Act is adopted.

models to learn language tasks (e.g., part-of-speech tagging, named entity recognition, translations, etc.), require access to a significant amount of data in the relevant language(s) in order to do so. While accessing publicly available language data is significantly easier for high resource languages such as English, French, Chinese, and medium resource languages such as Greek, Dutch, Urdu,[2] it is quite difficult to find sufficient publicly accessible data and/or datasets, especially annotated datasets in languages spoken across Africa (generally classified as low resource languages).[3] This is the case for various reasons including colonial legacies of African countries, inequality of language use in business and public settings, the geographical location and language identity of the developers of AI systems and NLP models, etc.[4] South Africa is no exception.[5] In such circumstances, it becomes imperative to find ways to improve the availability of, and access to datasets,[6] and take steps to increase innovation around collection, curation, annotation and classification of datasets in those languages when and where found.[7] In this regard, public bodies (including public service broadcasters)[8] and public funding[9] can play a significant role. As part of the daily functions of public bodies, they create and collect a significant amount of data covering many types of information including language information.[10] The nature and mandate of public service broadcasters across Africa make them a significant resource for datasets in African languages.[11]

The use of public sector data (particularly those from public service broadcasters) in NLP research implicates various legal frameworks such as copyright, privacy and data protection, competition law, contract law, etc. Chief among these is copyright since language data are represented in text, speech and audio-visual format, which by their nature, may be subject of copyright protection. Copyright law grants a bundle of exclusive rights to authors of protectable subject matter such as literary, musical and artistic works, sound recordings, cinematograph films, computer programs, broadcasts, etc.[12] These exclusive rights attaching to these works could differ depending on the work in question but generally include the rights of

[2] Joshi, "The State and Fate of Linguistic Diversity" pp. 6282-6283; Kruit, "Minimalist Entity Disambiguation for Mid-Resource Languages" 300.

[3] Marivate, "Why African natural language processing now?" 133-134. See also, Martinus, "A Focus on Neural Machine Translation for African Languages" 1906; Braun, "Open science in machine learning" 343.

[4] Birhane, "Algorithmic colonization of Africa" 389; Sanneh, "Translating the message: The missionary impact on culture" 42.

[5] Marivate, "Low resource language dataset creation" 2004.

[6] Marivate, "Why African natural language processing now?" 134-137.

[7] Ibid.

[8] Ibid.

[9] Ibid. Ncube, "Effects of the South African IP regime" 282.

[10] Lee, "Licensing open government data" 207; Marcowitz-Bitton, "Commercializing Public Sector Information" 412.

[11] See Marivate "Why African natural language processing now?"137-138.

[12] See section 2(1) of the South African Copyright Act.

reproduction,[13] adaptation,[14] broadcasting or rebroadcasting,[15] transmission in a diffusion service,[16] publishing,[17] etc. The exclusive nature of these rights means that anyone wishing to engage in activities covered by such rights with respect to a given work must obtain permission (i.e., a licence) from the relevant copyright owner of such work to avoid potential liability for infringement. From a copyright perspective, the processes involved in training and developing NLP models implicate the selection and use of copyright works (news articles, books, movies, television and radio broadcasts, etc.)[18] in circumstances involving the exclusive rights of reproduction, adaptation, etc.[19] It follows, therefore, that data science researchers seeking to train and develop NLP models including in African languages may require a licence from the relevant copyright owner(s) of the materials represented in the training data. Copyright owners (even public service broadcasters) could, on the basis of such copyright, refuse access to the data resources or require payment of licensing fees for NLP research activities unless copyright exceptions apply.[20] Alongside these issues of the scope of, and tensions between copyright protection and copyright exceptions are concerns about reusability, as well as endorsement when it comes to use and application of such data as secondary data. These issues are explored in this paper.

To provide a factual context for the discussion of these issues, the analysis in this paper is framed around the experiences shared by a number of data science researchers from South Africa, Kenya and Senegal during an academic conference held at the University of Pretoria, South Africa.[21] The paper zeroes in on a research proposal entitled 'Improving News Categorization, Translation, Named Entity Recognition, and Part-of-Speech Tagging with Natural Language Processing Techniques' (see Annex 1) prepared by the Data Science for Social Impact (DSFSI) research group of

[13] For the full bouquet of rights accorded to different categories of protectable subject matter, see sections 6-11B of the South African Copyright Act. The right of reproduction applies to all categories except sound recording and programme-carrying signals.

[14] Applies to all categories except sound recording, broadcasts, programme-carrying signals, and published editions.

[15] Applies to all categories except programme-carrying signals, and published editions.

[16] Applies to all categories except programme-carrying signals and published editions.

[17] Only applies to literary, musical and artistic works and computer programmes.

[18] See P1 Computational Research: Africa Examples, Right to Research in Africa Conf., Pretoria 23Jan2023: (YouTube 2023) <https://www.youtube.com/watch?v=rZ-3MHcu1oA> accessed 12 September 2023. As indicated in Annex 1, below, the training and development of the NLP models required access to and use of a dataset of news content (i.e., text, audio and video formats including transcripts therefrom) from the SABC News website (collectively, 'SABC News Content').

[19] See de Castilho, "A Legal Perspective on Training Models" 1268.

[20] See also P1 Computational Research: Africa Examples, Right to Research in Africa Conf., Pretoria 23Jan2023: (YouTube 2023) <https://www.youtube.com/watch?v=rZ-3MHcu1oA> accessed 12 September 2023.

[21] See "P1 Computational Research: Africa Examples".

data scientists at the University of Pretoria, South Africa.[22] This involved an attempted access by data science researchers in South Africa to use language data from public service broadcasters for research purposes. This attempt was met with a licence as a condition for access and use, raising questions around the scope of copyright ownership and the application of copyright exceptions, which could obviate licensing requirements.

The issues identified above extend beyond the research proposed by the DSFSI.[23] Many research institutions and scholars in South Africa and across the African continent have expressed frustration with the process and difficulties posed by copyright protection mechanisms in accessing data held by public bodies for purposes of NLP research, AI development and/or machine learning processes.[24] Although the example considered is South African and is shaped by the South African legal context, the issues discussed are of broader relevance across Africa. Recognising the prevalence of these challenges, this paper also makes recommendations for guiding principles around best practices.

The first part of this paper sets out the issues raised by data science research use of public service broadcasters' data using the DSFSI Proposal as an illustration. The second part discusses these issues from the perspective of South Africa's Copyright Act of 1978 as amended, drawing out its implications (particularly of its fair dealing provision) for data science research as contemplated by the DSFSI. In the third part, the paper acknowledges that South Africa is in the process of amending its copyright statute, and its Copyright Amendment Bill B13F-2017 proposes a departure from fair dealing exception to fair use exception. In this regard, the third part explains the boundaries of what data science researchers may do with copyright data within the fair dealing copyright exception (and the fair use exception, should it become law) and also points out areas of uncertainties in the application of the law on copyright exception to data science research.

