Unpacking the Possibilities of a Vernacular Language Archive

Fagan, Henry
Fagan.henry@gmail.com
McNulty, Grant
grant.mcnulty@uct.ac.za
Hamilton, Carolyn
carolyn.hamilton@uct.ac.za
Suleman, Hussein
hussein@cs.uct.ac.za
Archive and Public Culture Research Initiative

Abstract

The Five Hundred Year Archive is a research project of the Archive and Public Culture Research Initiative based at the University of Cape Town. In an effort to stimulate greater engagement with the deep southern African past, the project has created a corpus of vernacular resources ranging from the earliest available to 1910. It includes productions by an array of African intellectuals in a host of African languages. The vernacular corpus, with its rich metadata, constitutes an extended language and conceptual archive. It is useful to historians, but may also offer research possibilities in other fields, particularly if used in conjunction with contemporary computational methods.

Keywords: African Intellectuals, Vernacular Writing, Orthography, Metadata, Copyright

1. Introduction

The Five Hundred Year Archive (FHYA) [1] is a project of the Archive and Public Culture Research Initiative (APC) [2] at the University of Cape Town (UCT). It seeks to stimulate engagement with the deep past - the neglected eras of the southern African past before the advent of European colonialism. One of the main challenges of conducting research in this area is the severe lack of material, including written sources. What written sources there are have complex histories of production that need to be researched in their own right. While researchers might make use of non-textual material like objects and sound recordings, much of this is misidentified, undated, lost, disorganised or scattered in institutions across the world, making the material difficult to access and use. Working across media and disciplines, with local and international partner institutions, we at the FHYA have created a handful of digital projects that address these challenges.

2. Convening an archive for the deep past

One of these projects is the recently launched online platform, EMANDULO [3], which digitally convenes a growing collection of historical materials relevant to the deep southern African past and makes them available through a single, searchable database. The FHYA selects materials from existing archival or museum collections and galleries, as well as from personal collections. Partner institutions include the Wits University Historical Papers, the Johannesburg Art Gallery, the KwaZulu-Natal Museum, the Amafa/Heritage KwaZulu Natal provincial heritage conservation agency, the Swaziland National Archives, the Killie Campbell Africana Library, the Cambridge Museum of Archaeology and Anthropology, the Austrian Academy of Sciences, the Bews Herbarium at the University of KwaZulu-Natal and the Cambridge University Library. Sometimes individuals or institutions offer digitised or analogue materials to the FHYA. The materials range from hundreds of published texts to recorded oral histories, excavated objects, reports, theses, maps, photographs, images, and researchers' notes. The curation of these materials online offers an opportunity to reorganise and re-position the materials in ways that are designed to challenge the ways in which they have been framed historically by colonial knowledge practices.

The holdings include early productions of the past by African intellectuals who were brokering the history of past generations into the new colonial world. Many are texts written in vernacular languages, which scholars have generally ignored

