The Letters Column: A Remediation of the *Daily NATION* Newspaper, 1974 – 1978

By

Antony Oduor Owino

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Arts

in

Digital Humanities - Library and Information Studies

Digital Humanities
University of Alberta

Abstract

Digitizing newspapers has a decades-long history, particularly within the library settings that cater to needs of researchers and scholars. However, as scanned images, navigating such archives to locate news articles of interest, mobilize items by topic, authorship or correspondences is often disconnected and tedious without bibliographical indexes. Yet, digitization of this resource has benefited even less from such indexing due to a number of reasons including the volume of collections and nature of its content. The implication of which this resource and its inherent variety of genres remain obscured and underutilized. Here, the reputed letters-to-the-editor column becomes the focal interest onto which content is examined under bibliographical, media and technological perspectives. The end result of this thesis demonstrates the desired navigation strategies for users of digitized newspapers that takes the textual content, topics, authors, and the image itself into consideration. It illustrates interfaces enriched with referential metadata and capable of connecting threads of correspondence within a given collection.

The thesis is interdisciplinary by design and therefore examines the data from a variety of fields. Its historical context and material not only provides arguments that have generated and sustained past digitization projects, but reaffirms the newspaper as an important research resource. Tied to the research goals is automated indexing and how this approach is complicated by the type of content under exploration. To frame this properly, the exploration process for the user has to incorporate elements of the newspaper that made it useful in contemporary times as well as the interconnected interface found in web pages. The process of realizing this uses a number of computational tools and techniques (including computer vision, page layout analysis, and text analysis) to extract letters in the *Mailbox* column which are then indexed using a subject vocabulary.

This research attempts to give full treatment to an obscured text in the expectation that besides preservation, the digitized content is also useful already.

Acknowledgments

I would like to express my gratitude to Dr. Harvey Quamen (Home) and Dr. Ali Shiri (Specialization) for their work, knowledge, and patience as my supervisors. I would like to thank and recognize Dr. Maureen Engel, whose support and guidance from the very beginning have seen me to this moment. I would also like to thank Dr. Deb Verhoeven for chairing the committee. I could not have found a more supportive committee and program. I am profoundly grateful to Virginia Pow for succor in locating and obtaining data for this research.

I am indebted in manifold ways to close family members for their support and encouragement throughout my pursuits of the digital humanities. This accomplishment would not have been possible without them. Last but certainly not least, I wish to thank and acknowledge friends, colleagues, and administration who took an interest in this research project and provided words of encouragement along the way.

Table of Contents

List of Figures & Table

Chapter 1

1 Introduction

1.1 Statement of the problem

The scholarship of newspapers is and should always be a multidisciplinary endeavor. It is in the interests of tracing and interrogating the human experience that important work by historians and philologists as humanities scholars intertwine with the work of sociologists. While engaged in history-oriented exploration and research of digitized newspaper projects, an underlying theme again and again suggested a strong focus on visual representation in collection, preservation, organizing and description work. In other words, the lacking bibliographic records in digitized collections often minimize or ignore specific genres of text significant for analytical and historical interpretations. With a close examination, there exists an often obscured literary trove of reader communication in the “letters to the editor” column. Its utility as a popular and discursive feature of newspapers is rarely emphasized. Historian Stephanie Newell observes that our fascination for the contents and circulation often leaves readers (of those newspapers) out of the frame (Newell, 2011). Yet in the context of early African press, the centering of these letters is significant because their treatment as vibrant literary spaces will illuminate the people, and the popular discourses they produced and sustained. Moreover, the existing collections minimize the contexts, places, identities, and topics attributes relevant in systematic abstraction and retrieval of such texts.

Decades into newspaper archival and digitization projects, individual articulations are not only “illegible” for scholars at macroscopic scale, scholars are starting to locate and organize large collections by particular genres (Underwood, 2014). Ted Underwood, in the quest to understand genres in large collections, laments that our “existing metadata in digital libraries do not provide very strong support for selection by genre” (2014, 4). Beyond assembly of historical literature by genre, newspapers as naturally heterogenous documents pose both epistemological and technical difficulties for scholars. On the one hand, their marginal status as literature has led to a lack of theorization on the localized knowledges that they embody: the letters column in particular. And on the other hand, research on methodologies for reading the indeterminate and fragmented layout that characterizes newspaper pages remains an ongoing project. As other researchers recount, unlike books, magazines and journal papers, newspapers in general lack a standard layout, and contain extreme variations of text characters (Ferilli, Esposito, & Redavid, 2017). To date, there is no known "convention" within the indexing community to organize such letters with a basis on the subjects or authorship. Therefore the task of locating, organizing and describing individual texts in the peripheral genre of *letters* is aimed at attenuating space between the scholar as a user, and the texts; to possibly shift collections from objects of study into subjects of study. A diverse array of disciplines informs my thesis but the scope and goals are stimulated by media and information studies, and the methods digital humanities appropriate. The task of centering correspondences and their subjectivities is a step closer towards full treatment of newspapers as spaces of literary experimentation and articulation.[1]

1.2 Research Questions

As an exploratory study, the purposes are to apply methodological approaches on which the selves and subjects occurring in early newsprint can be distinctly articulated. Using computational tools and techniques, I seek ways to adapt referential and citational practices of the readership in order to refigure them as complete, indexable documents with metadata, and can be described with a form of bibliographic specificity. The significance of such a process has implications on how multidisciplinary scholars locate and situate particular texts within large collections. For humanities scholars, this is the ability to mobilize documents of interest within large newspaper collections and realize the promise of distant reading. The potential outcome for knowledge systems is the expansion of available bibliographical account in digitized collections that enable application of subject headings, indexing, and attribution on heretofore incoherent texts and "shadowy figures" (Newell, 2011, 26;

[1] Peterson et al. (2016, pp. 22 – 25) describe the early newspaper as the space where contemporary literary figures across Africa experimented with their voice – be it writing fiction, poems, and essays. Writers began their careers through the newspaper which gave them an audience, and a source of income.

Powell, 2012, p. 14). Towards these goals, my thesis will explore and address the following research questions:

1. How can we best extract readers and their letters in historical print newspapers?

    The main motivation as a digital humanist is to compile and extract information in the digital space. Historically, the letters columns were a popular subset of African newspaper content, I argue that it constitutes a genre with an array of inherent intertextual relationships involving personalities, events, topics and readers.

2. What do we do with the extracted (meta)data?

    This expands from one of the central claims of this thesis: Digitization of newspapers is incomplete until the contents are systematically accessible.

3. Finally, how do we organize and present specific documents from the digitized archive?

    Here I present digitized newspaper documents as multidimensional documents using systems of bibliographic control.

1.3 Scope of the Study

In considering the questions posed above, my thesis will discuss and engage with the utility of newspapers as a primary source for cultural researchers, historians and humanities scholars. I will introduce important work by historians and their discussions of early African print cultures. While a significant part of existing literature in this field focuses on West African literature, the scope of my research will be limited to English-language newspapers published in Kenya between 1974–1978 by the *Daily NATION*. In the end, it is my hope that the findings of my research also contribute to further investigations into print cultures in early Eastern Africa. Moreover, the methodology and framework employed in this research can contribute to what Stephanie Newell (2013a) calls “empirical enquiries into situated texts and readerships [and how they] can help scholars to comprehend the variety of relationships" (p. 6) that existed within printed texts. Even though historical newspapers featured comic strips, cartoons[2], and pictorials as indexable texts[3] in their own right, this thesis will focus on the account of language texts, particularly within the letters column.

The *Daily NATION* was chosen because of all the available collections of interest, it was found to have consistently run the letters column. However, this does not impede on the significance of methods described since the letters section overall maintained a relatively similar column layout in African newspapers that were examined. The choice of this collection and period was based not so much on the propensity to support the current thesis as on significance of the medium, textual consistency and appropriateness of the corpus. All were critical properties to a robust framework, methodology and methods which I explicate in the later chapters. The *Daily NATION* corpus is but an exemplar; it is representative, not comprehensive: a complete collection would run into decades. Thus, this is an experiment through which larger textual and analytical research on newspaper readerships can be seen from.

Although my research is preoccupied by the letters column found in the newspaper collection, references to editorials and other articles in the column were not the main focus and can only represent opportunities for future research. Here, letters to the editor become the basic unit for indexing with all the contingent theorizing of the practice applied to a newspaper column.

[2] For instance, Nicholas Hiley indexed cartoons and created an online catalogue from newspaper cuttings. As an image index, it allows researchers to make complex search and retrieval queries across a vast collection of cartoons and caricatures (Hiley, 2006).

[3] Text in humanities often takes the broader sense of a cultural object onto which methods of reading and interpreting can then be developed. This text includes images, sculptures, oral records or other records (Gardiner & Musto, 2015, 32-35).

1.4 Glossary of terms

Bibliographic: While this adjective usually refers to citationary practices which direct users to books, chapters, and articles in the scholarly sense (Atkinson & Hudson, 1990; Landau & Wanger, 1980), its use in this thesis is by all means auspicious to include nonbibliographic resources as well, within these practices. Thus, the data (nonbibliographic) under exploration are not conflated with the adjective proper but rather conferred this definition to resonate with the current research objectives.

Body text: Textual content of a news article besides the title. This may at times include the byline or other details of the writer including address and locale.

Clipping: Also known as press cutting, refers to the practice of cutting out articles from periodicals, particularly common in newspapers. Born out of necessity, a way to capture significant pieces in a publication, for preservation or reprinting in the case of early editors in African print cultures.

Document: This is a twofold concept: The adjective 'document' denotes a basic unit of textual record, for instance, a *document image* resulting from digitizing a single newspaper issue. The second meaning on which this thesis leans towards is a specific article: a letter-to-the-editor as it was registered in the *Mailbox* column of the daily.

Identities: This deals with the concept of personae invented by readers who contributed by sending letters to the newspaper column. The assumption is that most of these names were pseudonymous and therefore cannot constitute positive identification of the readers. Thus reference to full or initialized names are appreciated as biographical metadata for mere analytical purposes. Mullan (2007) and Newell (2013d) exhaustively discuss the concepts of anonymity and pseudonymity as used in print literature.

Narrative: A prose account of topics, events and personalities involved. Letters published made reference to specific topics in the *Mailbox* column, as such they formed a concise thread of letters covering different issues.

1.5 Motivation & Context

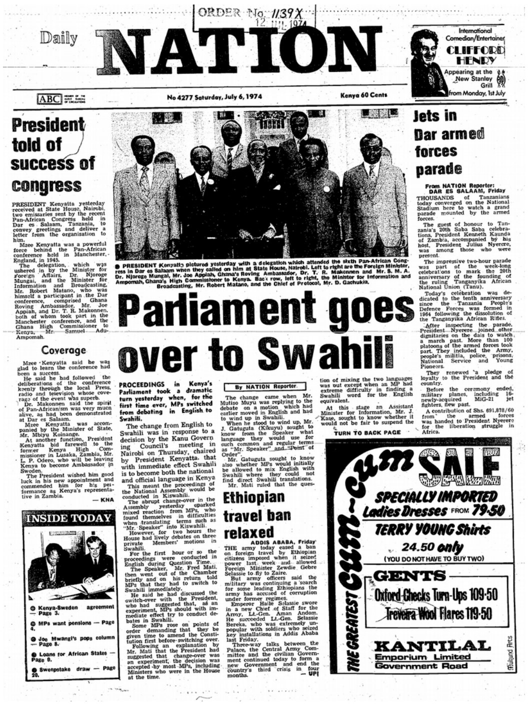
Humanities scholars and historians studying literary spaces in Africa continue to use newspapers as primary and secondary sources. Newspapers' documentary record is necessary to understand historical production of identities, representation, articulations and how readers interacted with the early periodical press in Africa (Barber, 2007; Newell, 2013a, 2013b, 2013c). The growing efforts to digitize periodical literature in the past few decades have resulted in a number of newspaper collection, preservation, reproduction and even indexing projects in and outside the continent (Howard-Reguindin, 2008; Simon, 2015). Many of these initiatives, for one reason or another, largely remain as reproductions of their original paper or microfilm. This minimizes their utility due to the lacking access strategies to contents, contexts, and the contributing readers.

For scholars seeking to trace and extrapolate the interactions of readers, newspapers are an important resource because as historical records, they generated and invited the public into their texts through such columns. This is a distinctive quality not found in other forms of literature. As Newell puts it in the introduction to *The Power to Name: A history of anonymity in colonial West Africa*: "Newspapers provide a substantial and unique resource for research into reader reception, cultural production, and political agency in Africa" (Newell, 2013d, p. 5). Newspaper readers articulated themselves as commentators on socio-cultural and political events of their time. For readers, the newspaper was a documentary record, a permanent register of opinions, and a channel for the expression of social, political and cultural realities (2013d). In the process, they engaged in a number of reflexive practices to quote and reference each other as well as foreground topics. Further, the creation of personae, use of proper or initialized names, and adoption of identities (pseudonymous and anonymous) altogether were attributes that respectively map to subjectivity, authorship, and agency.