I. THE INTERFACE BETWEEN DATA SCIENCE RESEARCH AND COPYRIGHT LAW IN AFRICA

Under the Copyright Act, a prospective user of copyright-protected content must, where the use implicates any of the exclusive rights (for example, reproduction, adaptation, broadcasting or rebroadcasting, transmission in a diffusion service, publishing, etc.) granted by copyright law, procure the consent of the relevant copyright owner or risk copyright infringement liability.[25] In essence, the legal authority of a copyright owner

[22] See https://dsfsi.github.io.

[23] For a fuller explanation of the practice of data science, see Marivate, "More than Just a Policy" 155.

[24] Marivate, "Why African natural language processing now?" 126; Hlomani, "Data Regulation in Africa".

[25] Section 23(1) of the Copyright Act.

to grant or decline a licence for the use of its protected material as training data, comes from sections 6 to 10 of the Copyright Act, which grants it the exclusive right to reproduce, adapt, broadcast, etc. its literary, musical and artistic works, cinematograph films, sound recordings, broadcasts, etc.[26] However, for data represented by copyright-protected materials, both general and specific limitations and exceptions have been established that exempt the text and data mining necessary for NLP research from copyright infringement liability.[27] These include provisions excluding certain kinds of ordinarily protectable subject-matter from copyright protection either because copyright has expired in the work or because the law explicitly or implicitly does not extend copyright protection to such works. In other instances, copyright exceptions such as fair dealing may apply to exclude ordinarily infringing activities from infringement liability and obviate the need to obtain a licence from the relevant copyright owners.[28] Evident from these experiences outside the African continent,[29] copyright limitations and exceptions including limitations as to scope of protectable subject-matter and exclusive rights can offer an enhanced access to data in African languages.[30] Exceptions offer the public access to copyright-protected works, which could in turn engender innovation and development. Put differently, copyright limitations and exceptions have the potential to help data science researchers overcome the problem of data paucity (discussed above) and enable them to easily conduct research that will benefit the country and the broader African society.[31] However, the implementation and use of copyright exceptions require an understanding of their scope.[32] Further, the applicability and/or application of copyright exceptions to data science research oftentimes depends on the entity undertaking the research, the nature of the research and the activities involved in the research. These are briefly described in this part.

A. The DSFSI Proposal and the experiences of other African data science researchers

In early 2023, a number of data science researchers from South Africa, Kenya and Senegal were part of a panel convened at an academic conference held at the University of Pretoria, South Africa.[33] These researchers shared their experiences on various research projects involving the collection and use of text, speech and audio-visual data ('language data') about African languages. One project involved the collection of language data on various

[26] Oira, "The dichotomy between signal and content" 414-415, 424-426.

[27] For example, the EU's Articles 3 and 4 of the Directive on Copyright in the Digital Single Market 2019 are dedicated to text and data mining exceptions.

[28] Greenleaf, "'Public rights' in copyright" 112.

[29] Erickson, "Copyright and the value of the public domain" 15-16.

[30] Ncube, "Intellectual property and Fourth Industrial Revolution technologies' 393 – 416.

[31] Although other legal considerations (attribution, source/endorsement, etc.) may still necessitate a licence or at least conditions/terms of use. See "Licensing open government data" 229-231.

[32] Okorie, "Fair use or fair dealing in Africa".

[33] See "P1 Computational Research: Africa Examples".

African languages from an internationally known religious organisation – Jehovah's Witness. Another project (which is the main focus of this paper) sought to collect a range of South African language data from South Africa's sole public service broadcaster, the South African Broadcasting Corporation (SABC). The other project sought to access data from academic publications in Africa.

These instances involved the collection of what is described here as "copyright data" – data which are or could be subject of copyright protection. DSFSI had, for several years before the January, 2023 conference, been seeking authorisation and consent from the SABC to use data in text, audio and video formats including transcripts therefrom) from the SABC News website (collectively, 'SABC News Content') to create annotated datasets and to develop NLP models to classify news content; translate news content from one language to another; perform named entities recognition (NER) on news content, and perform parts of speech (POS) tagging on news content to identify parts of speech.[34] At the conference, a former board member of the SABC, who was a serving board member at the time the leader of the DSFSI research group approached the SABC for authorisation was asked his opinion as to why the SABC did not seem inclined to provide the requested authorisation, and he stated that[35]:

> The SABC is one of the most commercially dependent public broadcasters on the planet. It is currently 80% dependent on commercial revenue. It only gets 3% of its revenue from the state… and in a situation like that with a massive public mandate and very little state support; what could be happening is pressure on the people sitting in that institution when they are approached by someone with genuine intentions like Vukosi start thinking "well, hang on a second, am I passing up the opportunity of potential revenue for SABC down the line? Am I gonna get into trouble if I give a blank cheque?". That's the first thing. And one can understand in a revenue constrained environment that they would be doing their job to see whether there is potential revenue for the SABC down the line. The second issue is on the legal side and I'm not a copyright lawyer and won't express an opinion but they were relying on the Copyright Act, on a section which escapes me now…

Following the January 2023 conference, the DSFSI prepared and shared with the SABC, a research proposal dated 5 April 2023 outlining the activities, methodology and expected outcomes of the NLP research it intended to undertake using the SABC News Content as training data.[36] According to the DSFSI, the SABC expressed willingness to grant a non-exclusive, non-transferable, non-sub-licensable, royalty free licence to the group to use the news content for the research as stipulated in the Proposal

[34] Ibid.
[35] Ibid.
[36] Annex 1.

on the condition that it would retain ownership of all its intellectual property rights to the News Content, any translated content pursuant to the tools created/developed from the NLP research and any annotated datasets from the research, which are deemed capable of commercialisation.

In view of the foregoing assertion, unauthorised use of the copyright-protected data *could* amount to copyright infringement.

B. Copyright subsistence and ownership

As indicated in section 23(1) of the Copyright Act[37] and confirmed by case law, to determine whether infringement has taken place, two conditions must be satisfied: first, it is necessary to establish that in relation to a protectable subject matter, the alleged infringer's behaviour or the behaviour of a person acting through them, falls within the scope of any applicable exclusive right;[38] second, such behaviour must be without the consent of the relevant copyright owner.[39] Infringement, therefore, presupposes unlicensed or unauthorised copyright use i.e., use that falls within the scope of the exclusive rights attaching to the work in question.