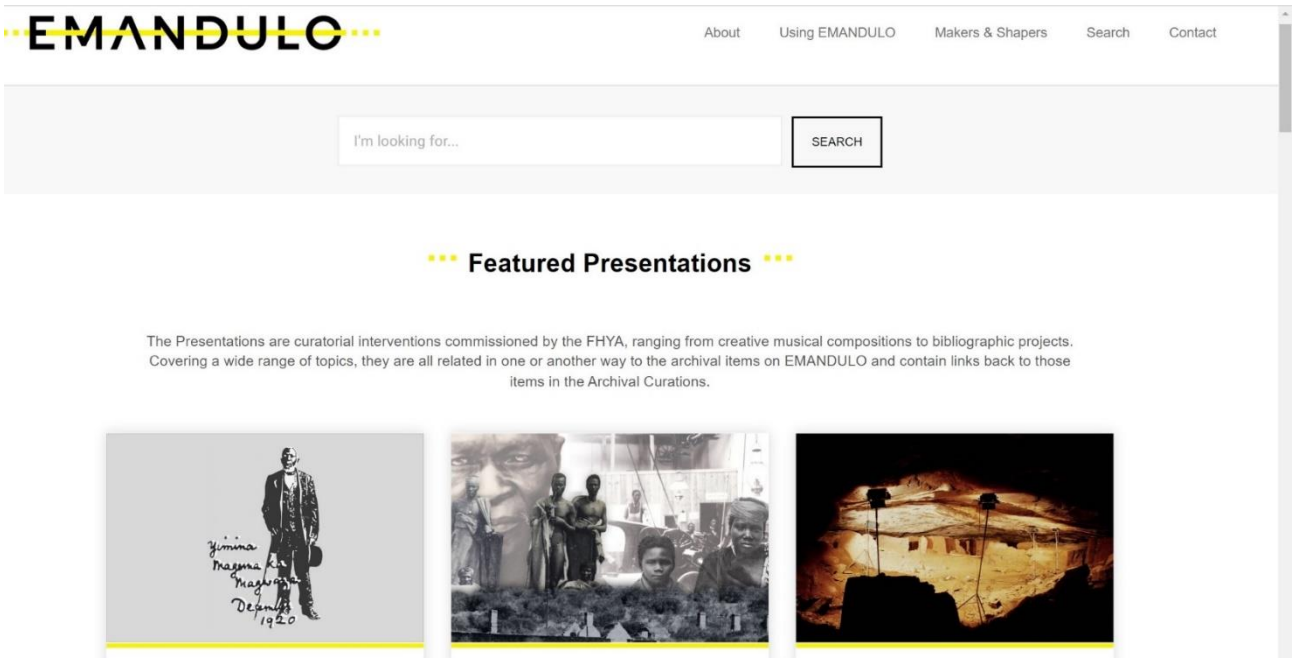


Figure 1: A screenshot of EMANDULO's home page.

as historical sources. On EMANDULO these materials are either presented in themed foci (see for example EMANDULO's collection of materials related to Magema Magwaza Fuze) [4] or are added to the "General Repository" (GR), which includes materials added both by the FHYA team (FHYA Depot) [5] and by registered contributors (Public Depot) [6].

Items in both the GR and the themed sections, like the Fuze archive, have extensive, searchable metadata which tracks, as far as possible, their archival histories - provenance and collection histories - foregrounding the changes to which they have been subjected over time. Provenance deals with the creator or originator of materials before they were lodged in an institutional setting while collection histories refer to the multiple hands that shaped them in each institutional context.

In this paper we focus on the vernacular texts on EMANDULO in both the themed sections and in the GR. We draw attention to the remarkable opportunities that they open up, not only for historians and historical linguists, but as vernacular resources for teachers and the many as yet unknown ways in which they might be used. Indeed, our aim in bringing a paper to this workshop is to use it as a prompt for wider discussion with a spectrum of experts about further areas of possible development and use.

3. Building the Vernacular Corpus

3.1 Selecting Materials

The FHYA selects materials that are pertinent to the five-hundred-year period before European colonisation in what is today the Eastern Cape, KwaZulu-Natal, parts of the Highveld, eSwatini and Lesotho, deliberately extending across modern-day borders that did not exist in the eras before colonialism. The vernacular texts, rendered in a variety of orthographies, occur in early versions of what later became standardised isiZulu, isiXhosa, seSotho, and siSwati. The corpus also contains texts in dialects such as 'siNguni' (Ndwandwe), sePhuthi and isiBhaca. It includes, amongst others, works by well-known writers like John Dube, R.R.R. Dhlomo, Magema Fuze, S.E.K. Mqhayi, Henry Masila Ndawo, and Thomas Mofolo, and many lesser known writers. The texts touch on a broad range of subjects ranging from histories, biographical writing and

historical fiction, to plays, social commentaries, and theology. Importantly, the corpus includes early dictionaries.

