

IsiXhosa Intellectual Traditions Digital Archive: Digitizing isiXhosa texts from 1870-1914

Jonathan Schoots

LEAP
Stellenbosch
University
jschoots@
sun.ac.za

Amandla Ngwendu

Department of African
languages and Literature
University of
Cape Town
amandla.ngwendu@
uct.ac.za

Jacques de Wet

Department of
Sociology
University of
Cape Town
jacques.dewet@
uct.ac.za

Sanjin Muftic

Digital
Library Services
University of
Cape Town
sanjin.muftic@
uct.ac.za

Abstract

This article offers an overview of the IsiXhosa Intellectual Traditions Digital Archive, which hosts digitized texts and images of early isiXhosa newspapers and books from 1870-1914. The archive offers new opportunities for a range of research across multiple fields, and responds to debates around the importance of African intellectual traditions and their indigenous language sources in generating African social sciences which is contextually relevant. We outline the content and context of these materials and offer qualitative and quantitative details with the aim of providing an overview for interested scholars and a reference for those using the archive.

1 Introduction

The IsiXhosa Intellectual Traditions Archive (IsiXIT)¹ is a collection of isiXhosa newspapers and books from the turn of the 20th Century, accessible online at <https://ibali.uct.ac.za/s/isixit/page/welcome>. This archive collects isiXhosa newspapers produced in the colonial period, covering 1870-1894 (with 1860-69 and 1895-1912 to be added in the future), as well as pioneering isiXhosa literature published in the early part of the 20th century. The archive allows users to browse images and to download MS Word text files² of early isiXhosa newspapers and books and collects metadata about each publication and its content.

There has been a great demand for deeper knowledge of African intellectual traditions, analyses, and histories. However, many key materials written in indigenous languages by African intellectuals have been inaccessible—available only as hard

copies carefully preserved in archives, or as expensive and limited subscription services.

The IsiXhosa Intellectual Traditions Archive is in the process of creating a substantial digital and textual database of these early isiXhosa texts, in order to make this data freely available to researchers and the general public. Our aim is to both increase access to these important historical materials, and to advance research by making source materials ‘research ready’ and removing barriers to access, thereby allowing a wide community of researchers access.

The archive currently includes the newspaper *Isigidimi sama-Xosa*³ (published from 1870 to 1888) and *Imvo Zabantsundu*⁴ (published from 1884 onward) as well as the books *Zemk’inkomo Magwalandini*⁵ first published in 1906 (Rubusana, 1911) and *Ityala Lamawele*⁶ first published in 1914 (Mqhayi, 1931). Other sources, including newspapers such as *Indaba* (1862-1865), and *Izwi Labantu* (1897-1909), as well as standalone works by early isiXhosa writers will be incorporated into the project over time. The entire archive can be easily browsed on our website (see above). We also make converted text files of these collections accessible for download using permanent DOI links (see footnotes above). These permanent links allow a stable way to access the files which will be maintained in perpetuity by UCT. Researchers can use these permanent DOIs to access and download the text files of the collection, and to view version changes over time. Researchers can also cite these collections to offer readers a permanent link to the data, thus supporting academic rigour.

¹In addition to the authors, the archive has been created with the support of the following student research assistants: Zimingtonaphakade Sigenu, Siphenkosi Hlangu, Sipile Nqiyama, Philisa Plamana, Sinovuyo Xhonga and Likhona Qazisa.

²We intend to expand download formats to include PDF and TXT formats in the future.

³Isigidimi sama-Xosa collection DOI:
<https://doi.org/10.25375/uct.22332271>

⁴Imvo Zabantsundu collection DOI:
<https://doi.org/10.25375/uct.22332268>

⁵Zemk’inkomo Magwalandini collection DOI:
<https://doi.org/10.25375/uct.22332286>

⁶Ityala Lamawele collection DOI:
<https://doi.org/10.25375/uct.22332277>

The newspapers and books in this collection represent the earliest written record of sustained intellectual and political debates by black South Africans. Produced in the eastern part of the Cape Colony (today's Eastern Cape), these texts capture the intellectual work of an emerging group of thinkers, activists, political leaders, and their wider public, as they developed new knowledge, oriented to understanding and engaging with colonialism and their changing social world from an African perspective. The isiXhosa newspapers were a central source of social, and increasingly political, reporting. They also make visible debates and dialogues of a wider public sphere through the printing of letters sent in from literate Africans from across present day South Africa, providing key insights on the development of African social and political discussion and theorizing. The archive also includes some of the earliest and most central isiXhosa book publications, published by leading figures of the African intellectual and political movement of the time. These publications offer one of the most systematic and complete records which reveal the development of African intellectual traditions in the face of colonialism.