The SABC News Content to be deployed as corpora in the training and development of these NLP models are potentially copyright protected. These are materials in text, audio and video formats, which as broadcast contents could qualify as literary works, sound recording and cinematograph film respectively.[40] The Copyright Act requires that for a work to be protected, it must be original and fixed in a material form and must be authored by a 'qualified person' (i.e., by citizens, residents and juristic persons).[41] While originality is not defined in the Copyright Act, relevant case law shows that it is to be assessed on a case-by-case basis and also that the leading standard is a "sweat of the brow" standard requiring some independent thought and intellectual effort by the author.[42] As materials commissioned and/or produced by the SABC and existing on the SABC News website, the SABC

---

[37] According to section 23(1) of the Copyright Act: *Copyright shall be infringed by any person, not being the owner of the copyright, who, without the licence of such owner, does or causes any other person to do, in the Republic, any act which the owner has the exclusive right to do or to authorize.*

[38] *Haupt t/a Soft Copy v Brewers Marketing Intelligence (Pty) Ltd and Others 908 JOC (A)*; *Jacana Education (Pty) Ltd v Frandsen Publishers (Pty) Ltd* 1998 (2) SA 965 (SCA); *Galago Publishers (Pty) Ltd and another v Erasmus* 1989 (1) SA 276 (A); etc.

[39] Ibid.

[40] See s1(1) of the Copyright Act as it defines literary works, cinematograph films and sound recording. Broadcast copyright in South Africa as it is defined and as the rights are structured in section 10 of the Copyright Act is essentially an ensemble of other works. Oira, "The dichotomy between signal and content" 414-415, 424-426.

[41] Sections 2 and 3 of the Copyright Act. Broadcasts and programme-carrying signals are exempted from the requirement of fixation.

[42] See *Haupt supra;* Geyer, "Determining Originality" 176.

News Content are fixed in material form and it is highly likely that they satisfy the requirement of originality and also, authorship by a qualified person.[43]

As explained earlier, the process of training and developing NLP models may result in the creation and annotation of datasets. In terms of the Proposal as embodied in Annex 1 of this paper, the expected outcomes of the data science research are the development of various NLP models for categorisation, translations, NER tags, POS tags; and the release of the derivative data/models under a permissive licence for other researchers to be able to use. These envisaged outcomes raise the question of IP (especially copyright) subsistence (i.e., are the datasets to be considered protectable subject matter?), authorship and ownership of NLP models and labelled datasets.[44]

On the issue of whether the labelled datasets constitute protectable subject matter, it must first be noted that while licensing is usually the go-to mechanism for permitting access to and use of labelled datasets, the datasets may not actually be copyright-protectable materials and hence appropriate objects for licensing.[45] However, where they are eligible for and attract copyright protection,[46] it raises not only the issue of copyright ownership of such materials but the rationale for limitations and exceptions.

The labelled datasets would be in text form and, therefore, could be a collection of literary works (specifically, a database) within the meaning of the Copyright Act. Section 1 of the Copyright Act defines "literary work" to include "tables and compilations, including tables and compilations of data stored or embodied in a computer or a medium used in conjunction with a computer, but shall not include a computer program."[47] As already discussed earlier, the standard for assessing originality which is necessary for copyright subsistence, is said to be the "sweat of the brow" test. This also applies to databases.[48] This standard is admittedly low especially for databases given

---

[43] . Oira, "The dichotomy between signal and content" 427-428.

[44] Neither the algorithms nor datasets are patentable inventions as they are excluded by s25(2) of the Patent Act. They also do not meet the criteria for trade mark protection or other recognised IPRs.

[45] *SOCAN v Bell* para 24.

[46] Even if the fair dealing exception did not apply, section 2(3) of the Copyright Act is clear that the fact that the making of a work infringes on an existing work is not by itself a relevant consideration in the determination of the eligibility of that new work for copyright protection.

[47] See also, *Kalamazoo Division (Pty) Ltd v Gay* 1978 (2) SA 184 (C); *Accesso CC v Allforms (Pty) Ltd* [1998] 677 JOC (T) (case No ii); and *Econostat (Pty) Ltd v Lambrecht* [1983] 89 JOC (W).

[48] Moleya, "Evaluating the copyright protection of databases" 56-79. See also, ***Waylite Diaries CC v First National Bank Ltd*** [1995] 1 All SA 451 (A); *Bosal Africa (Pty) Ltd v Grapnel (Pty) Ltd* 1985 (4) SA 882 (C) [893C]; *Human Sciences Research Council v Dictum*

their informational nature and the conflict between the interests of database developers and those of the public in accessing information contained in databases.[49] But, while there is merit in the argument that the sweat of the brow test is inappropriate for databases and that the protection of databases requires a high creativity-based standard instead, this paper focuses on merely acknowledging the possibility of and the basis for copyright subsistence in the labelled/annotated datasets as databases.

Following from the foregoing, the annotated datasets, if original, may qualify as a work to be protected under copyright law. Where that is the case, the author of such literary work (i.e., the dataset) would be the data science researcher who curated, created and/or labelled the datasets as envisaged by section 1(1) of the Copyright Act. By virtue of section 21(1)(a) of the Copyright Act, the author is the first owner of copyright unless where the exceptions apply. These exceptions relate to works created for publication in a newspaper, magazine or periodical; works created under a commission; works created in the course of one's employment.[50] As such owners, the exclusive rights to reproduce, adapt, broadcast the work, etc. belong to them. DSFSI indicates in the Proposal, its intention to share the annotated datasets publicly under a permissible licence for other researchers to use. The intention to licence presupposes that there are some possible proprietary rights held over such datasets. To the extent that the annotated datasets are protected under copyright law, the licensing of the datasets is within the purview of their rights as copyright owners.

Where the datasets lack originality and cannot be protected by copyright, it may be difficult for the DSFSI (and other data scientists in similar situations) to maintain copyright control over their datasets. However, they may still use licences to relinquish control over their datasets.

C. Copyright protection and NLP research

Machine learning has been useful across many sectors such as fraud detection in the financial sector; health diagnosis in the field of medicine, understanding text for spam detection, answering questions, grouping documents and sentiment analysis, etc.[51] Building and/or developing trained language models to perform these tasks requires a significant amount of data to be used as an input into the machine learning algorithm.[52] For NLP, the training data is usually labelled or unlabelled text. The dominant approach to training and developing NLP models involves the identification and selection

---

Publishers (Pty) Ltd (2003) 804 JOC (T) [809D].