To illustrate this, a reader wrote in response to an article published in the *Daily NATION* issue of July 9, 1974 (Figure 1.1) regarding parliamentary debate proceedings switching from English to Swahili:

> **POINT OF ORDER**
>
> IT appears our MPs are finding difficulty getting Kiswahili words for "Speaker", "Point of Order" etc. (DAILY NATION July 6). Here are a few suggestions: Speaker — *Msemi*; Point of Order — *Onyo*; Point of Information — *Kumbusho*; Division (of votes) — *Ita kura*; Motion —Utangulizi.
>
> These are but a few examples I am sure other readers, can suggest words to cover the whole Parliamentary language.
>
> **Abadi wa Funguni, Nairobi**

While this instance only mentions when the editorial was published, the reader concisely quoted both the article and date of issue. This mode of addressing the audience of readers, albeit crude, demonstrates the inherent intertextuality that we can only glean from close reading. Yet these widespread referential practices in the letters column were part of the culture mediated by print technology and they upheld topical threads across the dailies.



ORDER No. 1139X
12 JUL 1974

Daily NATION

ABC

No 4277 Saturday, July 6, 1974

Kenya 60 Cents

President told of success of congress

PRESIDENT Kenyatta yesterday received at State House, Nairobi, two emissaries sent by the recent Pan-African Congress held in Dar es Salaam, Tanzania, to convey greetings and deliver a letter from the organisation to him.

Mzee Kenyatta was a powerful force behind the Pan-African conference held in Manchester, England, in 1945.

The delegation, which was ushered in by the Minister for Foreign Affairs, Dr. Njoroge Mungai, and the Minister for Information and Broadcasting, Mr. Robert Matano, who was himself a participant in the Dar conference, comprised Ghana Roving Ambassador, Mr. Joe Appiah, and Dr. T. R. Makonnen, both of whom took part in the Manchester conference, and the Ghana High Commissioner to Kenya, Mr. Samuel Adu-Ampomah.

Coverage

Mzee Kenyatta said he was glad to learn the conference had been a success.

He said he had followed the deliberations of the conference keenly through the local Press, radio and television whose coverage of the event was superb.

Dr. Makonnen said the spirit of Pan-Africanism was very mucn alive, as had been demonstrated at Dar es Salaam.

Mzee Kenyatta was accompanied by the Minister of State, Mr. Mbiyu Koinange.

At another function, President Kenyatta bid farewell to the former Kenya High Commissioner in Lusaka, Zambia, Mr. L. P. Odero, who will be leaving Kenya to become Ambassador in Sweden.

The President wished him good luck in his new appointment and commended him for his performance as Kenya's representative in Zambia.

— KNA

INSIDE TODAY

- Kenya-Sweden agreement — Page 3.
- MPs want pensions — Page 4.
- Joe Mwangi's pops column — Page 8.
- Loans for African States — Page 9.
- Sweepstake draw — Page 20.

● PRESIDENT Kenyatta pictured yesterday with a delegation which attended the sixth Pan-African Congress in Dar es Salaam when they called on him at State House, Nairobi. Left to right are the Foreign Minister, Dr. Njoroge Mungai, Mr. Joe Appiah, Ghana's Roving Ambassador, Dr. T. R. Makonnen and Mr. S. M. A. Ampomah, Ghana's High Commissioner to Kenya. Back row, left to right, the Minister for Information and Broadcasting, Mr. Robert Matano, and the Chief of Protocol, Mr. D. Gachukia.

Parliament goes over to Swahili

PROCEEDINGS in Kenya's Parliament took a dramatic turn yesterday when, for the first time ever, MPs switched from debating in English to Swahili.

The change from English to Swahili was in response to a decision by the Kanu Governing Council's meeting in Nairobi on Thursday, chaired by President Kenyatta, that with immediate effect Swahili is to become both the national and official language in Kenya.

This meant the proceedings of the National Assembly would be conducted in Kiswahili.

The abrupt change-over in the Assembly yesterday sparked mixed reaction from MPs, who found themselves in difficulties when translating terms such as "Mr. Speaker" into Kiswahili.

However, for two hours the House had lively debates on three private Members' motions in Swahili.

For the first hour or so the proceedings were conducted in English during Question Time.

The Speaker, Mr. Fred Mati, then went out of the Chamber briefly and on his return told MPs that they had to switch to Swahili immediately.

He said he had discussed the switch-over with the President, who had suggested that, as an experiment, MPs should with immediate effect try to conduct debates in Swahili.

Some MPs rose on points of order demanding that they be given time to amend the Constitution first before switching over.

Following an explanation by Mr. Mati that the President had suggested that change-over was an experiment, the decision was accepted by most MPs, including Ministers who were in the House at the time.

By NATION Reporter

The change came when Mr. Mutiso Muyu was replying to the debate on a motion which had earlier moved in English and had to wind up in Swahili.

When he stood to wind up, Mr. J. Gatuguta (Kikuyu) sought to know from the Speaker what language they would use for such common and regular terms as "Mr. Speaker" and "Point of Order".

Mr. Gatuguta sought to know also whether MPs would initially be allowed to mix English with Swahili where they could not find direct Swahili translations.

Mr. Mati ruled that the question of mixing the two languages was out except when an MP had extreme difficulty in finding a Swahili word for the English equivalent.

At this stage an Assistant Minister for Information, Mr. J. Kase, sought to know whether it would not be fair to suspend the

TURN TO BACK PAGE

Ethiopian travel ban relaxed

ADDIS ABABA, Friday

THE army today eased a ban on foreign travel by Ethiopian citizens imposed when it seized power last week and allowed Foreign Minister Zewdie Gebre Selassie to fly to Zaire.

But army officers said the military was continuing a search for some leading Ethiopians the army has accused of corruption under former regimes.

Emperor Haile Selassie swore in a new Chief of Staff for the Army, Lt.-Gen. Aman Andom. He succeeded Lt.-Gen. Selassie Bereka, who was extremely unpopular with soldiers who seized key installations in Addis Ababa last Friday.

Three-way talks between the Palace, the Central Army Committee and the civilian Government continued today to form a new Government and end the country's third crisis in four months.

— UPI

Jets in Dar armed forces parade

From NATION Reporter:
DAR ES SALAAM, Friday

THOUSANDS of Tanzanians today converged on the National Stadium here to watch a grand parade mounted by the armed forces.

The guest of honour to Tanzania's 20th Saba Saba celebrations, President Kenneth Kaunda of Zambia, accompanied by his host, President Julius Nyerere, was among those who were present.

The impressive two-hour parade was part of the week-long celebrations to mark the 20th anniversary of the founding of the ruling Tanganyika African National Union (Tanu).

Today's celebration was dedicated to the tenth anniversary since the Tanzania People's Defence Forces was formed in 1964 following the dissolution of the Tanganyika African Rifles.

After inspecting the parade, President Nyerere joined other dignitaries on the dais to watch a march past. More than 100 platoons of the armed forces took part. They included the Army, people's militia, police, prisons, National Service and Young Pioneers.

They renewed a pledge of loyalty to the President and the country.

Before the ceremony ended, military planes, including 16 newly-acquired MiG-21 jet fighters, flew past.

A contribution of Shs. 691,878/60 from the armed forces was handed to President Nyerere for the liberation struggle in Africa.

Figure 1.1: "Parliament goes over to Swahili" frontpage article. (*Daily NATION*, 6 July 1974)

Commenting on a different topic, a reader in the *Daily NATION* newspaper published on July 17, 1975, in response to another reader illustrates the intra-referencing at play, replete with sententious rhetoric. The passage “Miss Salma Dathi’s letter (DAILY NATION, July 9)” points to content found in the daily.

> **Take care of kids**
> I SHOULD like to correct Miss Salma Dathi’s letter (DAILY NATION, July 9).
>
> Miss Dathi says that “It is so painful to know that teachers, who are very responsible people act so indecently”. Salma ought to comprehend the heavy burden teachers have while at school, and I fail to know why she has rushed to the Press to criticise our good teachers.
>
> Parents would be helped in the difficult task of bringing up children to be good and useful citizens with sound morality if parents themselves were ready to co-operate with teachers rather than criticising them.
>
> Many children are now in schools where the standard of behaviour would not be tolerated by the school committee and the teachers if these children were not given heavy punishment. It is for this reason that many people have started to urge as far as blaming teachers for nothing.
>
> Miss Salma should know that our children when at home are under the umbrella of their parents. During this period the parents ought to keep a keen eye on their children, particularly daughters, who might be loose to fall into temptation and get pregnant.
>
> **Stephen J. K. Ongaro,**
> **Kajiado.**

The editor of the column – as the gatekeeper – exercised a generous latitude on the nature of submissions to maintain a vibrant space for discussion both the societal minutiae and contemporary topical issues. As a way of discouraging prosaic banter, some limitations were imposed on the length and subject of the letters; as such, reader commentary was often topical, at times arguably intellectual in its delivery. But the topical nature of correspondences that were mediated by early printing press is largely ignored as a basic unit of documentary record in its own right, and collectively as pieces of literature enabled by their interconnectedness. Clearly, more accurate and improved access strategies make readers' articulations apparent and concise, and readily accessible for historians and humanities scholarship. The "mailbox" columns embody the notion of a convened *public* who regularly contributed to these columns in order to reify themselves and the space created by the editors. The public here, to channel Michael Warner's argumentation, is empirical rather than notional in that the readership was convoked in order to sustain the very discourse that conjured it (Warner, 2005, p. 67). In effect, these were the contemporary public intellectuals.

The letters column was a space where topical subjects as reprints, clippings, and editorials were sustained (James, 2016; Newell, 2011). An assemblage of these letters gives specific insight into the content, identities and relationships that were forged in those public correspondences. Thus, the bibliographic account with references and indexes is an attempt to reconstruct a continuous record of events and trace the history of information not so much as claims but a list of subjects with particular interests. In addition, an index of the column-readers registers their contributions fully with attributions, in the hope that it captures what Newell refers to: "author's performativity and playfulness which can be considered *aspects of* their authorial intentionality" (Newell, 2013d, p. 179). For these reasons the methodology seeks to uncover readers as contributors to the column and the newspaper as a whole. As literary specialist John Mullan notes: "Attribution was for centuries the common habit of readers, the consequence of having to read in the absence of the author's name" (2007, p. 5).

The digitized newspaper collection used in this reconstruction of narratives was obtained from the holdings of the Center for Research Libraries (CRL)[4]: a consortium of academic institutions with a vast collection of newspapers for researchers. As was the case in this and many other collections, the newspapers are available only as scanned PDFs. This characteristic of the archive is where the current research sought to engage with using a repertoire of methods and techniques. In three phases, the scanned PDF[5] files were processed recurrently to transform them into segmented letters, analyzed and

[4] CRL digital materials collections http://www.crl.edu/electronic-resources/collections.

[5] Also known as image PDF; scanned PDF is different from searchable PDF which has machine searchable texts. Scanned PDF is essentially an image where any recognizable textual characters are yet to be converted into searchable texts.

eventually indexable texts. The first phase tackled the issue of page layout analysis with the aim of identifying and collecting letters from the *Mailbox* column. Results from the first stage were subsequently processed using optical character recognition (OCR) tools and generated individual letters from the column. This stage not only produced texts that were now processable as a basic unit, but also generated machine-readable XML formats which relayed additional metadata about the letter.

The need for attribution also meant that titles, bylines, date references were extracted where possible. In the final phase, topic modeling and automated indexing approaches were utilized to create an index that interconnects the heretofore disconnected events, letters, and contributors. The automated tools employed in this thesis not only mimicked the theory and practice of information professionals, but also utilized heuristic techniques to determine the optimum values when matching names and assigning subject terms. The outcome of this pipeline was a demonstration of *mailbox* letters with individually assigned vocabulary terms from a thesaurus – among other access points.

1.6 Layout of Chapters

The thesis is organized into six chapters. As this introductory chapter has signaled, besides centering inherent genres found in newspapers as literature, the thesis is framed along a bibliographical lens and uses computerized tools of exploring the archive. Chapter 2 begins with the condition of the newspaper and its status as a documentary and historical record, more so in the African context. It also expands on the ongoing qualitative research on historical newspapers and the networks they conjured. Chapter 2 further delves into the review of indexing theories as they apply to digitized collections.

Foundational to my exploration of digitized newspapers is the concept of remediation. In Chapter 3, I engage with this theoretical scaffolding to hold the shift from print to digital media. Building on the preceding chapter, I will argue that accessibility in indexed texts using bibliographical tools entail reconstructing print relationships on a new medium. The chapter concludes by interpreting the remediation theory through digital methods, laying the groundwork for how the digitized collection used in this research will be explored.

In Chapter 4, I introduce my methodology to engage with the disciplines that inform this thesis, its techniques and methods. The choices made in this exploration and the interplay of techniques, algorithms, towards realizing the research questions will also be discussed in detail. This chapter will introduce the research data for examination before engaging with the methods. The methods of page layout and text analyses are central to the methodology, I elucidate them in each step in order to describe my thoughts and perceptions entangled with computational tools.

Chapter 5 presents the application of results found by this research. The exploratory nature gives warrant to discuss and represent how this process contributes to the utility and access of specific newspaper columns. It is also here that the output is put into perspective with the last research question. Chapter 6 consists of the summary and conclusion to the thesis research, and points to opportunities for further research.