2 IsiXhosa newspapers and books in historical context

The early isiXhosa newspaper press was a forerunner for African language newspaper publication on the continent. The first isiXhosa newspapers were published from 1837, and isiXhosa papers again took the lead in development as the first papers to be headed by African editors (Cagé and Rueda 2016; Switzer and Switzer 1979, see also Masilela 2009; McCracken 2015; Gilmour 2007). *Isigidimi sama-Xosa* had the first independent African editor, Elijah Makiwane, from 1876. The first black owned and run paper, *Imvo Zabantsundu*, appeared in 1884. These papers represent the first sustained platforms for African led journalism where an emerging class of isiXhosa speaking political leaders, activists, and intellectuals could share opinions, debate contemporary issues, and coordinate social and political activity. That they chose to write for a literate isiXhosa-speaking readership⁷ is also significant. Contributors to this newspaper

⁷We refer throughout the paper to an isiXhosa speaking community instead of amaXhosa people because the community was comprised amaXhosa as well as amaMfengu, abaThembu as well as other participants and were united by the shared language format.

community increasingly “focused on conscientising educated Africans in order to mobilise support for social change” (De Wet, 2021). These newspapers published articles and editorials, reports written by correspondents across the eastern part of the Cape Colony, and letters written to the editor which represent the voices of a broad range of the newspaper's readership (Switzer and Switzer, 1979, 40–41, 45–46).

Isigidimi sama-Xosa (trans: The Xhosa Messenger) was founded in 1870 by James Steward of the Glasgow Missionary Society, and was published at Lovedale, the pre-eminent center of African missionary education in southern Africa at the time (Attwell, 2005). First published as the isiXhosa section of a dual language newspaper, it became an independent publication under the editorship of Elijah Makiwane from 1876 (Switzer and Switzer, 1979, 45–46). For 14 years, from its founding in 1870 until the founding of *Imvo Zabantsundu* in 1884, it was the leading site for African intellectuals of the eastern Cape colony to publish social, religious, and historical writings. It created a space of debate and analysis on a host of social issues by publishing letters written in from a wider community including some contributors from across present day South Africa (Odendaal, 2013). *Imvo Zabantsundu* (trans: Native Opinion) was the first African owned and run newspaper in South Africa. It was founded by John Tengo Jabavu in King William's Town with its first edition published in November 1884. *Imvo Zabantsundu* offered the first platform for African journalism, political and social commentary, and opinions which were free from any missionary control or censorship, and played a pivotal role in explicitly foregrounding African political interests. *Imvo* was primarily published in isiXhosa, but also included a translation of the editorial into English and at times published other English articles, letters, and advertisements (Switzer and Switzer 1979, 40–41, Moropa 2010; Mkhize 2018).

The IsiXIT collection also includes some of the earliest publications of isiXhosa literature - texts which emerged from leading intellectuals of the same community as newspaper contributors. These texts have gone on to be cultural touchstones of isiXhosa literature (Jordan, 1973). The books in this collection currently include *Zemk'inkomo Magwalandini*, a collection of praise poems and other writings, edited by Mpilo Walter Benson

Rubusana one of the most important cultural and political leaders of the early 20th century (Jordan, 1984). The collection also includes Ityala Lamawele, the celebrated novel of Samuel Edward Krune Mqhayi, who is perhaps most remembered as the preeminent Xhosa poet and “African Shakespeare” (Mqhayi, 2009). The archive aims to host these and other writings of African intellectuals and social commentators whose writings significantly influenced African ideological and intellectual projects of the period.

3 Importance and possibilities of isiXhosa language historical archives

The immediate and practical goal of the IsiXIT archive is to make early isiXhosa texts authored by African intellectuals in the late 1800s and early 1900s available and “research-ready” for contemporary study. To the best of our knowledge there are no other open access digital archives for isiXhosa texts from the late 1800s and early 1900s that make OCRred texts available to users.⁸ However, the added value of this archive lies in its contribution to emerging African social sciences. Inspired by the isiXhosa writings of the African intellectual S.E.K. Mqhayi, and by the work of the social scientist Neville Alexander on multilingualism and decolonising academia, we realised that if we are going to decolonise social sciences and contribute meaningfully to the making of an African sociology then we need to engage with these and other historical vernacular texts. This work requires grappling with the socially constructed meanings of African sociological concepts as they are conveyed through these texts and in relation to historical contexts.