[49] Moleya "Evaluating the copyright protection of databases".
[50] See section 21(1)(b)-(d) of the Copyright Act
[51] Marivate, "Why African natural language processing now?" 128-130.
[52] Ibid.

of the training data (i.e., media in written, audio and video formats). The data then undergoes pre-processing i.e., conversion into a format that can be read by machines or by the NLP tools. The pre-processed data will then be annotated/labelled. Annotation or labelling involves a human or an NLP tool reading the files and assigning appropriate labels to various segments of the data based on pre-defined instructions – statistical data, grammatical rules etc. Thereafter, the NLP model is 'trained'. This involves using a software programme (i.e., the training tool) that applies a machine learning algorithm (a test dataset) to the annotated data to make evaluations and/or analyse the annotated data to extract appropriate characteristics.[53]

It is possible that the activities involved in accessing data, annotating data and generally training and developing NLP models could implicate the exclusive rights of reproduction and adaptation. Section 1(1) of the Copyright Act does not provide an exhaustive definition of reproduction. In *Media24 Books v Oxford University Press*,[54] to "reproduce" within the meaning of the Copyright Act was held to mean "to copy". In *Blind SA v Minister of Trade, Industry and Competition and Others*,[55] the Constitutional Court was called upon to inter alia determine whether the technologies and the activities involved in making accessible format copies of copyright-protected works for persons with visual disabilities amounted to reproduction and/or adaptation. The court noted that while "a content-based distinction between reproduction and adaptation will not always be definitive", adaptation involves some "interpretative engagement with the text so as to render its meaning".[56] Further, the court noted that with making copyright-protected materials accessible to wider audiences using technology, it may sometimes be a question of copying the work into another format and no more,[57] and at other times, it may involve more than mere reproduction and require some translation, transformation that requires interpretation.[58] The court concluded that a comprehensive and appropriate copyright exception would be one that speaks to both the right of reproduction and the right of adaptation.[59]

This paper agrees with the reasoning applied by the Constitutional Court and submits that in the present case, the process of text and data mining, processing and annotating could go beyond reproduction to also involve some interpretative exercise as to include adaptation of the copyright data. However, even where the unauthorised copyright use of a protected work satisfies the infringement criteria indicated above, there would be no

---

[53] Abebe, "Narratives and counternarratives" 329-331; Marivate, "Low resource language dataset creation". See also, de Castilho, "A Legal Perspective on Training Models".

[54] *Media 24 Books (Pty) Ltd v Oxford University Press Southern Africa (Pty) Ltd* [2016] JOL 36649 (SCA) at para 12.

[55] (CCT 320/21) 2023 (2) BCLR 117 (CC).

[56] Paragraph 84.

[57] Paragraph 85.

[58] Paragraph 87.

[59] Paragraph 89.

infringement liability when copyright limitations and exceptions apply.

In the light of the foregoing, a key interpretive issue is whether the research proposed by the DSFSI using copyright data from the SABC fell within the scope of copyright exceptions obviating the need for a licence or whether the circumstances of the proposed use were such as to warrant a licence. It is imperative that data science researchers engaged in developing and training natural language processing (NLP) models have certainty as to the copyright status of language data used for such activities. The answer to the questions whether a data science researcher in a public research institution like the one represented by the DSFSI may be held liable for copyright infringement because they, for research purposes, used copyright-protected broadcast news content to train natural language processing (NLP) models will inform the decisions and practices of data science researchers around the lawful reuse of existing materials in data science and AI development. Also, seeing as the copyright in the broadcast news content in the instant scenario is held by a public service broadcaster, the answer presented in this paper will further determine what approach public bodies should adopt in relation to data generated as part of the discharge of their duties as such public bodies also considering the overarching constitutional rights of access to information and cultural participation.

It is against the context presented above, that data scientists and others making decisions and policies on data use and governance in South Africa and across Africa including those relating to the DSFSI's research as indicated in Annex 1, will undertake their decision-making.[60]

II. FAIR DEALING AND NLP RESEARCH

Sections 12 to 19B of the Copyright Act provides for general and special exceptions from copyright protection where otherwise infringing activities would not be considered infringing. Section 12(1) refers to an exhaustive list of activities ('dealing'), which are to be considered within the parameters of fairness.[61] Fair dealing for the purposes of research or private study, personal or private use, criticism or review, reporting current events in the case of literary or musical or artistic works (section 12(1) and 15(4); broadcasts (section 18); published editions (section 19A) and for the purposes of criticism or review, and/or reporting current events in the case of cinematograph films (section 16(1), sound recordings (section 17) and computer programs (section 19B), would not be infringing. What is required there is that the activities alleged to be infringing are undertaken for any of the purposes referred to in that provision (i.e., private or personal use, research, criticism or review, reporting current events) and undertaken in a

[60] Ibid.
[61] *Moneyweb* 102.

manner considered fair.[62] While that provision does not indicate the meaning to be ascribed to "fair dealing", case law offers guidance.

A. Exempted fair dealing activities

In *Moneyweb v Media24 and another*, a South African high court had the opportunity to consider the scope of the fair dealing exception for the purpose of reporting current events and its applicability in addressing issues of copyright infringement liability. Starting with the relevant fair dealing purpose (i.e., reporting of current events), the court considered that even though a current event need not be one that occurred on the day of the report, such event must be relatively close in time to the report.[63] The court also indicated that the phrase 'reporting of current event' should be given its ordinary, wide meaning.[64] It has also been held that the fair dealing purpose of reporting current event requires an "element of currenthood" to be in the "predominant or material of the work".[65]

In *SABC v Via Vollenhoven* where the respondent sought to bring their showing of a work commissioned by the SABC within the purview of section 12 as a form of criticism or review, the court held that this was 'patently a sham'.[66] In arriving at this conclusion, the court considered the evidence that the respondent had, in an interview with some radio stations, insisted that it intended to get the story contained in the work out to the public. While the court declined to formulate a specific meaning, the court further noted that the purpose of dealing must encompass a "genuine purpose and not a pretext for a purpose which is not contemplated under fair dealing".[67]

In all, when it comes to the interpretation of the specific purpose – private study, personal or private use, criticism or review, reporting current events, research – the courts appear consistent in holding that the context of the work in question, and the facts existing at the time of dealing are important considerations. However, it is to be noted that within South African case law, the only fair dealing purposes that have found some judicial explanations are fair dealing for the purposes of reporting current events and criticism/review. The interpretation to be given to "research" has not been considered judicially

---

[62] *South African Broadcasting Corporation SOC Ltd v Via Vollenhoven and Appollis Independent CC and Others [2016] 4 All SA 623, paras 35-36; Moneyweb para 102. This is similar to the position in Canada. See CCH Canadian Ltd. v. Law Society of Upper Canada [2004] 1 SCR 339, para. 50; Society of Composers, Authors and Music Publishers of Canada v. Bell Canada [2012] 2 SCR 326, para 26 where the court noted: Unlike the American approach of proceeding straight to the fairness assessment, we do not engage in the fairness analysis in Canada until we are satisfied that the dealing is for one of the allowable purposes enumerated in the Copyright Act.*

[63] *Moneyweb,* 123.

[64] Ibid.

[65] *SABC v Via Vollenhoven,* para. 35.

[66] Paragraph 36.