Chapter 2

2 Literature Review

2.1 Introduction

This chapter examines scholarly literature around digitization of periodical press, and newspapers in particular. Due to the interdisciplinarity of this thesis, the relevant literature traverses the works of historians, philologists, textual scholars, digital humanists, and information scientists. I attempt to synthesize their interpretations and situate them within the purview of bibliographic practices. I will introduce and discuss the treatment of documentary records within library and information sciences (LIS) where large-scale archival and digitization of newspapers is situated. Considering the literature around automated indexing, I will cast light on LIS as an enabler of bibliographic work on vast digitized collections.

2.2 Newspaper as a record

Given that early newspapers largely constituted the register of obituaries, advertisements, and ordinances, their foremost role was to keep "record" of such daily events, lending themselves easily to the "newspaper as a record" axiom. More relevant for the current research goals, this "record" was also for people seeking to understand a particular time and place (Martin & Hansen, 1998). By indexing and sequencing events, newspapers became an enduring yardstick over a period of time. However, adopting this "record" perspective requires us to be aware of the challenges of periodizing, interpreting objects whose "status as productive literary forms" (Newell, 2011, 26) is less theorized. Newspapers are often studied as a unitary collection of data, rather than heterogenous publications with fairly constant titles.[6] Different types of newspapers existed as dailies or weeklies, and this influenced the topic they covered. In addition, newspapers like other serials were regularly published

[6] Bond sees having a constant publication title is one commonality between newspapers and other periodicals such magazines and journals (1969, para. 2-3).

and dated, making the currency of their reportage real. And so taken as a whole, content is where newspapers differed significantly with other periodicals such as journals (Bond, 1969). While book reviews, obituaries, poems, comic strips, and horoscopes were both important and distinctive genres in their own right, letters to the editor represented a form of interconnected and discursive print-mediated literature that is ongoing in the public sphere. This genric prescription is important and is alluded to by the repertoire of conventions used by readers in their contributions. As the results in this research will reveal, use of phrases like "allow me space ..." or "I refer to the letter of ..." for example were popular forms of addressing fellow readerships.

In *African Print Cultures*, Peterson, Hunter, and Newell promoted a number of solidarities that newspapers mediated by weakening spatial relationships between people. The underlying theme was that historical newspapers were more than appendages to the mainstream literature; they were assemblages of information and genres co-created by editors and their readers. Editors or readers thus clipped articles for reference and as a matter of cultural interest. The authors argue that when scholars simply treat newspapers as conveyances of data, they obscure the creative techniques by editors – clipping, reprinting, cut-and-pasting-citation – "to make connections and draw linkages" (Peterson, Hunter, & Newell, 2016, p. 1). This approach, they further argue, "makes it hard to see what African readers gained in their engagements with the media" (p. 1).

In digitizing records, technology provides the means to attenuate the cognitive space between text and the scholar by making it legible and parsable outside its print form. But accurate textual analysis is hinged on delineating contents, contexts and print-mediated selves; a daunting task that continues to attract growing research in particular genres (Underwood, 2014). The innate heterogeneity and irregular layout of content pose significant difficulties in large-scale analytical and empirical research. Content detection in historical literature and newspapers is a growing area of study, both in industries and academia, with extraction of novels and poems drawing much attention (Lorang, Soh, Datla, & Kulwicki, 2015; Underwood, 2014). Incidentally, Owens (2013) in a blog post called for systematic approaches aimed at freeing images from newspapers. Even so, significant work remains to be done on identification and description of reader-text interactions within newspapers.

2.3 Newspaper digitization

Seen as temporal recordings of the mundane, newspapers continue to play a key role for researchers and scholars in the fields of social sciences, humanities, and history as primary sources. The increasing scholarship in print subjects by cultural researchers has made newspapers valuable resources despite their ephemerality and skepticism in the onset of preservation and digitization projects (Bingham, 2010). Yet through these limitations and opportunities, the preservation and access of newspaper archives in the digital age – financial imperatives notwithstanding – have made little progress beyond the scanning phase of documentation. Subsequent processes in digitizing the voluminous archive is a complex undertaking considering page segmentation (physical and logical), labeling of various components, before finally extracting articles as complete documents (Zeni & Weldemariam, 2017). Since the early days of digitization, the bulk of these resources rarely served the needs of scholars in a number of ways: 1) the ability to perform searches using subject entry; 2) having online access direct to the content, and 3) a systematic structure to navigate the collection.

Digitized newspapers in both full text and scanned content versions present an attractive proposition and utility to scholars according to Bingham (2010). Especially in terms of locating content where organization, compilation and establishing relationships enrich their exploration. As mentioned above, the financial challenges and objectives of each project largely influence the level of digitization. Around Africa, few newspaper digitizations are online, and those that are, largely represent press clippings with limited forms of bibliographic access. The Mozambique History Net[7] is one instance consisting primarily of newspaper clippings from Mozambican and southern African sources covering broad topics such as economy, education, social studies and governance between 1960 – 2002.[8] Simon (2015) argues, despite the limited availability of digitized historical newsprint within Africa, this and other such projects are still fragmented in their collections.

In a notable paper presented at the 73rd IFLA General Conference panel on *African Newspapers: Access and Technology*, Howard-Reguindin (2008) reported on two digitization and indexing initiatives

[7] (2011, November 10). Mozambique History Net. Retrieved January 18, 2020, from http://www.mozambiquehistory.net/

[8] (n.d.). ilissAfrica - Detailed view: Mozambique History Net. Retrieved February 11, 2020, from http://www.ilissafrica.de/en/als/detail/125816

which need recognition as harbingers of the current research in newspapers. Set up in 2001, the Kenya Indexing Project (KIP) was the first such initiative in Kenya. It selectively compiled an online index of newspaper articles published in the country between 1986–2002. At its peak, approximately 70,000 newspaper articles had been manually digitized and indexed, covering cultural affairs in the country such as music, theater, dance, art and literature (Howard-Reguindin, 2008). The coverage of subjects was later expanded to include governance, law and gender issues. What stood out was the addition of control numbers, and use of subject terms based on the U.S. Library of Congress Subject Headings (LCSH), 24th ed[9] to index these press cuttings. Both scanned images of newspaper clippings and associated metadata were mounted on the indexkenya.org website.

The same paper by Howard-Reguindin also presented a similar initiative by the same project; however, unlike the previous archive that was online, the *gender issues* database was distributed free of charge on a CD-ROM (Howard-Reguindin, 2008). Figure 2.1 shows a copy of the digital library of selected newspaper articles. The portable digital archive was created in 2005, sponsored by the Ford Foundation and contained over 3,600 articles. Its reportage of gender issues between 1985–2005 was collected from seven newspapers based in Nairobi, some of which have long ceased publication. Being a derivative project of KIP, the subject headings on the index are fixed in time and lack the ability to keep up with current terminology (Nuckolls, 2015). Nonetheless, Howard-Reguindin (2008) concluded that such commendable work to clip and analyze articles was far from exhaustive against an "ever-growing backlog" (5). The conclusion from the report was clear that despite the sheer volume of newspapers' articles, there was significant merit for cultural researchers, literacy, governance among others in enhancing access to digitized newspapers.

As both initiatives demonstrated, the state of early digitization initiatives in Kenya – while selective and intentional – were not immune to the investments required to create and sustain these archives amid the increasing utility of technology.

[9] http://indexkenya.org/page.asp?name=about

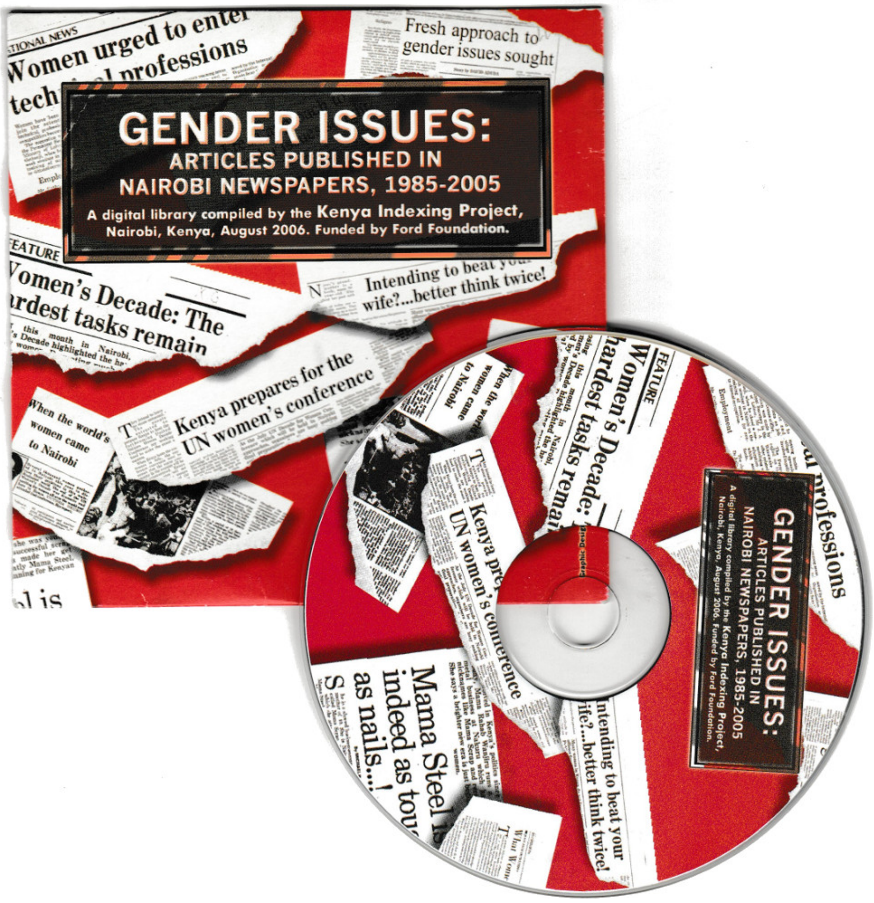

GENDER ISSUES:
ARTICLES PUBLISHED IN
NAIROBI NEWSPAPERS, 1985-2005
A digital library compiled by the Kenya Indexing Project, Nairobi, Kenya, August 2006. Funded by Ford Foundation.
Women urged to enter technical professions
Fresh approach to gender issues sought
Women's Decade: The hardest tasks remain
Intending to beat your wife?...better think twice!
Kenya prepares for the UN women's conference
When the world's women came to Nairobi
Mama Steel is indeed as tough as nails...!

Figure 2.1: Photograph showing a copy of the gender issues CD-ROM (January 11, 2020)

Though newspaper digitization initiatives in Kenya served as my point of departure, elsewhere in Africa studies examined the state of digitized newspapers to identify risks and strategies to address them. In cases where archives were mounted online, the content and functionality varied considerably to be regarded as reliable sources (Simon, 2015). Having partnerships with commercial publishers and other holdings presents a number of benefits including technical, administrative and resource sharing (Limb, 2005; Simon, 2015). James-Gilboe (2005) discussed how the value of digitized newspapers to end users was maximized by the ability to closely examine the coverage of news, and I add, helps gain a deeper understanding of correspondences. For scholars increasingly attuned to the hypermediated state of born-digital primary sources, digitization entail capturing, tagging and indexing all the information related to the extracted articles (James-Gilboe, 2005). Newspaper digitization, at the article level also enables classification into specific genres such as editorials, news, and opinions items (Klijn, 2008). An "article-focused" (160) digitization approach, as James-Gilboe

(2005) describes, augments the experience because the researcher can rely on keyword search to find other relevant articles; however, real value can be harnessed by identifying the various article components individually and as part of a corpus. In other words, it is through the transformation of individual components into sources of metadata that more robust access strategies and meaningful indexing is realized. It is noteworthy that technologies at the core of digitization have always imposed limitations that require us to balance accessibility and feasibility (Klijn, 2008).

Newspaper digitization remains a sparsely theorized area. The prolonged treatment given to journals and other publications is limited and selective even when applied to newspaper articles. A 2012 study by Krtalic & Hasenay has dealt with the value of newspapers as primary sources for research particularly in the social sciences and humanities. In the study, they conclude that bringing newspapers closer to researchers engender habits that spur and greatly extend the utility of collections (Krtalic & Hasenay, 2012). For indexers, newspapers evidently pose enormous collection, organization and indexing problems. Martin & Hansen (1998) identified key problems as content, fixity, reference, provenance and context (pp. 86-88, pp. 96-97). Being able to accurately refer to the sources in a fixed style across a collection are areas of concern particularly for historians using the archive. An index gives access to information at fixed positions, this information is then linked to similar information found in a collection. Whether in books or periodicals, the notion of fixity here relates to the unchanging nature of information regardless of the medium.

2.4 Refashioning letters to the editor

Digitization entails a purposeful reproduction of documentary records from analog form to digital. Be that as it may, accessibility and organization of individual documents, in most cases, is masked as a collection of digital images.[10] Textualizing document images which facilitates information retrieval is a basic task for the digital humanist, mostly because it makes the human record legible in mapping these expressions against the complex social-cultural system of knowledge organization common in academic literature. In the insightful collection of essays that is *The Power to Name,* Newell

[10] Digital image is used here because a scanned page in a newspaper may contain discrete texts, each constituting an individual document.

(2013d) states that only by modeling the creation of historical specificity of texts can we start to establish comparative studies with readerships in other localities, in the same period (pp. 180-181). As a digital humanist my central claim is therefore attached to the methodologies and methods that extract these letters as a genre, a dictionary of readers, and making intertextual characteristics clear and concise. Put differently, the objective is to explore a framework capable of distinguishing and identifying stylistic features (layouts, columns, texts) in the digitization process: for purposes of segmentation, extraction, analysis and indexing.