Tracing socially constructed meanings of key isiXhosa sociological concepts at critical moments in the history of an indigenous sociolinguistic grouping is about mapping what Raymond Williams (1981:109) in his famous book “Keywords: A Vocabulary of Culture and Society” refers to as the “modulations of the word through history”, which help us understand the term itself in relation to changing social conditions and lived experiences. We cannot generate African sociological theory without African concepts, because

⁸There is a growing effort by libraries, universities, and private companies to make digital scans of African newspapers available through paid (or limited but growing free) services. This project advances what is available by making OCRred texts available, for free, with an interactive website that facilitates rich navigation, searching and browsing.

concepts are the building blocks of theory. This scholarship, which relies on texts in indigenous languages, is grounded in African ontologies and what Mafeje (2000) calls endogeneity.

Up until recently, the challenge has been accessing and collecting these early isiXhosa writings and then converting image files of newspapers found in various library archives into a text format that is more easily researchable. The value of the IsiXIT digital archive gained considerable urgency after the devastating fire at the University of Cape Town in 2021 that damaged multiple buildings and destroyed many primary African collections, which were housed in the African Studies library.

IsiXhosa speakers have also alerted us to another advantage of drawing on the IsiXIT archive to explain the meanings of African sociological concepts and incorporate African languages in the evolution of social sciences on the continent (and elsewhere). It is a matter of identity affirmation. One of the authors of this paper recalls how one of his isiXhosa-speaking students articulated the ability of the archive and associated research to change her view of Sociology in relation to identity. She said that when she first studied Sociology she could not find herself in the discipline; she saw it as something foreign. In time and after working on the IsiXIT Digital Archive and reading the research outputs, her view changed: The archive helped her locate herself in the discipline and began to see Sociology as “something that is ours”. This insightful observation reminds us of the value of the IsiXIT Digital Archive as a source for decolonising our curricula. When Sociology and other social sciences are no longer foreign, and students can identify with them, and academia as a whole, then we are likely to see new contextually relevant innovative research that pushes the frontiers of knowledge both locally and globally.

To-date, numerous studies and research outputs have drawn on the textual data from the IsiXIT archive. These include, for example, studies on the socially constructed meanings of “Imfundo” (De Wet, 2021), “Impucuko” (Sigenu, 2021), and “Umsebenzi” (Qazisa, 2022), examinations of IsiXhosa tradition and culture as presented in African newspapers (Majokweni, 2022), and the role of African intellectuals, social networks, and newspaper discourse in shaping the innovation seen in emergence of African nationalism (Schoots, 2021).

4 Description of archival materials

4.1 Qualitative description of the materials

Newspapers: Both newspapers in this collection include a variety of published content, including leading newspaper articles, shorter articles and reports written by a staff of newspaper correspondents, letters written by the general public, and a range of announcements, advertisements, and government proclamations. Figure 1 shows a sample front page of one edition of each newspaper in the collection. To offer a taste of each newspaper's own specific style and ethos we now outline some of the content published in both Imvo Zabantsundu and Isigidimi sama-Xosa.



Figure 1: Front pages of Isigidimi (left) and Imvo (right)

Imvo Zabantsundu published a great number of announcements and advertising containing a wealth of information. Such announcements relayed important messages from local chiefs, alerted its audience to new consumer goods such as medications (traditional and Western) or wedding attire, and announced newly-opening shops, church services, and deaths, among many other topics. Adverts in the paper cover a wide range of places and companies, ranging from shipping or land transportation companies to companies selling clothing or new medications for animals. Announcements too reflect a wide range of topics, from updates about influential and famous people which reveal details and activities of their lives, announcements of open job positions, or the latest telegrams. All these adverts and information sought to help the readers stay informed in their everyday lives. At times these adverts specifically were directed to abantu abantsundu (Black/African people), showing the overall intended audience and sense of collective identity fostered by this newspaper. This advertis-

ing thus offers a window which extends beyond the regular publication of writing and ideas, showing the advertised image of the lifestyle, consumption patterns, and gender norms and expectations which were received by the readers.

Isigidimi sama-Xosa is particularly notable in the way that it tells stories. Through these narratives we learn details of events specific to that period, which reveal the ways people understood what was happening around them. The missionary foundation of Isigidimi is revealed in the frequency of church and Bible stories. Articles cover topics about churches and priests, including what they preached about. They also talk about the different chiefs and what they achieved, as well as discuss the wars of the period and the effects they had on the black people of the time. The paper also offers insights into the Lovedale Missionary school (where the paper was also printed) and offers rich information about the education of the time, including details about the students, such as the number of students of different races that study there. Articles also speak about quotidian topics like farming and the harvesting of crops.

Isigidimi sama-Xosa also includes a range of letters and adverts which give broader insight into debates around important social issues of the day (such as questions of drinking and temperance, or debates around cultural and religious practice), as well as offer reflections on the professions of the letter-writers (notably teachers). Announcements in the paper give information into important meetings and emerging organizations of the time, court cases, the opening of university, and available bursary, among other topics. Adverts reveal what is available at the market and even their prices, and show the sale of land, or medication.