We aim to include on the site *all* the vernacular texts we are able to locate from the earliest times to 1910. For the period after 1910 we include texts that for any number of reasons we, or our registered contributors, consider relevant to the study of the deep past in southern Africa.

3.2 Locating Vernacular Materials

The FHYA proactively seeks out texts by combing catalogues and databases (such as UCT Libraries or WorldCat) for keywords between particular dates. This enables us to locate appropriate materials, particularly published books. It also scours open-access repositories such as Google Books or the Internet Archive for texts whose copyright has expired. Suitable online finds are subsequently downloaded and added to EMANDULO. In many instances, the texts are contributed, or pointed to, by scholars who locate them in the course of their research. Before the UCT fire (Davids, 2021) we were able to get physical access to early books in the UCT African Studies Collection, which we then digitised. We sometimes access relevant books in other collections.

3.3 Digitising Materials

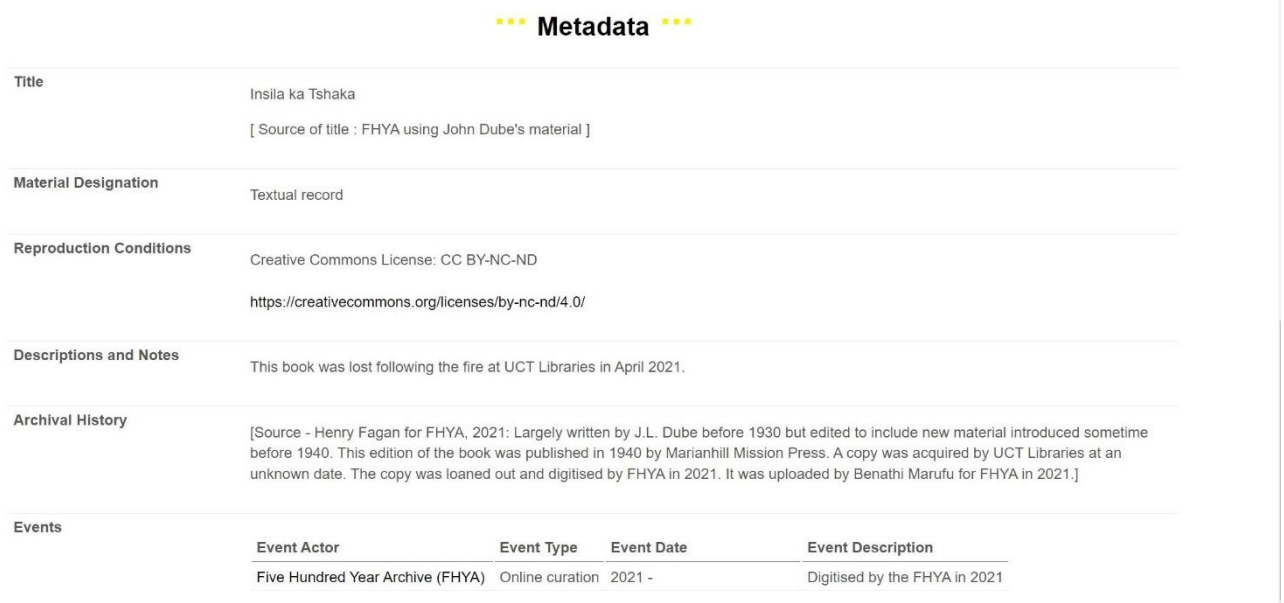
Where necessary the FHYA digitises the materials it acquires and creates PDFs. Most scans are done at a moderate quality and in black and white to ensure their file sizes are not prohibitively large when uploading them to EMANDULO. In rare cases, such as with books where visuals are the focus, the scans are made in colour.