Letters to the editor also comprise secondary moments of discursive production. Hence the definition of indexing adopted by this thesis: the listing of key entities (bylines, location, date references, titles) and concepts embedded in a text, which has parallels to Anderson & Pérez-Carballo's (2001) definition of simply indicating the topics, meaning and purpose in documents. Similar to the scholarly journal indexes that measure and compare a researcher's productivity in a field, the utility of such collections is not only the quantified aspects but also pointing to the topics an individual contributes towards. A newspaper reader for example, actively engages in topics about moral, cultural, governance, civic issues which become reference texts. The digital reconstruction therefore not only contributes to the "user-centered indexing" as defined by Fidel (1994), it also reifies the concept of centering selves and identities.

2.5 Automated indexing

In the fields of information sciences, indexing literature is traditionally seen as the process of 'summarization' (Hutchins, 1977). Human indexers are tasked with describing what the text is 'about' in order to facilitate retrieval. Hutchins notes index entries represent a semantic condensation of content from subject texts. As a well-documented process, it involves identifying the salient keywords or phrases in a text. This argument has been the foundation of previous experiments in automated indexing. Earlier attempts by Gerard Salton (1991) noted that among other successes, the promise of automated indexing lies in "the ability to identify useful relationships between different texts" (979). A subsequent study evaluated these successes in the generation of themes and textual summaries

(Salton et al., 1994). While the study restricted itself to texts, its methods would come to characterize the traditional indexing approach (Moens, 2000, p. 62; Suominen, 2019), where a subject or topic is entered in order to retrieve a list of all documents related to the topic of inquiry. The alternative indexing approach advanced by Soumimen (2019) inverts this strategy by taking in a single document as input and returns as output the most relevant subject terms to describe the document. I discuss this implementation in the methods section of Chapter 4; however, it is apposite to first review the discussions of intellectual indexing in relation to automated processing of symbols, as it were.

For decades, human versus automated indexing debates have revolved around the implication of each approach on a number of information variables. The method of indexing (intellectual or automated) as the foremost characteristic, and a meta-variable itself demands an extended brief to which I shall return to shortly, but the other important variables of note are size of document, specificity, exhaustivity, surrogation and indexable matter (Anderson & Pérez-Carballo, 2001; Moens, 2000). The reflexive nature of either indexing method is whether it derives a word verbatim from the document or assigns a concept to a derived word or symbol from the document. The former, in the traditional sense of automated indexing, used weights, frequency and other statistical parameters to rank words of importance while the latter assigns its concepts from an authoritative source, presumably a human indexer or controlled vocabulary to match with derived words. Ultimately, they both *indicate* where specific information is located through description or other characteristics (Anderson, 1985).

Book chapters, journal articles and even editorials are considered sizable documentary units compared to paragraphs, titles and other similar-sized passages. Specificity refers to the exactness of terminology used to represent the text. Exhaustivity determines how central or peripheral the topic is by the number of occurrences. Surrogation is an amorphous quality where specific pieces of the text become representations, or surrogates, in place of full texts. Thus characteristics of a text document, for example; title, author, author's affiliation, and size are individually or collectively considered representative terms of the subject matter: they become access points to information. Finally, and more nuanced, indexable matter refers to either distinct components in the document, or the full content that is considered for indexing. Nuanced because, on the one hand, it warrants glossing over whole documents with bibliographic qualities (topical, references, authorship), and, on the other hand, considers the entire document that embodies the content as a whole (Anderson, 1985, p. 297). As such the last two attributes have deeper implications to the document size and exhaustivity (Anderson, 1985; Anderson & Pérez-Carballo, 2001).

Subject indexing of published materials refers to assignment of keywords from a thesaurus to describe a document within a larger field of interest or subject (Lancaster, 2003; Ripplinger, 2001). This restriction to a controlled vocabulary which also puts a representative "information specialist between the text and the user" (Cleveland & Cleveland, 2001, p. 36) is what forms the basis of intellectual indexing; whose focus so far has been on larger documents. Because they rely on extended detail to better map against a predefined concept, larger documents are retrieved more reliably in such a system. Surrogation and specificity lend easily to smaller abstract-like documentary units where concepts are apparent, and the abbreviated form limits the variability of aboutness of the document, which in turn, merits less intellectual analysis. Aboutness as a concept in LIS refers to the intrinsic subject of a document, beyond what is explicitly mentioned (Campbell, 2013; Moens et al., 1999). Aboutness is also determined by the purpose for which the document is being used (Fidel, 1994). The purpose is often achieved through listing of entry terms. Consequently, according to Ripplinger (2001) automated and unrestricted indexing while successful in document retrieval, is predicated on surface-level properties of the document and therefore suffers from keyword ambiguities, and an undue burden of coordination to the user (Cleveland & Cleveland, 2001). This inability to univocally indicate information is, to the best of my knowledge, another area of research which continues to show sparse but fairly significant activity from various scientific and humanistic domains.

Research by Anderson & Pérez-Carballo (2001) shows that while intellectual and automated indexing approaches are common and produce different results, users find them effective for information retrieval. Fidel discussed the nature of automated indexing:

> Automated indexing is clearly document oriented, because it is based on stored text. At the same time, its dynamic nature and flexibility make it a promising approach to user-centred indexing… From its early stages, automated indexing promoted the idea that indexing and searching are two sides of the same coin, and that both are dynamic and interactive processes. Another fact indicates the user-centered nature of automated indexing: It has never addressed the issue of aboutness; the major quest has always been to find the technique that results in the "best" retrieval, not the one that represents documents best.
>
> Thus, the current research in automated indexing reflects a contradiction. On the one hand, it is the most user-centred approach because of its dynamic, helpful, and flexible nature. On the other hand, indexing is based solely on the text stored and is completely immune to the particular group of users and their queries. (Fidel, 1994, 575)

There is a lot to unpack in the above reference, however, important for my analysis is the tension Felix highlights towards the end, which while presented as opposing attributes do in fact describe the effectiveness as a form of indexing. While automated indexing has been around for decades, its theoretical and technical range continues to expand in tandem with our exploration of technology and techniques. Examining other recent indexing principles echoes similar contrasts. Birger Hjørland (2017) classifies current subject indexing theories into two broad categories (58-59). The first is content-oriented indexing, which according to Hjørland hinges on a document's inherent "subject." In contrast, request-oriented indexing involves attributing subject terms to documents by a specialist to facilitate the intended use. Hjørland, however, considers automated indexing as too contextual to fit either principle.

While manual subject indexing is a rigorous and intellectual process, its automated counterpart is largely algorithmic. Among the foremost implementation strategies include associative and lexical approaches (Toepfer & Seifert, 2018). Associative approaches rely on the similarity of terms in the document with descriptors in the controlled vocabulary. With a large collection of manually indexed documents as training data, machine learning techniques can be used to ascribe concepts to new documents. Lexical approaches, on the other hand, prioritize the observable features such as salient terms in a document terms to match them with the vocabulary (Suominen, 2019). Though each approach individually may be inadequate to effectively describe the contents of documents, Toepfer &

Seifert (2018) also envisioned a fusion architecture that is leveraged on the strengths of each individual approach.

2.6 Indexing and Other Disciplines

Textual scholar Jerome McGann (2014) assailed our predilection of associating the study of documents of cultural memory with institutional apparatuses, whose purpose is to maintain particular rules and standards (p. 81).[11] Hence, the areas of indexing, abstracting and classification as traditional roles of LIS are there to, as Andersen (2008) says, "mediate society and culture" (p. 97). While McGann's contention is understated and critical, what Andersen (2008) is suggesting, however, is not a departure from methods or practices of LIS, but confronting the ideologies these roles embody by their determinations (p. 105). Whereas Anderson & Pérez-Carballo (2001) noted that the primary focus on larger documents by humans is not a distinctive element of the type of indexing (human vs. machine), in principle I argue that paragraph-sized documentary units are largely determined by automated indexing due to a perceived lack of credence in the content of such documents. Others have implicitly disqualified automated indexing altogether based on this premise:

> There is no automatic indexing tool available that could produce the index in the back of this book. Many of the automatic indexing tools available are not intended to be used with book-length narrative text. Instead, they are designed to process massive amounts of textual data stored as titles of articles and reports. So automatic indexing is not *indexing* as the word is used in this book. In many cases, a more descriptive term for automatic indexing would be *automatic list generation*. (Mulvany, 1994, p. 245)

Indeed, one of the main principles of indexing as described by F.W. Lancaster (2003) is the document's length as a distinguishing property for identifying subject matters. This single dimension of text has implications on which documents are considered for abstraction at all, foreclosing the utility of indexing and inclusion of other texts within newspapers.

[11] Similar disciplinary contentions have been raised by Walter Mignolo in a 2014 interview with López-Calvo, asserting that research in the traditional disciplines has "developed" often to maintain the status quo (López-Calvo & Mignolo, 2016).

What and how then do we consider documents left out of this frame? Asking this question effectively requires that we challenge the constructs that "privilege contents of a text above its consumption" (Newell, 2011, 26). Recognizing the traditions that ignore newspapers and their readers have further implications in how we treat the collections as legitimate text within the frameworks of textual and knowledge reproduction. Following the conclusion by Anderson & Pérez-Carballo (2001) that, automated indexing works – albeit differently – just as well as intellectual indexing, to automatically retrieve and index individual letters from *mailbox* columns therefore goes beyond showing semblance to intellectual indexing. In this differentiated syntax with respect to indexing, paying close attention to the contexts, forms of address, reference and other forms of surrogation become important distinguishing features of information in documents. The inherent post-coordinate capabilities that come with automated indexing allow greater contextual inferences. For example, the reference of dates and names in newspapers correspondences as illustrated in the opening chapter. And as we shall see in Chapter 4, use of dates and quotation structures within body texts constitute a range of contextual features based on association.

2.7 To Acknowledge Bylines

Following an earlier suggestion in this chapter under the section on newspaper digitization, one could argue the lack of theory and practice of subject indexing in newspapers is partly due to a lack of tradition, and structural disaffinity to organized branches of knowledge (technology and science). In addition, the lack of descriptions, classifications, or indexes in and of newspaper collections obscure their organization as knowledge systems. Due to this fact, creating an index of the letters column simply mobilizes a particular subset within a domain of literature. More importantly such indexes not only locate documents of interest, but also acknowledge "the creation of documentary-mediated persons and selves" (Day, Buckland, Furner, & Krajewski, 2014, p. 37). The increasing relevance of technology as a tool in humanities scholarship is central to how we can refigure archived documents in the new digitized spaces. For example, the ability to situate common byline components such as name, organization, and locale. Woodruff & Plaunt (1994), for that matter, envisioned a geographical indexing that would extract words pertaining to geographic place names in order to spatially attribute the document. These components help to identify and attribute authors as well as determine the subject topics of interest; and for scholars, items of interest within the vast collections of digitized texts. Organizing collections on such scale helps to reckon with the scarcity of metadata in large digital collections, on a macroscopic (Underwood, 2014) and microscopic scale.

2.8 Conclusion

In order to absorb readers' contributions in the press, a paradigm shift is necessary because such texts exist outside the scope of conventional literature, in the material and institutional sense. As Newell and others contend, early printing press mediated particular subjectivities and genres in the form of readers' articulations characterizing a discursive literary space. Examining the state of early newspaper digitization has also allowed us to new cast light on the theoretical and practical challenges present in this area of research. This, I have argued, gives us the cultural warrant to examine indexing practices outside the traditional apparatuses. Nonetheless, the LIS practices that organize information, however, remain critical to realizing the objectives of this thesis. My interpretations and engagement with disciplinary discussions, and evolving research have been situated along these objectives; as well as perspectives from other disciplines which are intertwined with this research.

Chapter 3

3 Theoretical Framework

3.1 Introduction

In this chapter, I now engage with the theoretical framing based on remediation theory used in digital media studies, writing, and understanding of digital cultures. Systems of locating and accessing documents are subject to the media which information is encoded. While the preceding chapter has discussed organized text as a documentary unit, it is important to understand the implications of our pursuits in digitized collections this research is exploring. Not only does the newspaper adopt a new format when digitized, its new-found malleability can be refigured to acknowledge the newspaper's old interface yet also rival the characteristics of born-digital documents including full-text searches, hyperlinks, and descriptors. It is this framing that I draw on to provide access to narratives, subjectivities, and topics found in the letters column. However, I will argue that reimagining location and access within digitized collections–that is, between documents, between subject terms, and between surrogates–has to be reconciled with both media traditions. By doing this, I hope to strengthen previous research on newspaper indexing and also acknowledge the contributions of allied disciplines to the research.

3.2 Remediation and the twin logic

Digitizing a document fundamentally marks the transition from a print to a computer-mediated interface, this shift borrows from predecessors and contemporaries while embodying the forms, contexts and properties of the new medium. Remediation theory as advanced by David Bolter and Richard Grusin discusses how content shifts across media – focusing on how a new medium *represents* an older medium (Bolter & Grusin, 1999). A medium exists in relation to other forms of media past and present, it absorbs, borrows and repurposes other media forms. For example, Bolter and Grusin point out the cinema is remediating drama, photography and painting (p. 67); the same way "televised news programs feature multiple video streams, split-screen displays, composites of graphics and text—a welter of media that is somehow meant to make the news more perspicuous" (p. 6). As a theory, remediation becomes an approach for describing phenomena particularly in the increasingly digital culture and screen media. In *Writing Space*, Bolter (2001) further articulates that remediation involves the tension of homage and rivalry as a new medium imitates features of another old medium (p. 23). It is a mimetic relation between the old and the new, which follows N. Katherine Hayles (2004) media-specific analysis of the complex relationships in print and screen. Remediation is better grasped (or complicated) through its double logic of immediacy and hypermediacy.