[67] Ibid.

in South Africa. However, the meaning to be ascribed to "research" has been discussed in decisions of courts outside South Africa[68] as well as in scholarly literature.[69] Similar to the stipulation by the South African court in *Moneyweb* regarding the news reporting purpose, the Canadian Supreme Court held in *CCH v Law Society* that "research" must be given a large and liberal interpretation in order to ensure that users' rights are not unduly constrained".[70] In summary, research has been construed as:

- An activity not limited to non-commercial or private contexts.[71]
- Lawyers carrying on the business of law for profit and accessing materials for the purpose of advising clients, giving opinions, arguing cases, preparing briefs and factums.[72]
- Library staff making copies of the requested cases, statutes, excerpts from legal texts and legal commentary.[73]
- Including "many activities that do not demand the establishment of new facts or conclusions" and "can be piecemeal, informal, exploratory, or confirmatory" or be undertaken "for no purpose except personal interest".[74]
- An activity undertaken for the purpose of reaching new conclusions.[75]
- "those processes of study, experiment, conceptualization, theory-testing and validation involved in the generation of new knowledge"[76]
- consumers using music previews for the purpose of identifying which music to purchase.[77]

Applying the guidance above to the circumstances presented by the DSFSI's Proposal and data science generally, it is submitted that the activities

---

[68] See *CCH v. Law Society* supra; *SOCAN v Bell* supra.

[69] Appadurai, "The right to research" 167-177; Okorie, "Government Role"; Von Kries, "Defining commercial and non-commercial research" 60-74.

[70] *CCH v Law Society,* 51. Also, *SOCAN v Bell,* 20.

[71] Ibid. Although the nature of the research (i.e., whether for commercial or non-commercial purpose) is a relevant factor in weighing the fairness of the dealing. See SOCAN v Bell, 36.

[72] Para 51.

[73] According to the Canadian Supreme Court in *CCH v Law Society,* 64: Although the retrieval and photocopying of legal works are not research in and of themselves, they are necessary conditions of research and thus part of the research process. The reproduction of legal works is for the purpose of research in that it is an essential element of the legal research process".

[74] *SOCAN v Bell* supra at paragraph 22. See also, Oriakhogba 2023.

[75] SOCAN v Bell supra at paragraph 22.

[76] United Nations Educational, Scientific and Cultural Organization, Resolution 15 adopted by the General Conference at its 39th session, Annex II, United Nations Educational, Scientific and Cultural Organization's Recommendation on Science and Scientific Researchers (Oct. 30 – Nov 14, 2017), https://en.unesco.org/themes/ethics-science-and-technology/recommendation_science

[77] *SOCAN v Bell,* paragraph 30.

contemplated in the Proposal and involving the use of the SABC News Content qualifies as 'research' and is a purpose permitted as fair dealing under s12 of the Copyright Act.

B. The "fairness" of the dealing

Having established that the purpose is permissible under South African copyright law, it becomes necessary to apply the second ambit of the fair dealing exception: whether the dealing (even for the permitted purposes) is fair.

The Act does not, however, offer guidance as to the parameters with which to assess fairness. With regard to this, the courts in South Africa have, while cautioning against wholesale adoption of foreign jurisprudence on this, indicated the following factors as relevant: the nature of the medium in which the work have been published; whether the original work has already been published; the time lapse between the publication of the two works; the extent of the acknowledgement given to the original work; the nature and purpose of the use; the nature of the copyright work; the amount and sustainability of the use; the effect on the market and the value of the work; etc.[78]

Paraphrasing the application of these factors in *Moneyweb*, even where the dealing was for a purpose indicated in section 12(1) of the Act, such dealing will *not* be fair dealing where publication of the alleged infringing article was made within 1 day of publication of the original article; where the alleged infringer contributed little or nothing to what it had copied from the original article; or where what was copied was likely to be a substitute for the original article even if the original article was acknowledged as the source.[79]

The above position is supported by case law and literature from other jurisdictions outside South Africa that adhere to the fair dealing approach.[80] In this regard, a dealing with musical works for research purposes was found to be fair dealing and permissible because: the real purpose of using the work was for research; there were "reasonable safeguards" in place to ensure that the work is actually used for that purpose;[81] the character of the dealing was such that only single, temporary copies were distributed;[82] the amount or quantity of the work taken as part of the dealing is small when compared against the entire work taken from;[83] the use/dealing was the "most practical, most economical and safest" way to achieve the "ultimate purpose" of the

---

[78] *Moneyweb* paragraph 113; *SABC v Via Vollenhoven* paragraph 36.

[79] *Moneyweb*, 127-131. In *SABC v Via Vollenhoven*, the court did not proceed with the interpretation and/or application of these factors because of its finding that the respondent's dealing did not fall within any of the purposes recognised under section 12 of the Act. See paragraph 36.

[80] Rosati, "Copyright reformed:" 33-54.

[81] *SOCAN v Bell* supra para. 36; *CCH v Law Society* supra para. 66. These resonate with *SABC v Via Vollenhoven* where the court noted that Via Vollenhoven's real motive was to get the work out there.

[82] *SOCAN v Bell* supra paras. 37-38; *CCH v Law Society* supra para. 55.

[83] *SOCAN v Bell* paras 39-43; *CCH v Law Society* supra para. 113.

dealing,[84] the nature of the work was such that it should be widely disseminated (not conflating or mistaking availability with dissemination),[85] and the effect of the dealing on the original work was not adverse or competing.[86] In considering the purpose of the use, the relevant perspective is that of the person using the work and regard must be had to the "real purpose or motive" behind using the work.[87]

The interpretation from case law and literature offers significant guidance as to how to fairly deal with a protectable subject matter. However, there are jurisdictional differences in *what* (or *which*) fairness factors are applied even though significant similarities exist in *how* the identified factors are applied. What is evident from the foregoing is that in South Africa, the factors are interlinked but not applied cumulatively.[88] As held by the court in *Moneyweb*, these factors are not exhaustive and one factor may be more important than the other.[89] In other words, once the purpose of dealing is within the purposes listed in the statute, what is needed to determine whether that purpose as expressed in the dealing is fair is an inexhaustive list of factors that do not necessarily apply cumulatively. The factors that are relevant and applicable therefore depends on the nature of the work and the context of the use.[90] In essence, a cumulative and exhaustive approach to identifying *and* applying the fairness factors is not warranted by case law and would represent a lack of the contextual questioning and iteration necessary for constructing fairness especially in the current technological landscape in South Africa and globally.[91]

C. Fair dealing purposes + fairness

In relation to relevant factors to consider in determining whether the dealing with the SABC News Content for research purposes is fair, it is submitted as follows:

- The nature of the medium in which the works have been published in the case of the SABC News Content is the internet, specifically on the

---

[84] *SOCAN v Bell* paras 44-46; *CCH v Law Society* supra paras. 57,114.

[85] *SOCAN v Bell* para 47.

[86] *SOCAN v Bell* para 48.

[87] *SOCAN v Bell* paras 33-34.

[88] Contrast with the US approach where the copyright statute provides for the four fairness factors to be applied cumulatively. See Ginsburg, "Fair use in the United States" 265-294.

[89] *Moneyweb* para 113.