The FHYA applies optical character recognition (OCR) to all texts, making them machine readable and thus searchable on EMANDULO. Although the FHYA's scanner is capable of recognising characters in several languages, vernacular languages (aside from isiXhosa) are not intrinsically recognised by the software. There are also certain things that undermine the accuracy of the OCR. These include materials where the text is faded, those with rare and elaborate old fonts, and outdated or non-standard orthography. OCR problems double as search problems, as EMANDULO's search function cannot correctly index and search the affected text. For items where the text is difficult to discern, the FHYA has begun to introduce typescripts, where full text is made accessible as a related document. In some instances, handwritten texts have typed up summaries produced for one or another purpose in the past. Sometimes transcripts or summaries are contributed by registered users. Consequently, the original items are indirectly searchable.

Some items are old and fragile, which makes scanning them difficult. This may necessitate imprecise scanning (such as slightly skew pages) specifically undertaken to prevent straining the pages or the frames of delicate books. Many materials are already damaged and might have torn or missing pages, or faded text. In such cases, where possible, the FHYA will incorporate the missing page from a better quality copy or a later edition of the production. Such interventions are, of course, carefully noted in the metadata.

3.4 Adding Metadata

Materials sourced by the FHYA for the GR include all known details about each particular item. The FHYA team then researches and adds additional information to the metadata of a particular item so that users of EMANDULO know its archival history – when it was created, by whom and under which conditions, and what happened to it over time. In each case the source of the metadata is provided. Enriched metadata records thus make visible the various 'hands' involved in making and shaping the material over time. This also allows for different copies of the same item (with unique archival histories) to be acknowledged individually. In other words, it is the FHYA's emphasis on metadata that allows it to treat its materials as archival objects.



Metadata

Title	Insila ka Tshaka [Source of title : FHYA using John Dube's material]
Material Designation	Textual record
Reproduction Conditions	Creative Commons License: CC BY-NC-ND https://creativecommons.org/licenses/by-nc-nd/4.0/
Descriptions and Notes	This book was lost following the fire at UCT Libraries in April 2021.
Archival History	[Source - Henry Fagan for FHYA, 2021: Largely written by J.L. Dube before 1930 but edited to include new material introduced sometime before 1940. This edition of the book was published in 1940 by Marianhill Mission Press. A copy was acquired by UCT Libraries at an unknown date. The copy was loaned out and digitised by FHYA in 2021. It was uploaded by Benathi Marufu for FHYA in 2021.]

Events

Event Actor	Event Type	Event Date	Event Description
Five Hundred Year Archive (FHYA)	Online curation	2021 -	Digitised by the FHYA in 2021

Figure 2: A screenshot of the metadata of an individual item.

The FHYA records metadata meticulously and in a consistent style based on the International Council on Archives' General International Standard Archival Description (ICA, 2000). The ICA-Atom metadata format has an emphasis on encoding archival processes related to objects and collections. The Europeana Data Model (EDM), which also focuses on cultural archives, extends that notion to cater for the principles of Linked Open Data, with the intention of supporting distributed interoperable archives. The specific metadata format used in EMANDULO is, however, not crucial, as the key elements can easily be mapped between ICA-Atom and EDM.

At the systems level, Europeana and EMANDULO are driven by very different objectives, with the former designed for costly large scale distributed collection management and the latter for resilient smaller collections in low-resource environments.

It is sometimes necessary to enter metadata that departs from the characteristics of the physical object. For example, colonial-era books with titles that are extremely offensive in a contemporary context are slightly altered (by using inoffensive substitutes) such that their titles are not explicit but still recognisable. Problems with metadata accuracy can arise when items are mislabelled by institutions, or when additional data about the materials cannot be located. In such cases, the FHYA looks to clarify and put down as much useful information as it can. The final step entails uploading material to EMANDULO, where it is accessible to the public. EMANDULO users can browse the items' metadata and search through metadata and OCR-generated texts.