In summary, these papers offer a revealing window into the events of the period. They make visible the ideas of both African intellectuals and a letter writing public in the eastern part of the Cape Colony and richly reveal a picture of daily life of the time through adverts and announcements.

Books: Zemk'inkomo Magwalandini (1906) was written by Mpilo Walter Benson Rubusana. Literally translated the title of this book would be "The Cows have Gone, you Cowards!" In this book the author reprimands isiXhosa-speaking people for allowing their heritage to be lost. The text also argues for a hybrid combination of a re-imagined isiXhosa culture and sense of self that appropriates

Newspaper Title	Date Range	Publication	Pages per edition	Total Editions	Total Pages	Total Words
Isigidimi sama-Xosa	Oct 1870 - Dec 1888	Monthly	8	202	1543	1,619,386
Imvo Zabantsundu	Nov 1884 - Dec 1894	Weekly	4	518	2065	5,473,187
Total				720	3608	7,092,573

Table 1: Details of digitized newspapers

Book Title	Author (Year)	Total Pages	Total Words
Zenk'inkomo	Rubusana (1906)	583	62,568
Magwalandini	Mqhayi (1914)	204	31,228
Total		787	93,796

Table 2: Details of digitized books

(and integrates) aspects of English and isiXhosa culture.

The book is a compilation of works by a number of authors edited by Rubusana. Substantial selections include religious poetry. There are also three famous sections by William Wellington Gqoba; two of which are entitled Ingxoxo Enkulu Yom-Ginwa nom-Kristu (Big Discussion between the “Pagan” and the Christian) and the third is called Ingxoxo Enkulu Ngemfundo (Big Discussion on Education).

Ityala Lamawele (1914) was authored by Samuel Edward Krune Mqhayi and includes fiction, history and poetry. The title means “the lawsuit of the twins”, which is taken from the story narrated in the first half of the book. The plot is inspired by the biblical story about twins in Genesis Chapter 38. Mqhayi uses different aspects of the story to present a picture of the operation of customary law among isiXhosa-speaking people, in particular, its democratic character, as well as their social life during the reign of King Hintsa. The second part of the book is a historical account of the relations between the amaXhosa, the amaMfengu and the colonialists. The remainder of the book focuses on the political manoeuvring by the English in fuelling tensions between various isiXhosa-speaking groupings up to the 1917 sinking of the troopship the SS Mendi (there were 607 black South African troops on board, all of whom perished). The book ends with short biographies of the new leaders of

Paper	IsiXhosa Section	English Section	Xhosa / Eng (%)
Isig	1,600,272	19,114	98.8/1.2
Imvo	3,986,546	1,486,641	72.8/27.2

Table 3: Language composition of newspapers

the “reaction to conquest”.

4.2 Overview by the numbers

The collection represents a significant contribution to making isiXhosa historical text available. Notably, the collection already includes over 7 million words in newspapers alone and an additional almost 100,000 words of book materials. This sizable collection is thus useful for large scale isiXhosa computational text analysis. The collection also has considerable time coverage: the newspaper collection currently spans 25 years from 1870 to 1894, covering an especially important period of the formation of new social, political, religious, and other institutions. The two books also reflect some of the most famous early 20th Century isiXhosa literature. We intend to extend this collection as we continue to grow the archive.

Table 1 and 2 report the key statistics of the archive’s current collection of newspapers and books. Table 3 shows an estimate of the total isiXhosa language section words and English language section words. This is assessed at the paragraph level, and reflects whole blocks of English or isiXhosa text, not the use of single words. This analysis shows that Isigidimi sama-Xosa is published almost entirely in isiXhosa. The approximately 1% of English text comes from the very few letters, advertisements, or government announcements printed in English. Imvo has a far more substantial, although still minor, English language section, on average covering about 27% of the printed content. This included the editorial translated to English as well as other letters, articles, or adverts.

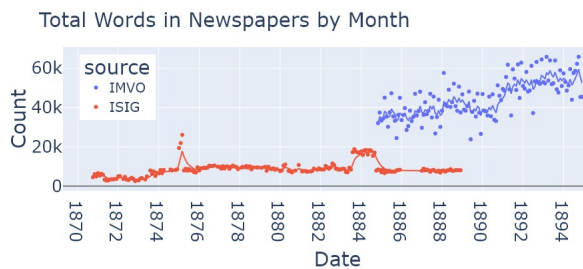


Figure 2: Count of all words in Isigidimi and Imvo

Figure 2 shows the total count of all words (isiXhosa and English) for both papers over the period, grouped by month. Isigidimi is shown in orange and Imvo is shown in blue. The dots represent the count of words, and the line represents a smoothed moving average. While both publications have approximately the same number of words per edition, this monthly grouping shows that the total volume of Imvo was far higher than Isigidimi due to its weekly publication. This also shows that, while Isigidimi published monthly for most of its lifespan, the paper was published twice monthly between July 1883 and September 1884.