[90] Shay, "Exclusive rights in news" 594-596. In p.596, noting that while the factors from foreign case law are relevant, this is in addition to some other factors are "*relevant to the particular form of potential fair dealing.* This aligns with the reasoning and approach of the court in *Moneyweb*: According to the court in paragraph 113, "the factors relevant to a consideration of fairness *within the meaning of section 12(1)(c)(i)* include…" (emphasis added).

[91] The proposed fair use provision in South Africa's Copyright Amendment Bill B13F-2017 seems aligned with *Moneyweb* that the factors are neither cumulative nor exhaustive. See Part III, below.

website of the SABC. Given the open nature of the internet and the statutory role of the SABC, there should be no unfairness in accessing and using the works for purposes of NLP research. Given the NLP research context as indicated in the Proposal and also given that the relevant parties [researchers in a public research institution and the PSB] are public institutions whose mission is essentially to serve in the public interests, the factors considering whether the original work has already been published and the time lapse between the publication of the two works are not as relevant. Even if they were, the SABC News Content have been published.

- The consideration of the amount and substantiality of use requires nuancing if it is to be relevant. Because of the nature of NLP research which requires access to the entire work before pre-processing and processing can determine which parts make it into the training data and the NLP model, a consideration of the amount and substantiality of the use in determining fairness of the dealing would be irrelevant and miss the point entirely if the quantity used is the focus without due qualification. As a first step, NLP research would always involve use of a substantial, if not the entire portion of the work. Taking "amount and substantiality" of use into consideration without discountenancing factual, unoriginal elements of the given works, would defeat the purpose of the research fair dealing.[92] Even if this factor were considered relevant in the case of NLP research, it is submitted that this use should be considered fair. In terms of quality, the dealing for the purposes of NLP research does not take the expressive elements of the works in their traditional context. In terms of quantity, this is also the case when the amount taken from *each* work is not significant when viewed against the entirety of the work.[93] As the court in *Moneyweb* noted, one must look at the dealing vis-à-vis each work.[94]

- The purpose and character of the use are relevant considerations. When fair dealing is for the purpose of research, the central issue is whether the inquiry-based nature/purpose of research justifies the use of the original work bearing in mind the intended effect of research. In the instant case, the nature of the use of the SABC News Content is to be considered fair, being a research project as clearly set out in the Proposal and also to promote the use of South African languages which are currently low-resource languages. As discussed above, case law in South Africa and elsewhere shows that the real purpose of the dealing is a relevant factor and the assessment of this factor is to be

[92] Margoni, "A deeper look" 689-690. Some scholars have argued that this would still remain a relevant consideration for any work but the weight to be attached to it would be quite low. For instance, Shay, "Exclusive rights in news" 598.

[93] See *SOCAN v Bell*, 39-41; *CCH v Law Society*, 65.

[94] Paragraph 116. See also, *SOCAN v Bell*, 43.

made from the perspective of the user.[95] In the present case, the purpose of mining and scraping the works from the SABC News website is truly and primarily for research in the setting of a public research institution and to enable other NLP researchers working on South African languages to be able to do so. Furthermore, even though the research is not commercial research or undertaken for commercial purposes, the Proposal indicates that there are reasonable safeguards in the form of a permissible licence to ensure that the resulting datasets will only be used for research purposes.[96]

Regarding the character of the dealing, it is submitted that the fact that the NLP research process involves reproducing the News Content in machine-readable format and then annotating them for machine learning tasks meant that copies made of the original works could not be used for the traditional purpose of music, text and/or video files.[97]

- Regarding the alternatives to the dealing factor, it is submitted that very little weight should be attached thereto in the NLP research context especially NLP research involving African languages. Unlike the news reporting environment where a user could practically apply their own expression and originality to using/reproducing published news report, the NLP research environment requires that the content be accessed in its raw form, labelled and used. Even if more weight is attached to this factor, it is submitted that for NLP research in South African local languages, there are no sufficient and suitable non-copyright equivalent of the work that could have been used. Further, the processing of the content is necessary and crucial to achieve the ultimate purpose. Scraping the data, annotating them and using them to develop and train NLP models is the only practical way that data researchers conduct research in natural language processing.

- Applying the nature of the work factor, which examines whether the work is one which should be widely disseminated, it is submitted that the nature of the work (literary, musical and artistic works, sound recording and cinematograph film and broadcasts in South African languages) is one which should be widely disseminated for the benefit of South Africans. It is indisputable that the dissemination of works in South African languages is desirable and beneficial and while these works are easily available on the SABC News website, it does not mean that such works are accessible and widely disseminated especially in the technology and computational environment. Unless these works are processed in machine readable format and for

---

[95] *SOCAN v Bell,* 33-36*; CCH v Law Society*, 66.

[96] See Annex 1.

[97] Margoni, "A deeper look" 688. In *SOCAN v Bell*, it was accepted that streaming as opposed to downloads meant non-duplication or further dissemination by users. See paras 37-38.

machine learning, the work will not be disseminated for NLP research purposes. Also, given that the works in current form are not easily used in the machine learning environment without processing, NLP activities are of immense benefit to promoting further dissemination.

- The nature of the work factor is linked with the factor that requires a consideration of the effect of the dealing on the potential market for, or value of, the original. In the case of the NLP research as proposed in Annex 1 and for data science research in public research institutions generally, it is submitted that because of the conversion of the News Content into annotated datasets for NLP, it can hardly be said that annotated datasets are in competition with the use and enjoyment of the individual broadcast content itself. And since the effect/outcome of the NLP research is to increase access to and dissemination of South African languages – an outcome within the purview of the SABC's mandate, it cannot be said the datasets have a negative impact on the works.[98]

Also, since the research outcomes (per the Proposal) is to automate the production of news and "help individuals and organisations find news articles that match their interests using a search engine and produce more high-quality content in less time", it cannot be said that the research has a negative impact on the SABC News Content. In the instant case, the works are being used in a specific manner: factual and informational elements are taken, not to cut out the copyright owner's primary market for the work but to enable uses in a setting where the copyright owner does not operate. Moreover, if the SABC were to operate in that setting, it cannot charge the public a fee for such services because its statutory mandate requires it to make those contents available and accessible to the public. Even though a licensing market could exist for the SABC in terms of the SABC News Content, it is not a licensing market that a public service broadcaster such as the SABC should exploit to the detriment of research, transformation and decolonisation.

- On the factor relating to the extent of the acknowledgement given to the original work, it is submitted that the acknowledgement proposed to be given to the SABC News Content will be sufficient and fair as the storage and management of the annotated data sets will indicate the SABC News website as the source of the annotated data sets and a link to the SABC News website will be provided.[99]

In the light of the foregoing, a data science researcher in a public research institution such as the members of the DSFSI research group, using copyright-protected broadcast news content to train NLP models would not be infringing on copyright. The activities contemplated in the Proposal and involving the use of the SABC News Content qualifies as 'research' and

[98] *SOCAN v Bell*, 47-48.
[99] *Moneyweb*, 126.

involve dealing fairly with the News Content. Such use is non-infringing and does not require a licence.