To the user, immediacy wants to reproduce old media in a transparent fashion as if unmediated. As a style of visual representation the goal of immediacy is "to make the viewer forget the presence of the medium … and believe that he is in the presence of the objects of representation" (Bolter & Grusin, 1999, pp. 272-273). This is evident in artistic work such as paintings which establish distance using perspective to look "real." The key to achieving immediacy in technology mediated objects is "placement of the user in the same virtual space as the represented object" (Samuels, 2006, p. 54), moreover, it is worth noting there exists other interpretations of immediacy beyond visual representation or perfect transparency. In audio for example, binaural soundscape is perhaps the most potent representation of immediacy through an immersive, and stereophonic reproduction. Fagerjord (2003), extending what the progenitors have suggested, further observed that alternative routes to immediacy also confront rather than dispel the presence of the medium. For Bolter & Grusin (1999), contemporary paintings as a matter of fact achieve immediacy not by denying its mediation but by acknowledging it (p. 58).

Pursuing immediacy, according to the theorists has often been through "the interplay of aesthetic value of transparency with techniques" (p. 24). These techniques of perspective painting, erasure, and automaticity were prevalent in early forms of media (photography, film, television) in the digital environment. If we are to fully interact with texts, it is through such techniques can we succeed in reimagining the old media. As a matter of fact, this is manifest in our encounters with digitized newspapers where, as scholars we discover and extract textual relationships, genres, and subjects; we are filled with a sense of immediacy that is apparent yet naturally absent on print. But transparent immediacy according to Bolter and Grusin is a shroud, not always a perfect representation of the old medium, "*immediacy* is our name for a family of beliefs and practices that express themselves differently at various times among various groups" (p. 30). OCR software, for instance, remediates historical texts because scholars not only "read" vast collections of digitized texts, they can also subject the texts to searching, mining, concordances and textual analysis. Such tools give us the ability to reveal patterns within texts, allow us to collapse breadth and depth which would otherwise be laborious if not "impossible to discern with the naked eye" (Röhle, 2012, p. 67). To pursue the next level of interaction with texts requires us to engage with hypermediation, the other logic of remediation.

Hypermediacy seeks to remove the authenticity of the old media. Its goal as a style of visual representation therefore "is to remind the viewer of the medium" (Bolter & Grusin, 1999, p. 272). We become aware of the new medium by how it extends the dimensions of a digitized artifact. Because mediation is at the core of digitization projects, hypermediacy entails all the refiguring to enrich our experience interacting with media. As such, the goal of this logic is to multiply "the signs of mediation and in this way [tries to] reproduce the rich sensorium of human experience" (p. 34). Through close reading on print media, this experience is innate since concepts, meaning and intertextuality are captured as information but the experience on a textual plane has to be reified with hyperlinks. If we are to trace these elements in the letters column (for example which/who topic/reader or author they were addressing in earlier issues), then this is not mere enrichment of the media, on the contrary, it generates new perspectives that stretches the claim of *this* old media to literature. The characteristics of the old medium are thus borrowed and reorganized. Hypermediacy from this lens is the result from the interplay of texts, techniques and processes especially in interacting with digitized collections to make textual relationships concrete.

Perhaps more pertinent for this thesis, subscribing to *remediation* theory means confronting disruptions arising from translating textual expressions through algorithms – the process of digitization. It is important to be aware these ruptures are results of a dialogical relationship between perspectives and processes, and work by “honoring both the social and the technological” (Rutenbeck, 2009, p. 92). Though “ruptures” in Bolter and Grusin’s position refers to technological limitations and notional “great potential” (Bolter & Grusin, 1999, p. 22), herein methods engage with technology as a continuum of practices (cutting and clipping) and processes (coding languages and algorithms) enacted on images in the pursuit of remediation. Coding is a translative language, which also allows us to critically engage with computers in the process. The challenge of translating human perspective into computer instructions requires us to reckon with the tools and methods of remediation.

3.3 Tools of Remediation

Interacting with computer vision tools that detect, filter, transform, and segment page images may be considered an aspect of transparent immediacy; which scholarship with digital methods entail. Owing to the fact that we achieve immediacy by juxtaposing the real and representational. The range of algorithms rendered by computer vision tools (e.g., OpenCV or Scikit-image[12]) allow us to analyze the many dimensions of digitized texts as documentary records. Mediation is therefore not dismissed or forgotten but constantly acknowledged as part of the methodology. These tools evoke transparency through visual manipulations of media such as those in the archives, in turn the inherent tensions are contrasted. Fagerjord (2003) engages with remediation theory in a more direct approach by arguing: “the core difference between hypermediacy and immediacy in all their different and confusing guises is their different strategies for achieving an unmediated authentic experience in ourselves. It is our experience that is the locus of remediation” (p. 304). Fagerjord asserts that Bolter and Grusin’s concept of “reality” is not stable, it shifts across texts, readers, viewers, and even media. In the context of digitized newspapers, we realize that rather than cohere, the two logics of remediation are sufficiently intertwined to represent the intricate relationships of texts and techniques. On the one hand, the digitized newsprint page usually in image form (a.k.a. page image) is reproduced exactly as

[12] OpenCV and Scikit-image are open-source computer vision and image processing libraries commonly used in detection, segmentation and manipulation of images. Although they both share similarities in terms of functionality, the methods in the current research are based on OpenCV.

an original print, giving visual access to full content; yet, on the other hand, as a bibliographic record, attributes from an individual article / correspondence are used to establish intertextuality in the collection.

The paradigm of remediation and its inherent twin logic is also useful in interrogating human-technology interactions at large. Bolter (2001) notes that remediation is manifested in how the concept of *hypertext*[13] is articulated by linking texts in the digital space analogous to printed footnotes (p. 28), showing the implicit and explicit claims of immediacy and hypermediacy. In *"Q i-jtb the Raven": Taking Dirty OCR Seriously*, Ryan Cordell (2017) begins by stating that: “We must understand mass digitized texts as assemblages of new editions, subsidiary editions and impressions of their historical sources, and these various parts require sustained bibliographic analysis and description” (p. 190). Cordell’s argument invokes the twin logic of remediation by how it honors the sources (albeit implicitly) and declares the imitations variably as editions and impressions. To faithfully remediate media is to suggest going beyond immediacy explicitly and more, it attests to the possibilities of generating meaning and connections anew within digital environments.

The computer is a multimedia device, a medium, for it constitutes the various types of media (text, images, video) within the digital space. However, since the currency of Bolter and Grusin’s theory is prefaced in visual forms of representation, we need to understand their definition of a medium. According to Bolter and Grusin (1999): “a medium is that which remediates. It is that which appropriates the techniques, forms, and social significance of other media and attempts to rival or refashion them in the name of the real” (p. 65). Therefore the computer is, furthermore, a tool and an automaton (Andersen, 2003, pp. 184-185) whose relationship with other media (old and new) is based on respect and rivalry. Cordell (2017) affirms this by pointing out that far from being just a window to the archive, the computer is an integrated system for remediating the archive (p. 192). As a medium, it is not wholly reproducing techniques for capture and digitizing analog materials, conversely it refashions and encapsulates practices into a new virtual space. Digitization stimulates the blending of forms, contents, functionalities, and contexts of textual expression and interaction in ways not

[13] Hypertext as a network of links on the world wide web that connects topics of significance and make the structure of linked documents transparent (Bolter, 2001, pp. 27-28).

possible before (Rutenbeck, 2009, p. 91). Media theorist Lev Manovich (2001) further notes two important ramifications that occur with digitization: 1) the image is formalized into mathematical terms which makes it, 2) “subject to algorithmic manipulation” (p. 27). Manovich concludes that “*media becomes programmable”* (p. 27). Indeed, this is evident in how contemporary social media platforms such as Instagram evoke nostalgia, distortion, and surreality by applying algorithmic ‘filters’ on photo images. In such instances of remediation, filters efface algorithmic mediation to evoke nostalgia yet the very filters are reproducing an old medium: photography.

3.4 Devices of Access

While perfect immediacy of the page image is apparent in our interactions with content on a display screen, it is by applying algorithmic manipulation that we can fully reproduce the intertextual relationships, extend dimensions and reenact textual practices on a hypermediated plane. In 2000, Gary Bradski published a paper entitled *The OpenCV Library*, a journal paper announcing an image processing library. As the first release of the library, its original goal was to advance the field of computer vision, and at the core was the ability to detect, recognize and segment an object/human/face in an image. Since its initial release, several iterations have expanded its utility across industries and academia. Other similar and equally robust open-source libraries have been developed since then such as the Scikit-image processing library by van der Walt et al. (2014). Nonetheless, the expansion of algorithms within OpenCV library in recent years has spurred further research and application into numerous fields such as machine learning and real-time video processing, the core morphological algorithms that detect, transform, and fragment objects in an image have become powerful tools in abstracting objects in images and video. More pertinent – particularly in the following Chapter – is the language of contours: one way computers perceive boundaries of a shape, draw an outline, or identify contiguous zones.

Advantages of the digital context are in the capacity to appropriate such techniques, technology, forms to generate scholarly or bibliographical digital editions of the ‘popular press’ as subjects worthy of analytical inquiry. One may also attribute the macro-level inaccessibility of entire genres of opinions, correspondences, or similar texts to: 1) focus on “page image” interfaces common in large collections; and 2), the limited application of scholarship devices of citation (Cordell, 2017, 193). Yet, on the other hand, the domain of humanistic scholarship views locating and compiling of information as a *sine qua non* to scholarship (Nichols, 2007), and so is the relevance of interdisciplinarity in mediating human expressions as they shift across media. To enhance and analytically consider documentary records also means to reconstruct the shadowy “selves” beneath the overlay of the image; it means to embody the devices that collate subjects, indexes, and the authors as references to other topics, texts, and readers. As Day, R. E., Buckland, M., Furner, J., & Krajewski, M. (2014) aptly suggest, successful mediation of documents and their textual elements in turn leads to subjective desires being effectively and efficiently mediated (p. 40).

To engage with textualities and analyses require articulation of content in the pieces that constitute a document within a newspaper column. Such that title, body text, byline, and other references generated within the text can be robustly determined. OCR results for the most part are remarkable in textualizing books, magazines and newspapers; obviously, hinged on the material state of the historical documents (Cordell, 2017, 194). Through the intermediate processes of segmenting page images, there lies a promise of remediating the newspaper column. By focusing on physical and logical explication, it lays the foundation for meaningful OCR-ing and subsequent approaches such as text analysis, topic modeling, and the very automated indexing tools. All these approaches are remediating print just not limited to the visual sense, but as textual representations for retrieval, discovery, indexing and referencing. They recontextualize texts.

3.5 Conclusion

My framework is motivated by how the shift in mediation from print to digital (on computer web) does facilitate access through bibliographic apparatus; that is, enables automated indexing from reconstructed texts. While I have attempted to frame computer mediated documents in this chapter, the goal is not to extol the state of the art in tools and techniques, rather to argue that as the antecedents to digitize texts, our parsing of *images* informs our treatment of consequent texts: for analytical, bibliographical, and even empirical purposes.

Chapter 4

4 Research Design

4.1 Introduction

Because the thesis set out to remediate newspapers, the letters column in particular, we can now begin to trace this process. The following exploration will negotiate the multilayered landscape of humanistic scholarship on which remediation of textual media takes place. Although the literature was situated within the purview of bibliographic objectives, the framing has provided the supporting theoretical assumptions. Having extended the disciplinary scope to include praxis of digital humanities, the following sections attempt to systematically describe the processes that refigured the letters column. The exploratory design of this research therefore draws from the current methodological and technical principles of computer vision, traditional library and information sciences, as well as studies in an attempt to reconstruct a scholarly digital edition within historical newspapers. As such, its outcome embraces bricolage; from the malleable and shifting nature of allied disciplines, goals, and tools that continue to expand scholarship in digitization.

Letters to the editor as a collection represent a genre and documentary record of human expressions. With this premise, the following section outlines the overarching methodology before introducing the actual data in detail, and then followed by an account of their processing. Throughout these sections, I describe the pipeline as an interplay of visual, structural and algorithmic methods of remediating a specific collection.

4.2 Methodology

It is helpful at this point to recapitulate the three research questions: 1) how do we best extract and compile letters to the editor both as cultural texts and documentary records; 2) what access points does the extracted data present for documentary analysis based on structural components such as topics, titles, “body text”, bylines, authorial identities, date references and other metadata? And 3) how do we organize and represent such documents within a bibliographic framework?