Editions missing from the collection: While we have aimed to have as complete an archive as possible, at times some editions are no longer available in hard copy at South African institutions. Most notably, the collection is missing the entirety of the year 1886 editions of Isigidimi, which we have been unable to locate in hardcopy original. A small number of other individual editions are missing due to the inability to access hard copy scans. However, these missing editions are sporadic and limited.

5 Overview of the digitizing process

The goal of our digitizing process has been to create materials which are easily accessible and offer a high quality replica in both image and digital text formats. We start with high quality scanning undertaken by the National Library of South Africa of original hard copy documents, producing a high resolution .tiff image of no less than 400 dpi. From this image we undertake Optical Character Recognition (OCR) using ABBYY FineReader 15. We use ABBYY's built in language model for both English and isiXhosa during this scanning process. These scans then go through a two stage checking, correction, and quality assurance process by our team of isiXhosa language experts. The first stage involves working to correct any character errors in the OCR process. Each page is read over by one of

our team of first language isiXhosa speakers. They pay particular attention to areas where the ABBYY program itself has highlighted uncertainty, but also check the entire text correcting errors. This results in the completion of the first quality assurance stage. The text is then output as a .docx file with the aim of preserving the layout of the original image. A second stage of correction and quality assurance is then undertaken on these documents, ensuring that the layout and formatting of the document is readable and reflects the original, while also seeking to catch any other outstanding problems. With the completion of this second quality assurance stage, the documents are completed. We then undertake an additional audit of around 20% of these final documents to ensure that we are producing high quality output and to catch any systematic errors in the process.⁹ These documents are uploaded to our hosting site with additional metadata created for each newspaper edition. They are also hosted as complete collections for download on our repository hosted on UCT's Zivahub which is accessible via the website and via permanent DOIs (see footnote 3-6). Metadata for each newspaper edition and each book includes the available information on the title, date issued, volume and edition numbers, editor or author, publisher, media type (newspaper or book), extent or number of pages, and available formats of the files. More details on this hosting are outlined in the following section.

6 Showcasing the archive with Omeka S and IIF on Ibali: UCT Digital Collections

In 2021, the Digital Library Services (DLS) department at the University of Cape Town Libraries launched a university-wide showcasing platform for the university's digital collections, here: <https://ibali.uct.ac.za>. The site is called Ibali: UCT Digital Collections (isiXhosa for 'story') and it runs on a set of semantic web technologies called Omeka S and IIF. Ibali is part of the Libraries' drive to nurture an Open Access space where digital collections can be created, curated, published

⁹We currently rely on human correction by isiXhosa experts rather than an automated approach. Because we correct all OCR with human correction we do not calculate the OCR error rate. We do not currently include automated spell checking due to limited tools as the orthography of turn of the 20th century isiXhosa is different from present day isiXhosa. However, we recognize that automatic error detection and statistics would be valuable for the project and hope to explore automated checking and reporting approaches in the future.

and showcased. It is a highly collaborative and flexible, future-thinking online repository space. Since its launch a number of diverse collections have already been showcased on Ibali - including a library of open access resources focusing on climate change, an archive of an active theatrical research project and the IsiXhosa Intellectual Traditions under Colonialism collection.

The main architecture of Ibali is the open source software Omeka S¹⁰. Omeka S is a web publishing platform for GLAMs (Galleries | Libraries | Archives | Museums), designed to create relationships between objects in collections as well as describe them through linked open data resources on the internet. The 'S' in Omeka S stands for 'semantic', as in connecting to the semantic web, where data in web pages is structured and tagged. Its primary focus is on organising elements of a collection such that the links in between items and the greater elements of the internet are strengthened, allowing for much more relevant searches and deeper explorations. It relies heavily on metadata and allows creators of websites great flexibility to set up metadata templates to consistently describe their items. These templates are constructed out of internationally recognized metadata standards (Dublin Core, schema.org) as well as customised ontologies. It makes Omeka S a piece of software that is both heavily customizable, while at the same time being rigorous and interoperable with metadata.¹¹

Creating a collection showcase site such as one for IsiXIT is done in partnership with Digital Library Services. Initially, we engaged in a collection/site interview where key aspects of the collection are identified together with desired characteristics of the site. It was identified that the ability to engage with the newspapers visually while having the ability to download the working files was key to create an interesting collection website. The process of building the site onto Ibali was split into three steps: 1) Prepping the Metadata Entries 2) Packaging Data and Metadata for Upload and 3) Website Design. Steps 1 and 2 would be repeatable to accommodate batches of completed newspaper editions.