3.5 Copyright

For the General Repository, the FHYA focuses on published texts that are no longer subject to copyright. Copyright for archival materials, like those housed in the Fuze archive, is less clear cut. We have received a legal opinion that if an archival document has been available in a public institution for 50 years or more, it can be treated in the same way as published material. In certain instances, it is necessary for the FHYA to apply to holders of copyright for permission to upload material. Occasionally, institutions may claim copyright over materials to which they may not have a legal right. In cases where the copyright of the material is unknown but the item is conjectured to be in the public domain, the FHYA uploads the item with a short disclaimer outlining that it will be taken down should there be a valid copyright assertion. All materials uploaded to

Another issue arises when institutions produce and then exhibit digital copies of materials but refuse to share those digital copies with the public on the basis that they are a production of the institution. In circumstances such as these, public institutions exert controls that may not be in the public interest. The FHYA works to generate debate and discussion when such cases arise, with a view to shifting policies.

4. A vernacular language and conceptual archive

The FHYA has made a point of locating and uploading to EMANDULO sources in local African languages and allows registered users to do the same. We value contributions from a broad spectrum of users, from professional academics to those with an interest in history and culture. Registered users acknowledge that they either own the copyright of contributed material or that it is out of copyright and that there is nothing racist, sexist or homophobic in what they contribute. All contributions are moderated before being made publicly available. The collected texts form a rich and specifically tailored corpus - one that will grow as more resources are added. Given the FHYA's attention to the production of rich metadata, as much as possible that is known about the material - its dates of production, of publication, etc. - is easily established. This makes it possible to track how vernacular languages were being used at particular points in time and how language use has evolved.

A new generation of scholars working in the APC are paying productive attention to concepts within vernacular writing, and more specifically, concepts in motion across time. As these concepts are infused with particular connotations and meanings, they are not easily translated, but attention to their vernacular usage can bring their historical 'lives' into view. The term 'umbuso' in what is now the KwaZulu-Natal region offers a case in point. The nineteenth century saw dramatic changes in the nature of political power in the region as independent rulers were replaced by colonial governments and the chiefs they appointed. The earliest African writers, using local languages, recorded their understandings of what umbuso was in the eras immediately before their own time. In part this was because they were concerned to engage critically with newly introduced colonial ideas about rule, which the colonial authorities represented as being based on traditional African rule.

Terms like umbuso that were previously used to refer to significantly different forms of rule, were pressed into service in communication about new forms of colonial rule. While the term stayed constant over time, its meaning was changing. The searchability of carefully dated texts in our corpus makes it possible for historians to track the changing meanings of the term over time. What is considered by many today to be a traditional form of rule by chiefs can thus be shown to have a specific, changing history across time. As EMANDULO helps to illustrate within the southern African context, shifts in meaning and orthography take place as political contexts evolve.

Standardised forms of African languages often coalesced around mission stations in particular geographical centres. For example, the Lovedale Missionary Institute situated in what is now the Eastern Cape region became a hub for the production of isiXhosa writing during the early twentieth century (Peires, 1980). Likewise, *Ilanga Lase Natal*, a newspaper founded by John Langalibalele Dube and Nokutela Dube, became a spring-board for early isiZulu writing and isiZulu intellectual discourse (Hughes, 2011). By bringing together materials that show dialectical and orthographic variations, the FHYA corpus lends itself to ongoing work in mapping and periodising these developments. Subject to careful historicisation and contextualisation, the kinds of resources found in the GR are green fields for those interested in the past and in historic changes in orthography and the meaning of terms and concepts, especially when used in conjunction with contemporary computational methods. In

this respect the FHYA corpus of digitised vernacular texts constitutes an extended language and conceptual archive.

5. Early Computational Experiments using FHYA data

The corpus is already proving useful to computer scientists seeking to develop machine-based interventions to advance new research possibilities. As the initial archive was developed, a number of different research problems were identified and explored at the intersection of the digital humanities and computer science.
The text and image collections were used as the basis for experiments with search technology to determine how end users react to multimodal search - where users can enter textual queries or submit pictures, or both - and if specific common search algorithms are applicable to historical South African texts (Singh, 2022). It was determined that some techniques (like stemming) are applicable, but others (like thesaurus use) require specialised development. In addition, users showed a preference for textual search where concepts were abstract but image search where concrete visual representations were possible.