III. SOUTH AFRICAN NLP AND FAIR USE

The focus of the preceding parts of this paper has been on NLP research conducted by data science researchers in the setting of a public research institution. Further, the interpretation has been on the basis of the current copyright statute in South Africa. These two considerations leave various questions unaddressed such as whether data scientists working in environments outside public research institutions can claim the fair dealing exception and whether or not the interpretation proposed in the preceding parts of this paper would change (for better or worse) under the fair use exception proposed in the Copyright Amendment Bill B13F-2017. Clause 15 of the Bill proposes the deletion of the current fair dealing exception and in its place, the insertion of a fair use exception covering an open-ended list of permitted purposes and including an equally open-ended list of factors to be taken into consideration in determining whether a given use is fair

With respect to the proposed fair use exception, the fundamental differences when compared with fair dealing are that it provides a non-exhaustive list of fair use activities[100] along with an equally non-exhaustive list of factors to be taken into consideration in determining whether or not a use is fair.[101] The effect of these is that there would now be statutory basis for considering not just research but a plethora of activities as fair use activities. Further, there would also be statutory basis for taking all relevant factors into consideration in determining fairness. In essence, the proposed fair use provision, by broadening the exempted fair use purposes and providing for a non-exhaustive, non-cumulative list of factors to considered in determining fairness, both retains and strengthens the interpretation proffered above for the use of copyright data in data science research.

CONCLUSION

Research in African NLP especially as represented by the Proposal of the DSFSI research group in Annex 1 requires data science researchers (and any person assessing data science research use of copyright works) to determine the scope of the fair dealing exceptions and how they apply to the use of copyright-protected works in the context of emerging technologies. The particular context of South Africa and African NLP also present an opportunity to reflect on how public institutions including PSBs should

[100] According to the opening part of the proposed section 12A(a), "*[i]n addition to uses specifically authorized, fair use in respect of a work or the performance of that work, for purposes such as the following, does not infringe copyright in that work…".*

[101] According to the opening part of the proposed section 12A(b), "[*i]n determining whether an act done in relation to a work constitutes fair use, all relevant factors shall be taken into account, including but not limited to…".* See also Okorie, "Fair use or fair dealing in Africa".

discharge their statutory mandates in the copyright environment.

Insofar as the fair dealing exception is concerned, case law in South Africa, while limited, aligns with foreign decisions particularly those on fair dealing in Canada and the UK where the specific dealing must first be one of the permitted purposes and such dealing must be fair. In this regard, it is required to consider the perspective of the user and the context of the use/dealing. The considerations for determining the fairness of the dealing are not fixed and vary depending on the nature of the work. The upcoming Copyright Amendment Bill also upholds this position.

With specific regard to the question of whether data science research in the context of public research institutions and involving use of copyright data held by public bodies, is infringing, this should be answered in the negative: a data science researcher in a public research institution such as the members of the DSFSI research group, using copyright-protected broadcast news content to train NLP models not be infringing on copyright as the activities contemplated in the research and involving the use of copyright-protected content, qualify as 'research' and involve dealing fairly with the content. Such use is non-infringing and does not require a licence. Moreover, where copyright subsists in the outcomes of data science research, the relevant copyright owner (barring contractual overrides) is the person(s) who created or authored those outcomes.

To conclude, in terms of South African law, it would be fair dealing (or fair use when the Copyright Amendment Bill becomes law) for data scientists in public research institutions to use copyright data as training data for NLP research. Based on the discussion in this paper, such data scientists should also consider the following aspects when using copyright-protected materials as training data for African NLP and data science research generally especially where copyright in those materials are held by a public body:

- the need to indicate and acknowledge the source of the training data;
- the need to safeguard the resulting annotated datasets and trained models for research purposes;
- the need to inform, clarify and/or caution users of their research outcomes and outputs as to the extent of accuracy and the limits of possible uses/reuses of the underlying data.

Annex 1

References

National Statutes

Copyright Act 98 of 1978.

Copyright Amendment Bill [B13B 2017].

Higher Education Act 101 of 1997.

Intellectual Property Rights from Publicly-financed Research and Development Act 51 of 2008.

Case Law

Accesso CC v Allforms (Pty) Ltd [1998] 677 JOC (T) (case No ii).

Blind SA v Minister of Trade, Industry and Competition and Others (CCT 320/21) 2023 (2) BCLR 117 (CC).

Bosal Africa (Pty) Ltd v Grapnel (Pty) Ltd 1985 (4) SA 882 (C) [893C].

CCH Canadian Ltd. v. Law Society of Upper Canada [2004] 1 SCR 339.

Econostat (Pty) Ltd v Lambrecht [1983] 89 JOC (W).

Galago Publishers (Pty) Ltd and another v Erasmus 1989 (1) SA 276 (A).

Haupt t/a Soft Copy v Brewers Marketing Intelligence (Pty) Ltd and Others 908 JOC (A).

Human Sciences Research Council v Dictum Publishers (Pty) Ltd (2003) 804 JOC (T) [809D].

Jacana Education (Pty) Ltd v Frandsen Publishers (Pty) Ltd 1998 (2) SA 965 (SCA).

Kalamazoo Division (Pty) Ltd v Gay 1978 (2) SA 184 (C).

Moneyweb (Pty) Limited v Media 24 Limited and Another [2016] 3 All SA 193.

Society of Composers, Authors and Music Publishers of Canada v. Bell Canada [2012] 2 SCR 326.

South African Broadcasting Corporation SOC Ltd v Via Vollenhoven and Appollis Independent CC and Others [2016] 4 All SA 623.

Waylite Diaries CC v First National Bank Ltd [1995] 1 All SA 451 (A).

Treaties, Resolutions, Declarations, International Official Reports, etc.

European Union's Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information stipulates that the Directive does not apply to "documents held by public service broadcasters and their subsidiaries, and by other bodies

or their subsidiaries for the fulfilment of a public service broadcasting remit".

United Nations Educational, Scientific and Cultural Organization, Resolution 15 adopted by the General Conference at its 39th session, Annex II, United Nations Educational, Scientific and Cultural Organization's Recommendation on Science and Scientific Researchers (Oct. 30 – Nov 14, 2017), https://en.unesco.org/themes/ethics-science-and-technology/recommendation_science

Books

OH Dean: Handbook of South African Copyright Law (Juta) p1-95.

Oriakhogba, Desmond. *The Right to Research in Africa: Exploring the Copyright and Human Rights Interface*. Springer Nature, 2023.

Von Kries, Caroline, and Gerd Winter. "Defining commercial and non-commercial research and development under the Nagoya protocol and in other contexts." In *Research and Development on Genetic Resources*, pp. 60-74. Routledge, 2015.