The prep for the metadata entries required looking at the appropriate metadata fields to create a template that would align itself with the publication

of the items of the collection (predominantly newspapers). We were able to select metadata properties from the ontology standards of Dublin Core and Schema.org to include entries for vol number, issue number, editor, language, etc. As Omeka S allows for the capturing of items which are not just media, we looked at creating items for the people who were instrumental in the publishing of the newspapers such as John Tengo Jabavu. With Omeka S functioning as a database, we could also work on linking every single issue of a newspaper to its edition, thus allowing users to quickly jump back and forth between a single issue and its edition. This allowed us to organise and keep track of the different newspaper editions, so that the database queries underneath Omeka S could return user queries based on their search. All of this metadata was mapped to a spreadsheet so that the team could populate the entries with new batches of newspaper editions.

Another column within the spreadsheet contained the names of the transcribed WORD files relevant to that issue (which would be preloaded onto the server). Using the CSV Import module of Omeka S, the columns in the spreadsheet would be mapped to the different metadata fields, and each row would be considered a new item. Omeka S then created items for each of the issues, with the metadata linking the word transcript files. Together with the included OCR module, it also looked at the imported Word Documents and extracted the text so that a full text internal search could be made on the site.

Once these items are uploaded, Omeka S automatically creates website pages for each item. The look and feel of these pages is dependent on the look and feel of the site. In Omeka S this is managed through Themes, with many of them being available freely online. They are also heavily customizable with some programming experience in a combination of CSS, HTML and PHP. Within the Omeka S site creation tool it is also possible to create stand alone pages which can highlight certain aspects, or to create landing pages for the collection. The theme can then be customised to present a particular look and feel to the site, and we worked together on finding design choices that would make navigation easy as well as present the newspapers and their media in an interesting way.

A couple of Omeka S modules also enhanced the process, such as the ability to quickly create an index list of items based on shared metadata. With

¹⁰Visit <https://omeka.org/s/> for a tour of Omeka S.

¹¹Omeka S also has a large community developing additional open source modules and themes which add extra functionality to the main software architecture.

this module we created an index tree for each year that would allow users to quickly jump to an issue in a particular month of a particular year. Another important consideration was allowing multiple languages to be presented on the page, as is the case with having website text in IsiXhosa as well as English.

6.1 Facilitating both exhibition and analysis: combining Omeka S exhibition with Figshare repository

Our overall aim to make materials easy to access for both researchers and a general public leads to two goals: 1. Exhibiting and creating user friendly access to the materials and 2. making the underlying data easily and permanently available to the research community.

We use Omeka S for exhibition and broad access. This offers us a location for a public facing exhibition of these materials on a platform that facilitates intuitive and visually appealing exploration as well as metadata presentation and a robust faceting search for processed articles.

We use UCT's digital repository Zivahub, hosted on the Figshare repository system, to make the whole collection easily accessible for download for scholars who will use these materials offline for both quantitative analysis (such as linguists or computational social scientists) and qualitative analysis (such as historians or Media studies scholars). Zivahub offers a permanent record of the available files, including DOIs which will always reference the materials, and a permanently accessible version history. This allows researchers to both access and reference the materials in a format that will always be accessible to the scholarly community. ZivaHub will ensure that our dataset will always be available no matter what versions or formats of the data we share. This will assist the project in making the data FAIR (Findable, Accessible, Interoperable and Reusable).

7 Example of article search with keywords

We are also developing extended search functionality in this archive which is currently offered on the 'Article Search' tab of the site. By using the power of Omeka S and combining it with work done to catalogue, label, and extract information of the articles of the newspaper, we are able to offer a powerful search feature which allows users

to explore the papers in new ways. This search is already available for both Isigidimi and Imvo for the year of 1885¹², and we intend to expand this detailed metadata in the future to encompass more of the collection.

This search functionality allows the user to search for keywords together with a set of five additional labels which capture the type and content of each newspaper item. This includes capturing the 'type' of article (advert, letter, editorials, etc) and sub-types (or topics such as education, law services, politics, or marriages). These two columns were formatted as a controlled vocabulary, meaning that the entries came from a limited pool of options. Additional search labels also include author, location, and language (isiXhosa or English) of the newspaper item when available.