With these questions in mind, the analysis of documents was built on a methodology as diverse as the array of tools and disciplines that have informed the research so far. The methodological approaches of computing and information professionals (Bradski, 2000; Hebert, Palfray, Nicolas, Tranouez, & Paquet, 2014; Ferilli, 2011; Ferilli, Esposito, & Redavid, 2017; Hung-Ming Sun, 2005; Klijn, 2008; Konya & Eickeler, 2014; Lorang, Soh, Datla, & Kulwicki, 2015; Smith, 2009; Wang & Srihari, 1989; Wong, Casey, & Wahl, 1982) and how they have dealt with the challenges that pervade page image/layout segmentation. Combined with OCR processing strategies (Cordell, 2017; Koistinen, Kettunen, & Kervinen, 2017; Reul, Springmann, Wick, & Puppe, 2018; Smith, 2007). And the techniques that made the results of automatic transcription legible for bibliographic interpretation and control (Koistinen, Kettunen, & Kervinen, 2017; Reul, Springmann, Wick, & Puppe, 2018; Smith, 2007). Consequently, the overall methodology was divided into three stages encompassing image analysis, text encoding and analysis, and finally automated indexing. As mentioned earlier, each stage exploited a set of techniques and practices worthy of reflection as they related to the questions of the current research. Nonetheless, it is fair to note this methodology carried assumptions which are largely specific to this collection.

4.2.1 Seeing Images vs. Coding Images

The first phase dealt with the intricacies of transcribing human perspective on a computational environment. My goal in this phase was to segment and extricate individual letters published by the editor. This was done over a number of steps that included computationally detecting layout columns, extracting text regions, fragmenting text logically, and assembly of logical segments into letter articles. While the human perspective of boundaries is distinct in semantic and visual sense: that is, we can tell logical boundaries on page image content, pixel values to the human are fluid owing to the discreteness of color representation. In other words, the computational form requires an explicit grasp of the image boundaries, which despite rendering tractability, limits the scope for nuanced interaction (McCarty, 2004). This attempt to honor antecedent practices of newspaper clipping, as it were, within a computational model is fraught with difficulties of a representational and semantic nature. Williard McCarty (2004) asserts there are epistemological implications when we are forced to "confront the radical difference between what we know and what we can specify computationally" (p. 256). McCarty is, in similar but other words, echoing the distinctions described by N. Katherine Hayles (2004) between the visible page image on the screen and its underlying code (78-79). Hayles argued that code often is "invisible to the casual user" (79), but can be represented or reconstructed through special techniques and software. This rendering of intermediate image representations as part of methodology becomes an opportunity to clarify assumptions, reflect on the malleability of digitized texts and also acknowledge our "altered way of thinking" (McCarty, 2004, p. 256), inside the virtual space. Even so, at the core of the rule-based system[14] is the constant reification of binarism: creating dichotomies between texts and non-texts, title and body text, foreground and background. As I illustrate later in the methods, handling the vagaries of content segmentation while constrained by the binary logic was an interpretive process of observing repeated patterns and structures: it was an exercise in reflexivity.

Page segmentation is one technique of image analysis that is most relevant to the current research. Whether physical and logical, it gives us the capability to handle the image representation in more meaningful ways. Where objects are not only manipulable, they can be extricated as texts and labeled as such based solely on detectable color properties. I used rule-based approach as opposed to machine learning in this stage for a couple of reasons:

1. The page layout of the letters column had subtle changes over the years (especially in the header section), though infrequent, the use of a machine learning approach had less feasibility due to computational costs of testing and retraining models. The size of each distinguished set would constrain effective training to result in reliable segmentation. Further, because supervised machine learning works by educing and

[14] Rules-based system is a static system with conditional steps defined by a human, as opposed to a system that "learns" by training.

generalizing, its effectiveness and ability to extrapolate are predicated on having considerable training data (Esposito, Ferilli, Basile, & Di Mauro, 2008; Ferilli, 2011).

2. I had a good grasp of topological structures within the collection, as the methods section will demonstrate, it provided flexibility in adjusting and testing rule changes for nuances that would otherwise require retraining machine learning models. In other words, the ambiguity that favors unsupervised machine learning methods (Konya & Eickeler, 2014) was less pronounced.

When Furmaniak (2007) experimented with a combined approach of rule-based and machine learning methods, the prevailing assumption was that newspaper layout and topological structure are generally consistent. The approach proposed by Furmaniak combined unsupervised machine learning, and block segmentation (the primary approach used in the current research) with a language similarity component. In the latter step, texts obtained from *text blocks* are compared with each other to determine their semantic relatedness. For example, if adjacent blocks mention the same word multiple times, they are considered part of the same article. Despite this approach showing high accuracy results according to Furmaniak on newspaper articles, the heavy reliance on language similarity of body text, though effective, fails to account for other components of the article. The title or byline for instance, was not considered as logical and constituent part of a segmented block.

As the methods section will show, the letters column as a matter of fact represent the more consistent parts in a newspaper. Patterns of textual content therefore revealed structural rules as discernible and somewhat stable; titles then become integral pieces for establishing sequential relationships and boundaries of content across columns.

4.2.2 Transcribing Images

In the second phase, I built directly on the results of image analysis – which were segmented letters. This began by actual extraction of textual content through OCR software. As the text encoding and analysis phase, a set of pre- and post-processing steps were involved for bibliographical and curational purposes. First, the vagarious results from the OCR process demanded that particular emphasis was given not only to removal of characters considered as textual noise (e.g. {, *, °, =, @) from the OCR, but also attend to the delicate and problematic process of automated spell-checking. Problematic because though it was necessary to ensure coherence and improve the quality of text, there was a risk of miscorrection (as an automated process) particularly on local terminologies and references. The second step attempted to split OCR outputs into distinct text blocks of titles and body texts.

Inasmuch as the extrication of these components was more systematic, through heuristics and abduction[15] of font size and whitespace, determining bylines from body text presented a challenge since for the most part, they were intertwined pieces of text. Body text was heterogenous and comprised both *body text* and *byline*. Further, OCR results did not faithfully represent the textual topology, that is; italics, bold or normal font sizes, which in this instance would have served as a potential byline discriminant. The intention was to glean into both biographical and bibliographic data in the network of readers as identities that contributed to the newspaper column.

4.2.3 *Mailbox* Indexing

The third phase of the methodology was concerned with approaches for organizing and assigning index terms to this collection. This stage made use of existing algorithmic tools to assign bibliographic, referential and intertextual features. Following the suggested alternative subject indexing approaches in the literature, automated indexing explored the use of a controlled vocabulary. The letters were labeled with subject terms using the well-known UNESCO Thesaurus.

The first edition of the Thesaurus was published in English in 1977, based on concepts largely derived from analysis of UNESCO documents and publications (Garrod, 2000). Outside the international agency, its appeal to other organizations and institutions for subject analysis and retrieval of documents has been attributed to its reasonable size (currently 4421 terms) and greater

[15] Dixon posits abduction as the “seeing [of] patterns where there are patterns and creating the correct interpretation” (Dixon, 2012, 201-202). According to Dixon abductive reasoning involves spotting patterns and relationships in data, forming a foundation on which hypotheses are created; the other two types of reasoning (induction and deduction) are therefore methods by which they are proven.

accessibility by “staff who [are] not indexing specialists” (39). It has elsewhere been applied in a number of cultural heritage digitization projects (Kapsalis, 2019; Shiri et al., 2010). The current edition is available in Russian, French, Spanish as well as English. All the Thesaurus concepts / terms are organized into seven knowledge categories namely “Education”, “Science”, “Culture”, “Social and human sciences”, “Information and communication”, “Politics, law and economics”, “Countries and country groupings”; which are further divided into 88 micro-thesauri (Martínez-González & Alvite-Díez, 2019). By forming a hierarchical structure, each micro-thesaurus is univocally a subset of the super groups, and part of the larger UNESCO Thesaurus. For example, “Languages” micro-thesaurus belongs to the “Culture” super group as well as mapped to the broader concept of “Bantu Languages” which is part of an even broader concept of “African Languages”.

Although any thesaurus could have been used including a custom one, the UNESCO thesaurus was both effective and readily coincided with the automatically derived top topics. The labeling of letters utilized a comprehensive solution developed by Osma Suominen at the National Library of Finland.[16] Annif, as it is called, was suited for this task because it integrated and supported a variety of existing algorithms in natural language processing for the specific purpose of automated indexing. Built in Python programming language primarily as a microservice, it is used to describe documents given a controlled vocabulary as input. Annif is multilingual by design and has a built-in trilingual controlled vocabulary based on the YSO – General Finnish Ontology.[17] However, Annif implementation at the time lacked support for hierarchical structure found in the UNESCO Thesaurus, therefore a flat list vocabulary was constructed from all categories / concepts / classes for subject indexing. Approximately 4,400 English terms were extracted from the Thesaurus for subject indexing.[18]

[16] As an open source tool with institutional affiliation, Annif is an active project with regular updates, integrations and well is well documented (see https://github.com/NatLibFi/Annif/).

[17] The General Finnish Ontology (YSO) is published and maintained by the National Library of Finland. As of this writing, it contains over 27,000 general concepts in three languages namely: English, Finnish and Swedish (see https://finto.fi/yso/en/).

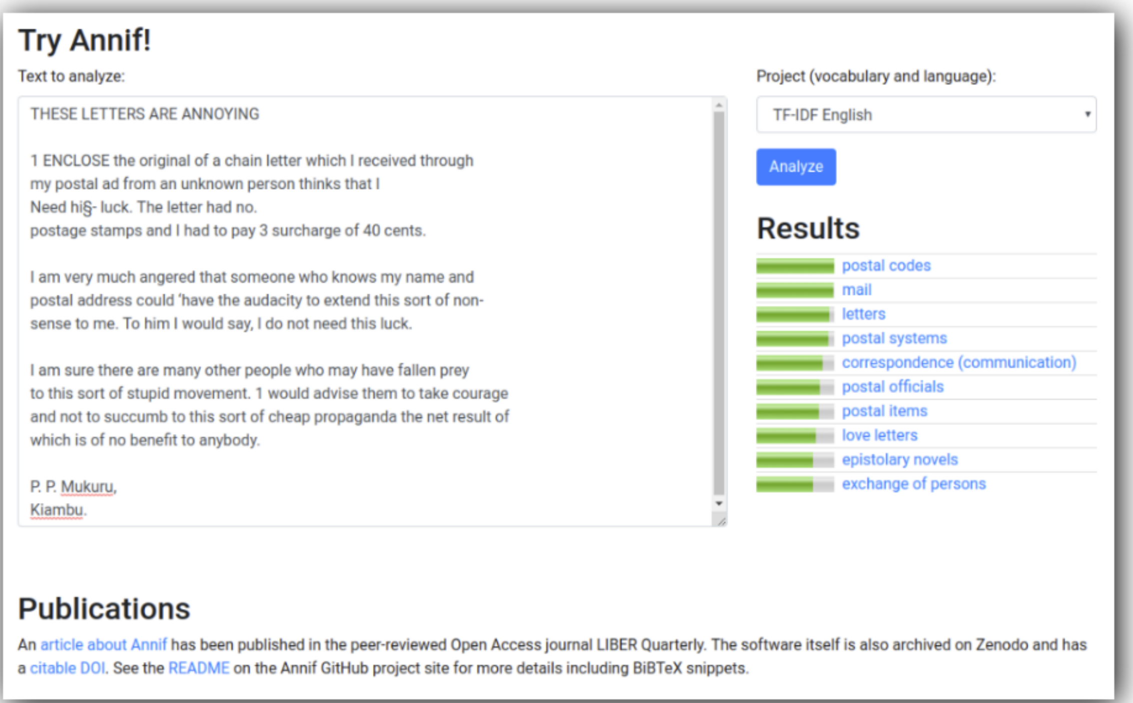
[18] The Jupyter Notebook for this process can be found at: https://github.com/ooduor/lettersiterate/blob/master/notebooks/create_annif_unesco_vocab.ipynb (Antony, “Create Annif TSV-format Vocab from UNESCO Thesaurus”).

Because Annif was built as an automated subject indexing tool. It integrated a number of intermediate steps under its *analyze* command which: converted document contents into tokens, normalized singular and plural terms, standardized text to lowercase, and determined lemma.[19] The resulting textual representation was treated simply as a bag-of-words[20] – devoid of context, syntax or order of terms within document texts. This new representation (a.k.a. vector space), allowed calculations to establish similarities and co-occurrences of terms as well as probabilities when assessing relevant documents (Ignatow & Mihalcea, 2017). According to Suominen (2019), the motivation behind Annif was to allow input of a "single document and the output would be the most relevant topics for that document" (5), in a way the goal was "to turn a traditional text index on its head" (5) – where topic is used to locate documents. To achieve this goal, however, required Annif to be trained prior using a subset of the documents that were labeled with the vocabulary.

Since the first prototype in 2017, the Annif project has evolved to support a growing number of algorithms that implement automated indexing from lexical, associative, and a combined approach. The "Try Annif" feature on the Annif homepage (see Figure 4.1), allows users to try these algorithms and subject vocabularies by choosing one of each against a document text which is then indexed and the output is a list of key topics ranked from the most relevant. Each of these algorithms, or backends as they are called, consist of different strategies for indexing a given subject vocabulary. For example, one of the backends uses Term Frequency - Inverse Document Frequency or *TF-IDF*, a method of characterizing topics based on the frequency of rare terms in the document with diminishing influence if the term occurs in more documents.

[19] Lemma refers to the base or canonical form of a word. Inflected words are normally grouped together to reduce morphological variation in the text.