Journal Articles, Book chapters, Internet sources, etc

Abebe, Rediet, Kehinde Aruleba, Abeba Birhane, Sara Kingsley, George Obaido, Sekou L. Remy, and Swathi Sadagopan. "Narratives and counternarratives on data sharing in Africa." In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp. 329-341. 2021

Appadurai, Arjun. "The right to research." *Globalisation, societies and education* 4, no. 2 (2006): 167-177.

Birhane, Abeba. "Algorithmic colonization of Africa." *SCRIPTed* 17 (2020): 389; Marivate 2023 (n7) p. 165.

Caroline Ncube and Isaac Rutenberg 'Intellectual property and Fourth Industrial Revolution technologies' in Z Mazibuko-Makena and E Kraemer-Mbula, E. (ed.s) *Leap 4.0: African Perspectives on the Fourth Industrial Revolution* (2020) MISTRA 393 – 416.

de Castilho, RE, Dore G, Margoni T, Labropoulou P, and Gurevych I, 'A Legal Perspective on Training Models for Natural Language Processing'. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA); Miyazaki, Japan, 2018 de

Erickson, Kris, Paul J. Heald, Fabian Homberg, Martin Kretschmer, and Dinusha Mendis. "Copyright and the value of the public domain: An empirical assessment." *Intellectual Property Office Research Paper* (2015): 15-16.

Geyer, Sunelle. "Determining Originality in South African Copyright

Law: Is It" or"," and", or Something" More"?." *THRHR* 85 (2022): 176.

Ginsburg, Jane C. "Fair use in the United States: Transformed, deformed, reformed?." *Singapore Journal of Legal Studies* Mar 2020 (2020): 265-294.

Greenleaf, Graham, and Catherine Bond. "'Public rights' in copyright: What makes up Australia's public domain?." *Copyright: What Makes Up Australia's Public Domain* (2013): 111-138.

Hlomani, Hanani, and Caroline B. Ncube. "Data Regulation in Africa: Free Flow of Data, Open Data Regimes and Cyber Security." (2023). Available at: https://publication.aercafricalibrary.org/server/api/core/bitstreams/22843761-f593-4859-8199-cfa750491c15/content

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. "The State and Fate of Linguistic Diversity and Inclusion in the NLP World." In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6282-6293. 2020.

Kruit, Benno. "Minimalist Entity Disambiguation for Mid-Resource Languages." In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pp. 299-306. 2023.

Lee, Jyh-An. "Licensing open government data." *Hastings Bus. LJ* 13 (2016): 207.

Marcowitz-Bitton, Miriam. "Commercializing Public Sector Information." *J. Pat. & Trademark Off. Soc'y* 97 (2015): 412.

Margoni, Thomas, and Martin Kretschmer. "A deeper look into the EU text and data mining exceptions: harmonisation, data ownership, and the future of technology." *GRUR International* 71, no. 8 (2022): 685-701

Marivate Vukosi. "Why African natural language processing now? A view from South Africa #AfricaNLP." *Leap 4.0. African Perspectives on the Fourth Industrial Revolution: African Perspectives on the Fourth Industrial Revolution*(2021): 126.

Marivate, V., Sefara, T., Chabalala, V., Makhaya, K., Mokgonyane, T., Mokoena, R. and Modupe, A., 'Low resource language dataset creation, curation and classification: Setswana and Sepedi' (2020) *arXiv e-prints*, pp.arXiv-2004.

Marivate, Vukosi, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho Mokgonyane, Rethabile Mokoena, and Abiodun Modupe. "Low resource language dataset creation, curation and classification: Setswana and Sepedi." *arXiv preprint arXiv:2004.13842* (2020).

Marivate, Vukosi. "More than Just a Policy-Day to Day Effects of Data Governance on the Data Scientist." *Data Governance and Policy in Africa* (2023): 155.

Martinus, L. and Abbott, J.Z., 'A Focus on Neural Machine Translation for African Languages' (2019) *arXiv e-prints*, pp.arXiv-1906; Braun, Mikio L., and Cheng Soon Ong. "Open science in machine learning." In *Implementing Reproducible Research*, pp. 343-365. Chapman and Hall/CRC, 2018.

Moleya, Ndivhuwo Ishmel. "Evaluating the copyright protection of databases in South Africa: a comparative analysis with the European Union." *South African Intellectual Property Law Journal* 8, no. 1 (2020): 56-79.

Ncube, Caroline, Lucienne Abrahams, and Titilayo Akinsanmi. "Effects of the South African IP regime on generating value from publicly funded research: An exploratory study of two universities." *Innovation and intellectual property: Collaborative dynamics in Africa* (2014): 282-315.

Oira, Hezekiel, and Lonias Ndlovu. "The dichotomy between signal and content as basis of broadcast copyright: a Kenyan and South African perspective." *Obiter* 39, no. 2 (2018): 399-429.

Okorie, Chijioke. "Government Role in Realising A 'Right' to Research in Africa." (2023).

Okorie, Chijioke. "Fair use or fair dealing in Africa: The South African experience" in Bosher, Hayleigh, and Eleonora Rosati, eds. *Developments and Directions in Intellectual Property Law: 20 Years of the IPKat*. Oxford University Press, 2023.

Open Government Partnership: https://www.opengovpartnership.org.

P1 Computational Research: Africa Examples, Right to Research in Africa Conf., Pretoria 23Jan2023: (YouTube 2023) <https://www.youtube.com/watch?v=rZ-3MHcu1oA> accessed 12 September 2023. As indicated in Annex 1, below, the training and development of the NLP models required access to and use of a dataset of news content (i.e., text, audio and video formats including transcripts therefrom) from the SABC News website (collectively, 'SABC News Content').

Rosati, Eleonora. "Copyright reformed: the narrative of flexibility and its pitfalls in policy and legislative initiatives (2011–2021)." *Asia Pacific Law Review* 31, no. 1 (2023): 33-54.

Sanneh, Lamin. *Translating the message: The missionary impact on culture*. No. 42. Orbis Books, 2015.

Schonwetter, Tobias, and Caroline Ncube. "New hope for Africa? Copyright and access to knowledge in the digital age." *info* 13, no. 3 (2011): 64-74.

Shay, R. M. "Exclusive rights in news and the application of fair dealing." *SA Mercantile Law Journal* 26, no. 3 (2014): 587-605.

South African Centre for Digital Language Resources website: https://www.sadilar.org/.

Tisani, N. "African indegenous knowledge systems (AIKSs): another challenge for curriculum development in higher education?: perspectives on higher education." *South African Journal of Higher Education* 18, no. 3 (2004): 174-184.

Van Dijk, Machiel, Richard Nahuis, and Daniel Waagmeester. "Does public service broadcasting serve the public? The future of television in the changing media landscape." *De Economist*154 (2006): 251-276.