The 'type' and 'subtype' labels were developed through human qualitative analysis and coding of newspaper articles by project researchers and research assistants. The 'type' category was developed to capture different newspaper item type categories such as articles, letters, adverts, announcements, etc. The 'sub-type' category was created to capture the primary topic or content of the item. The controlled vocabulary of each of these sets were created, tested, and revised through a pilot coding of four editions. The two coders both covered all editions and met to define a list together with the supervisor. This was then used to code the full dataset. Regular meetings were held during the full coding process to ensure common coding practices and inter-coder reliability, and to assess if additional labels needed to be added to the sets. Finally, the labels were checked and cleaned by a supervising researcher, using text analysis approaches to remove spelling errors, standardize formats of names and places, and clean and merge subcategories as necessary to create a standardized and computer readable label set.¹³

We use this coding together with the tools of Omeka S to offer a search page which allows researchers to select from these identified labels, thus

¹²1885 was selected as it is the first full year that both Isigidimi and Imvo were published (after Imvo's founding in Nov 1884)

¹³We hope in the future to incorporate automation into this label generation process to more quickly label additional texts. We are exploring using the human coded labels as training data for a supervised machine learning approach to label recommendation. As isiXhosa language NLP tools advance we will also explore Named Entity recognition and topic identification approaches among others, which can further enrich the search functionality.

facilitating the cross search of articles by any combination of article type, subtype, author and location while also searching for any keywords in the text. This then returns links which take the user to the text of the specific articles and links to the display of that newspaper.

To offer an example of the utility of this search: a user might search widely, perhaps for all use of the term ‘*umsebenzi*’ (work), seeing how newspaper contributors use this concept. The user might also use a much more focused search, perhaps looking only at editorials which mention “Sprigg” (the Prime minister of the Cape), or letters which discuss “*Rulumente*” (parliament). A search might combine labels, searching for all announcements and advertisements which come from King Williams Town or Ngqamakwe. Searching by combining language and labels attached to articles offers an exciting opportunity to quickly move through the newspaper collection in a new and non-linear way, and enables easy research and exploration of specific topics as well as content, people, or places. This search capacity is thus an exciting and powerful extension of the archive which supports research. Although generating this newspaper item level information is time consuming, we are currently seeking ways to expand this level of search to all newspapers in the archive.

8 Possibilities for research using the archives

The isiXhosa materials held in this archive have great potential to open up exciting new avenues for research across a number of disciplines, including Sociology, Political Science, History, Literature and Languages, Linguistics, Media studies, Data Science and many more. Possible examples of research are wide ranging. We imagine that these sources may be useful for questions which range across topics such as: discovering specific historical details, following the writings of individuals or small groups, studies of advertising, analysis of the linguistic structure of historical isiXhosa, the development of isiXhosa intellectual concepts and analysis, or building computational tools for the analysis of isiXhosa text, to name just a few examples. Such wide ranging research possibilities will inevitably be undertaken from a diverse range of disciplinary and methodological perspectives.

For this reason we have aimed to produce materials in a format which allows a broad range of en-

agement. For example, historical analysis might prioritise the exact replica of the image to preserve the maximal detail of the source, while computational text analysis approaches might be happy to strip the material to represent only the text of the corpus. We have aimed to host the material in a format which seeks a middle ground by making the digitized and OCRed newspaper pages available for download in a .docx format.¹⁴ This format preserves the layout of the original newspaper page, allowing qualitative readers to use word searching techniques while still experiencing the layout and style of the original paper. For researchers who are interested in only the textual corpus, the text can be extracted from these documents to support corpus linguistics or other computational text analysis methods. We also allow users to view the image scans on the website in a lower resolution JPG format which facilitates a “reading room” experience, allowing researchers browse and peruse the newspapers as they please. However, we do not offer access to image download, as the high resolution original images remain the property of National Libraries South Africa. To maximise access to the materials in the future we will explore making other download formats available including .pdf and .txt versions which might be suitable for different research needs such as preserving the visual architecture of the newspaper (pdf) or accessing the text only corpus (txt).

9 Conclusion

This paper has outlined the scope, content, and context of the materials collected in the IsiXhosa Intellectual Traditions Archive. We intend this to serve as a reference for scholars already utilizing the archive and as an overview for scholars interested in new research possibilities created by the archive. We believe this archive offers resources which may advance research in a wide range of fields, and we hope to continue to expand our coverage in the future to incorporate all major isiXhosa newspapers and books from 1860-1914.

Limitations

This paper outlines an archival collection with the intention of detailing useful information for a range of research fields. However, the focus of the infor-

¹⁴Although MS Word .docx is a proprietary format, most people have access to a version of the software. However, we are looking into using other open formats in the future.

mation has been shaped by the specific perspective created by our own limited disciplinary lens, grounded in the fields of Sociology and African Languages and Literature. For this reason, while we seek to provide adequate information for a broad audience, we may not provide the information or metrics most desirable to researchers in other fields. In addition, the information we provide now is a static snapshot of an evolving archive which we hope will expand. The information provided here will always provide a useful overview, but the specific details will continue to evolve as the archive grows.