[20] Bag-of-words refers to the representation of document text as a mere set of terms that occur in the document. This treatment in the topic modeling process allows us to establish co-occurrence of terms.



Try Annif!

Text to analyze:

THESE LETTERS ARE ANNOYING

1 ENCLOSE the original of a chain letter which I received through
my postal ad from an unknown person thinks that I
Need hi§- luck. The letter had no.
postage stamps and I had to pay 3 surcharge of 40 cents.

I am very much angered that someone who knows my name and
postal address could 'have the audacity to extend this sort of non-
sense to me. To him I would say, I do not need this luck.

I am sure there are many other people who may have fallen prey
to this sort of stupid movement. 1 would advise them to take courage
and not to succumb to this sort of cheap propaganda the net result of
which is of no benefit to anybody.

P. P. Mukuru,
Kiambu.

Project (vocabulary and language): TF-IDF English

Results

postal codes
mail
letters
postal systems
correspondence (communication)
postal officials
postal items
love letters
epistolary novels
exchange of persons

Publications

An article about Annif has been published in the peer-reviewed Open Access journal LIBER Quarterly. The software itself is also archived on Zenodo and has a citable DOI. See the README on the Annif GitHub project site for more details including BiBTeX snippets.

Figure 4.1: Annif's "Try Annif" interface at (http://annif.org/) allows you to enter text and select a vocabulary to index with. The results show terms based on YSO English vocabulary for a letter extracted from the current collection.

The derived terms in this case create an index for matching term frequencies in new documents about specific subjects (Suominen, 2019). As one form of associative indexing, TF-IDF was used to leverage on the different topics found in individual letters, which in turn were mapped to thesaurus terms to label the training set of documents. Once the documents were trained and evaluated, the backend could be used to assign subject terms to an unseen document. Worth noting, Annif also supports *Ensemble* backends where results from multiple backends are combined to mitigate drawbacks of one strategy and maximize the strengths of individual strategies.

The methods that follow entail the steps in each of the three phases described above. I begin by describing the newspaper briefly mentioned in the introductory chapter, and how it was collected for the current research.

4.3 Data Collection

The above methodology and research questions were conducted on a single newspaper publication; The *Daily NATION*, an English-language newspaper whose history goes back to the early 1960s.[21] Its emergence marked a new era of English-language publications in the Kenyan press, after decades of growth in African-controlled vernacular press.[22] (Gadsden, 1980) This growth Gadsden attributes partly as a reaction to "the political frustrations suffered by Kenya's Africans during these years" (516). The need to address wider audiences contributed to the shift towards Swahili – the lingua franca – and English-language newspapers. As such, the *Daily NATION* enjoyed wide readership and had a long-running letters-to-the-editor column.

Although the original research was interested on the aforementioned Kenya Indexing Project (KIP) online database. The project's indexed collection of newspaper clippings and subject access did not have full content mounted at the time. Due to the availability of subject indexes by specialists and a wide variety of local newspapers, this collection would have been helpful in scope (multiple newspaper publications) and for comparison between indexing methods in addition to the exploratory capacity similar to that of the current collection. While current research in essence explores locating units of texts within page images, locating primary data, in this case, partial or lack of full content is itself a challenge. Still, a number of newspaper holdings exist that have digitized newspapers dating back to the beginning of the 19th century. The majority of these holdings consist of consortia of academic institutions. The CRL is one such consortium whose large holding and network of newspapers this research relied on to gather the corpus.

Complete issues of the digitized newspapers published from 1974 to 1978 were downloaded at the beginning of the research. The individual issues in the CRL collection were organized incrementally and greatly eased the extraction of issues within the period of interest. Even though the

[21] For a comprehensive background on the *Daily NATION* newspaper and printing press in Kenya in General. Kenya Press, Media, TV, Radio, Newspapers—Television, circulation, stations, papers, number, print, freedom. (n.d.). Retrieved February 10, 2020, from http://www.pressreference.com/Gu-Ku/Kenya.html.

[22] According to Jackson (2007), vernacular press here refers to local language publications that are not designated as an official language by the state (p. 91). This can also be expanded to include any language not lingua franca.

quality of the scanned images was consistent and legible across the years, initial close examination revealed the quality of scans deteriorated over the years in terms of visual noise. This observation, while intriguing at first, impacted the pre-processing of page images for physical and semantic (or logical) segmentation workflow. It was yet another testament to the vagaries of the digitized texts, and precarious state of source materials.

Initial preparatory work involved a number of transformation steps on the files. Since every daily issue was a complete file with all the pages in print sequence, each file had to be split into separate pages. As I mentioned in Chapter 1, the original file format was scanned PDF, which was not compatible with the intricate segmentation workflow. Thus, the next step converted individual PDF files into PNG bitmap image (as opposed to JPEG) format in order to maintain adequate fidelity of the original file with a decent size (in bytes) of the resulting image. The economy of computer vision algorithms hinged on exploiting the quality and size of image files to maximize accuracy and tractability in digital contexts: a premise on which the subsequent workflow of OCR-ing, would benefit from. See Appendix A for more on the data collection and processing scripts.

The *Mailbox* column as it was named in the daily, consistently appeared on the same page, of every weekday over the years. This was serendipitous given that methods in this research were also enriched by such observations in the collection, it allowed recurrent strategies to automate extraction of all the *Mailbox* page images from the corpus. The benefit of this process not only reinforced the presence of useful corpus patterns, but also revealed the synergistic entanglements of “close” and distant reading that are part of these digital methods. It was a form of orienting with untheorized collections of texts while being cognizant of patterns (Cohen, 2009, 59), only this time, patterns observed are structural, and embedded within the corpus.

4.4 Methods of Page Layout Analysis

The *Mailbox* column as shown in Figure 4.2 (a - d) was a five-column structure, some letters occupied the entire length of a single column, or a part of it while others span multiple columns: such variations were constant. Preprocessing of these *Mailbox* images began by collapsing the grayscale gamut of each page image into a monochromatic representation – a binary image consisting of pixels with only two colors – background (white) and foreground (black) color. The importance of this thresholding technique was setting the stage for algorithmic manipulation of any visible object into numeric values including, detecting presence of boundaries or edges between visible objects, and filtering out visual noise on the image.

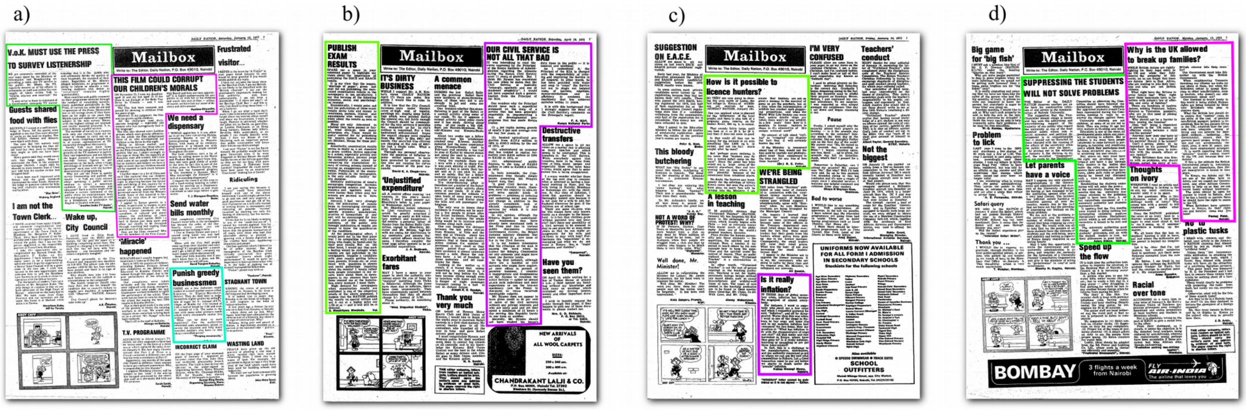


Figure 4.2: Sample of *Mailbox* column with varying single and multicolumn letters. (a) Highlighted letters attempt to show how variable the formats were on different issues spanning single or multiple *physical* columns. (b) Full-column letters, and full-length as well as multi-column letters in (c) and (d).

As I had discussed earlier, such rules and heuristics in preprocessing are necessary to translate the differentiated human-computer perception of objects. The process entails subsuming or excluding foreground from background colors in order to expose lines and shapes. Canny edge detection algorithm, devised by John F. Canny (1986), was used to extract edges based on low and high threshold values. These two values help establish lines: big enough to consider; those to ignore, and those that are below the upper threshold but are connected to it. Tuning these values individually was not practical, they were empirically determined from each page image to optimize results. Figure 4.3

shows the resulting edge image, austere and juxtaposed with the original image to make the meaningful boundaries clear.

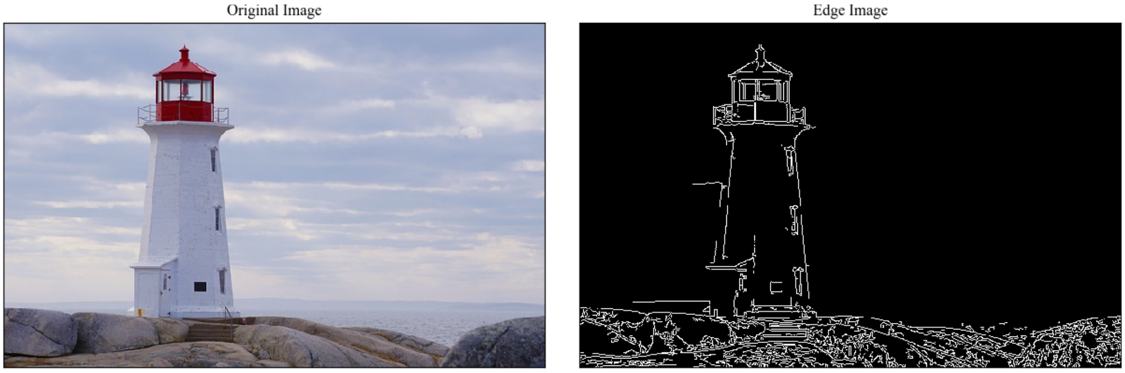


Figure 4.3: Canny edge detection algorithm applied to a photo image and the converted into creating a binary mask. Coldhamr (Photographer). (2017, February 28). Photo of a lighthouse [Photograph]. Retrieved from https://pixabay.com/photos/lighthouse-nova-scotia-canada-ocean-2101763/

The preprocessing described thus far was adequate for determining simple lines and polygons, since non-text objects complicated human-computer translation of the letter column. While the presence of comic strips, advertisements, and other imagery within this column call to mind the heterogeneity of newspaper content, they also lend themselves well to the bi-leveling narrative – "bi" because of continuous need to determine text and non-text regions on the image – and how this binary epistemology is based on the practice of juxtaposing and opposition. For example, to juxtapose the edge image to the original image as shown in Figure 4.4 suggests an unapparent opposition within the image, although to the reader it appears as mere inversion of colors. Computationally, however, transformations on the edge image have made explicit boundaries of every character at a pixel level by removing noise from the foreground objects.

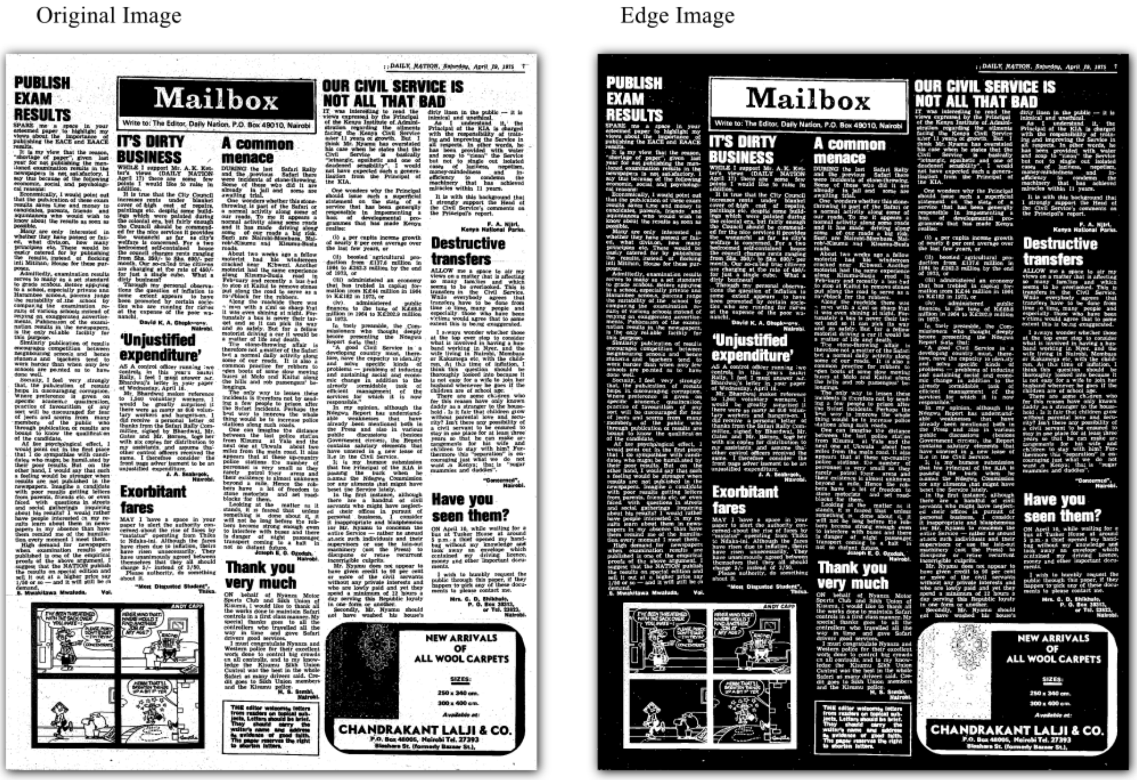


Figure 4.4: Canny edge detection algorithm applied to page image.