As metadata was being created by the FHYA team, it also became apparent that linking entities (e.g. people, places) where there is orthographic variation is a human-intensive task. Some computational linguists have recently explored the use of automated Named Entity Linking based on machine learning techniques. Experiments were therefore conducted to test the applicability of machine learning techniques to disambiguate vernacular representations of names in FHYA documents and metadata using statistical language models such as BERT and XLM-R (Dunn, 2022). Results from this work have highlighted that techniques that work well for European languages need further refinement to adequately handle the morphological characteristics of Nguni languages.

These early experimental studies demonstrate how the corpus developed by the FHYA is an enabler of research in computational disciplines and establishes a necessary symbiotic relationship between collection development and algorithm development in vernacular languages.

6. Conclusion

Our vernacular language corpus - an extended language and conceptual archive - recognises variability, fluidity, and historical linkages between non-standard forms of language across time. It is proving to be a useful teaching resource and we expect it to facilitate historical research. As the possibilities of our vernacular corpus extend beyond the confines of the discipline of history, we seek insights from other disciplines concerning its potential.

We are especially interested in finding out to what extent, and how, this corpus might be of interest to linguists and language specialists. What might we do to improve its usability for researchers and teachers outside of history? To what extent is the capacity of the corpus to track orthographic changes and the emergence of standardisation of interest, and can we enhance the ability of the system to facilitate this in any way?

Notes

[1] See http://www.apc.uct.ac.za/apc/research/projects/five-hundred-year-archive

[2] See http://www.apc.uct.ac.za/

[3] See http://emandulo.apc.uct.ac.za/

[4] See http://emandulo.apc.uct.ac.za/metadata/Fuze/index.html.

[5] See http://emandulo.apc.uct.ac.za/metadata/FHYA%20Depot/index.html

[6] See http://emandulo.apc.uct.ac.za/metadata/Public%20Depot/index.html

Acknowledgements

This research was made possible with the support of the FHYA team and the APC's broader network of researchers and students. Special thanks to Benathi Marufu for her work with the vernacular materials.

References

Archive and Public Culture Research Initiative 2022, EMANDULO, viewed 22 August 2022, http://emandulo.apc.uct.ac.za/

Davids, N 2021, 'Reflecting on the devastating UCT fire', *University of Cape Town News*, viewed 23 August 2022, https://www.news.uct.ac.za/article/-2021-06-23-reflecting-on-the-devastating-uct-fire

Dunn, JWD 2022, *Evaluating Automated and Hybrid Neural Disambiguation for African Historical Named Entities*, Master's dissertation, University of Cape Town, Cape Town.

Hamilton, C and McNulty G 2022, 'Refiguring the Archive for Eras before Writing: Digital Interventions, Affordances and Research Futures', *History in Africa*, first view, pp. 1-27.

Hughes, H 2011, *First President: A Life of John L. Dube, Founding President of the ANC*, Jacana Media (pty) ltd, Johannesburg.

International Council on Archive 2022, *ISAD(G): General International Standard Archival Description*, viewed 1 September 2022, https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition

Morelli, E 2022, *Exploring uMgungundlovu*, viewed 23 August 2022, https://studio-emandulo.uct.ac.za/fhya-exploring-umgungundlovu/

Peires, J 1980, 'Lovedale Press: Literature for the Bantu Revisited', *English in Africa* 7, no 1, pp. 71-85.

Ramji, H 2022, *Nongqawuse and the Great Xhosa Cattle Killing* , viewed 23 August 2022, https://studio-emandulo.uct.ac.za/nongqawuse-and-the-great-xhosa-cattle-killing/

Singh, SH 2022, *Investigating user experience and bias mitigation of the multi-modal retrieval of historical data*, Master's dissertation, University of Cape Town, Cape Town.