Ethics Statement

We provide access to digitized versions of isiXhosa texts which are all in the public domain. We do so for free with the intent to make materials available for both researchers and the general public. It is thus our aim to expand access to historically difficult to access materials, in a language which has faced a lack of research resources. Our hope is that this gives more access to historically marginalized researchers and publics. This project is not subject to ethics review as there are no living subjects discussed in these materials. The authors have acquired permission from quoted participants to use these quotes and conversations.

Human coders and language editors who have worked on this archive are all employed as members of the project. All coders and editors are first language isiXhosa speakers and are either students or researchers.

Acknowledgements

This article was written with partial support from the National Research Foundation of South Africa, grant number 138493.

References

- David Attwell. 2005. The Transculturation of Enlightenment. The Journal of Tiyo Soga. *D Attwell, Rewriting modernity. Studies in black South African literary history (Pietermaritzburg, KwaZulu-Natal University Press, 2005)*, pages 40–41.
- Julia Cagé and Valeria Rueda. 2016. The long-term effects of the printing press in sub-saharan africa. *American Economic Journal: Applied Economics*, 8(3):69–99.
- Jacques P. De Wet. 2021. [Social construction of the meanings of imfundo by African intellectuals in the](#)

[Cape Colony at the turn of the twentieth century.](#) *Canadian Journal of African Studies / Revue canadienne des études africaines*, pages 1–23.

- Rachael Gilmour. 2007. [Missionaries, Colonialism and Language in Nineteenth-Century South Africa.](#) *History Compass*, 5(6):1761–1777.
- Archibald Currie Jordan. 1973. *Towards an African Literature: The Emergence of Literary Form in Xhosa.* University of California Press, Berkeley.
- Pallo Jordan. 1984. Zemk'inkomo magwalandini. The life and times of WB Rubusana (1858-1936). *Sechaba*, 4.
- Archie Mafeje. 2000. Africanity: a combative ontology. *CODESRIA Bulletin*, pages 66–71.
- Anelisa Majokweni. 2022. *We are the voice of our people: IsiXhosa tradition and culture as presented in African newspapers: 1874-1890.* Honours Thesis, Department of History, Stellenbosch University, Stellenbosch.
- Ntongela Masilela. 2009. [The Vernacular Press and African Literature.](#) *Unpublished paper*, pages 1–13.
- Donal P. McCracken. 2015. [The imperial British newspaper, with special reference to South Africa, India and the 'Irish model'.](#) *Critical Arts: A South-North Journal of Cultural & Media Studies*, 29(1):5–25. Publisher: Routledge.
- Khwezi Mkhize. 2018. 'To See Us As We See Ourselves': John Tengo Jabavu and the Politics of the Black Periodical. *Journal of Southern African Studies*, 44(3):413–430.
- Koliswa Moropa. 2010. [African voices in Imvo Zabantsundu : Literary pieces from the past.](#) *South African Journal of African Languages*, 30(2):135–144.
- Samuel E. K. Mqhayi. 2009. *Abantu Besizwe: Historical and biographical writings 1902-1944 SEK Mqhayi.* Wits University Press, Johannesburg.
- Samuel Edward Krune Mqhayi. 1931. *Ityala Lama-wele.* Lovedale Press, Lovedale, South Africa.
- André Odendaal. 2013. *The founders: The origins of the ANC and the struggle for democracy in South Africa.* The University Press of Kentucky, Kentucky.
- Likhona Qazisa. 2022. *Socially Constructed Meanings of Umsebenzi and Associated Terms.* Honours Thesis, Department of Sociology, University of Cape Town, Cape Town.
- Walter Benson Rubusana. 1911. *Zemk'iinkomo Magwalandini.* Selwood Printing Works, Frome.
- Leo Jonathan Schoots. 2021. *Novelty, Networks, and the Rise of African Nationalism: African Intermediary Intelligentsia and the Making of Political Innovation in Colonial South Africa (1860-1890).* Ph.D., The University of Chicago, United States – Illinois.

Zimingtonaphakade Sigenu. 2021. *Socially Constructed Meanings of Impucuko in a Comparative Historical Analysis*. Master's Thesis, Department of Sociology, University of Cape Town, Cape Town.

Les Switzer and Donna Switzer. 1979. *The Black press in South Africa and Lesotho: a descriptive bibliographic guide to African, Coloured, and Indian newspapers, newsletters, and magazines, 1836-1976*. Boston: Hall.