Collecting non-texts involved two algorithmic steps which started by extracting lines and polygons. OpenCV is optimized for feature extraction, the first step used probabilistic algorithms to detect lines in the image. Line segmentation algorithms in OpenCV determine lines based on minimum length and maximum gap between points considered to be on the same path. The second step involved finding all shapes with contiguous boundaries, also known as contours, and then computing the size of each bounded area. Whereas edges as illustrated on Figure 4.3 appear like strands of pixels on the image – contours expand this concept by connecting edges into closed shapes. I shall return to the landscape of contours later when addressing its utility in recognizing shapes. Due to the text-heavy nature of these page images, contours for non-texts were significantly larger so any contour exceeding the area of an average contour was extracted to a derivative mask. As shown in Figure 4.5, both masks were combined to form the basis for areas to exclude from the image.

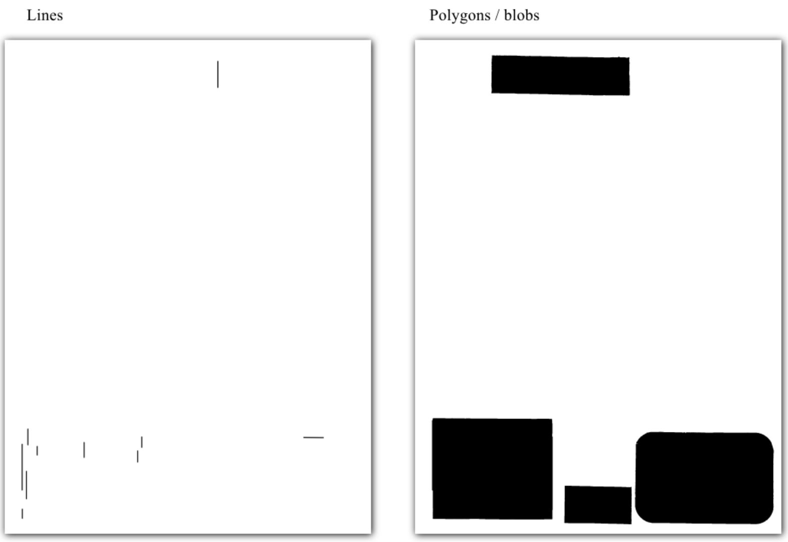


Figure 4.5: Mask images derived from the binary image. They comprise lines/edges and polygons.

Masking as an image manipulation technique transformed the subject image using a mask to exclude in the resulting image. A mask therefore was a binary image of the same dimensions as the original, created to perform conjunctive or disjunctive operations. In one such operation, when the mask is applied to another image, specific objects on the mask that overlap with the binary image are excluded. The secondary utility of masks rested on the ability to perform arithmetic functions successively to obtain a desired outcome. Figure 4.6 illustrates the application of masking in this instance.

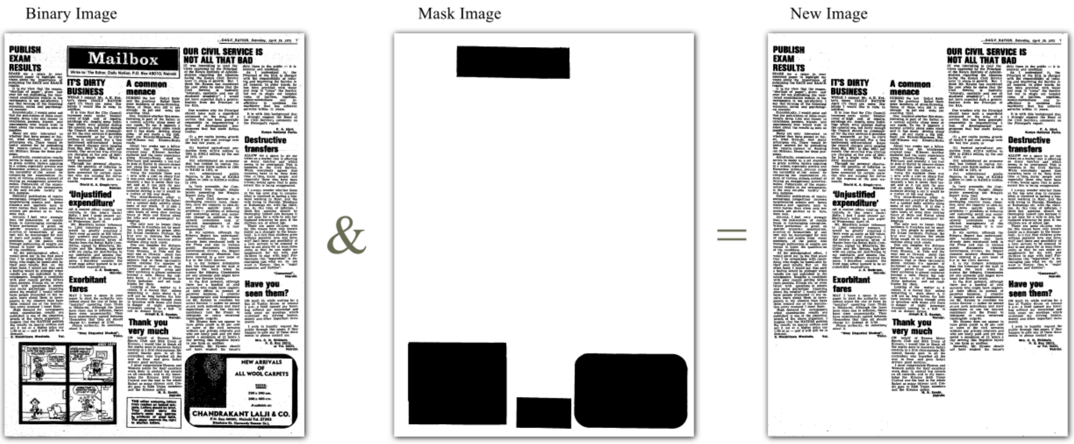


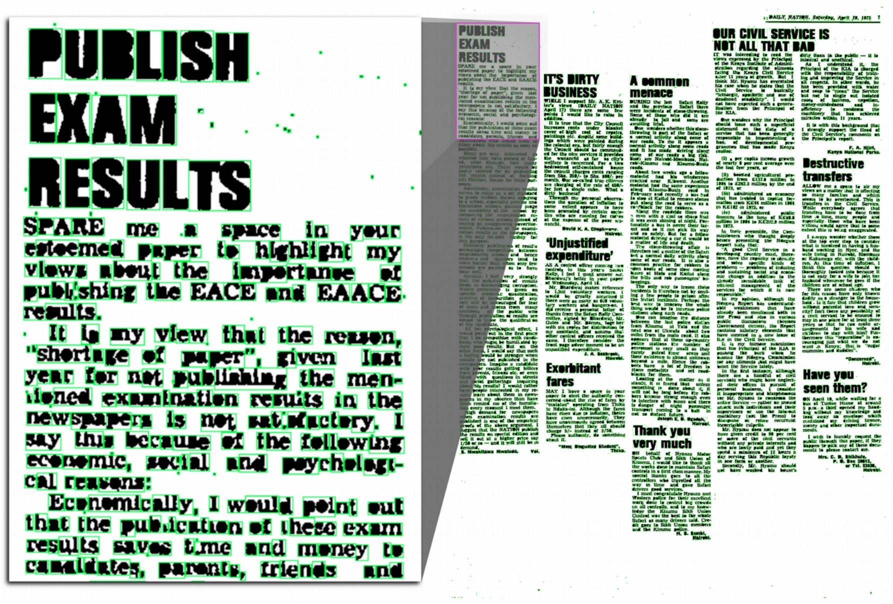
Figure 4.6: Imposing a mask onto a binary image. Overlapping pixels in the mask are excluded in the new image.

By making the new image content homogenous–that is, text only–it proffered qualities parallel, in some ways, to the juxtaposition and opposition of titles and body texts. Whereas the new image marked a departure from physical segmentation of the image (delineated text from non-text), it also signaled the change into logical and semantic segmentation of the column. This is an important point to note because though image segmentation is at the core of OCR processing (Cao, Prasad, Natarajan, & MacRostie, 2007; Smith, 2009), it often presupposes texts as *only* disjointed by columns and non-texts: giving less attention to the specific contexts of texts (within the page) than to recovery of text blocks and characters. For example, the text-only image at this stage was not only comprised of "meaningful components" (Cao, Prasad, Natarajan, & MacRostie, 2007, para. 1), it is still obscured to reveal the variety of textual relationships (between topics, authors, dates) yet to be parsed.

Owing to its continued use in this section, it is appropriate to revisit the concept of "contours" as earlier promised to add clarity for the reader. Contours are at the core of shape detection and analysis in computer vision providing a number of features for detecting, analyzing, and handling image objects. These include ascertaining area, perimeter, and bounding box in an image. This last characteristic in particular exposed the following features and properties useful for this research:

1. Contour number *(#)* automatically assigned on the bounded object
2. Box X-coordinate *(x)* starting from top-left corner of the document,
3. Box Y-coordinate *(y)* starting from top-left corner of the document,
4. Bounding box height *(h)*
5. Bounding box width *(w)*

The bounding rectangle was calculated to encompass the contour shape, located by x, y coordinates as well as both height and width measurements. These properties facilitated two preprocessing steps using a number of heuristics. First, the average contour height was representative of the body text, and as such contours that make up a title were identified by a much bigger height. Subsequently, the remaining contours were determined as body text; this was not before discarding specks and other residual visual noise that had persisted, which were significantly smaller (in size), from the image. From this interaction, titles and body texts emerged as two disparate streams of grouping content blocks. But insofar as contours succeeded in projecting (see Figure 4.7) such distinctive properties, they also highlighted texts as absolute characters whose contexts this repertoire of methods is seeking to reconstruct. This interaction between texts and techniques reveals how the human-technology dyad fluctuates between what is apparent and it being observed computationally.



PUBLISH EXAM RESULTS

SPARE me a space in your esteemed paper to highlight my views about the importance of publishing the EACE and EAACE results.

It is my view that the reason, "shortage of paper", given last year for not publishing the mentioned examination results in the newspapers is not satisfactory. I say this because of the following economic, social and psychological reasons:

Economically, I would point out that the publication of these exam results saves time and money to candidates, parents, friends and

IT'S DIRTY BUSINESS

A common menace

'Unjustified expenditure'

Exorbitant fares

Thank you very much

OUR CIVIL SERVICE IS NOT ALL THAT BAD

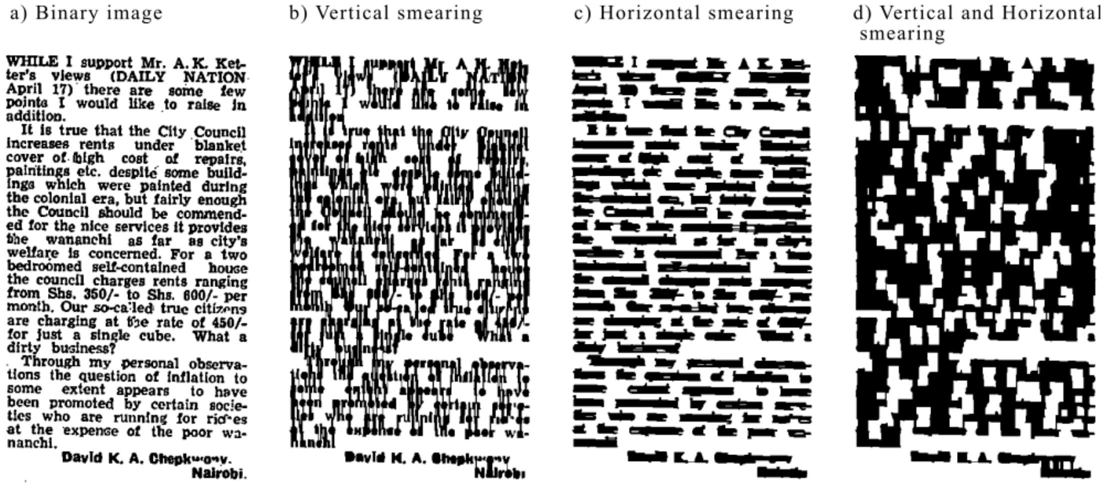
Destructive transfers

Have you seen them?

Figure 4.7: Text image with all identified contours bound within the image. Bounding rectangles around contours approximate the height and width of each contour.

The multi-column structure or Manhattan page layout of the column also revealed a consistent standard discriminant of text blocks. Baird (1992) defines the Manhattan page layout as one in which all text blocks are isolated “by a set of horizontal and vertical straightline-segments drawn through a white space” (2). This structure entails a set of conjectures according to Baird. The first is that white space is a generic layout delimiter and secondly, the background is simpler than the foreground. In addition to these, the different text sizes established above become an ingredient for Run-Length Smearing/Smoothing Algorithm (RLSA), a common algorithm used in image document segmentation. At a low level, the algorithm as advanced by Wong, Casey, & Wahl (1982) was applied to the binary image of black and white pixels based on a predetermined threshold; adjacent pixels are converted to black while those that are already black remain unchanged. When black pixels are “smeared” in this manner, it enables connection between textual characters to form blocks of text.

The bounding boxes described earlier were at the character level; however they did not lend themselves well to semantic segmentation of content blocks. Use of RLSA intensifies and extends text blocks to group them as homogeneous. Vertical smearing yielded better results on the body text image. This is because the body text mask gained extra white space above and below, after removing all titles in the mask at hand. Conversely, the absence of body text in processing the titles mask made horizontal smearing more effective in merging characters that constitute each title. As Figure 4.8b shows, vertical smearing conjoins the three paragraphs as one block of text through merging of adjacent pixels. While horizontal smearing was effective in merging title characters, it was less suited for body text because besides failing to conjoin rows as a unified block as shown in Figure 4.8c, it was also less effective in threading all the characters within a row. Increasing the smearing parameter ran the risk of overlapping adjacent columns in the process, it was important for the solution to allow more control on the direction of the smearing.



WHILE I support Mr. A.K. Ketter's views (DAILY NATION April 17) there are some few points I would like to raise in addition.

It is true that the City Council increases rents under blanket cover of high cost of repairs, paintings etc. despite some buildings which were painted during the colonial era, but fairly enough the Council should be commended for the nice services it provides the wananchi as far as city's welfare is concerned. For a two bedroomed self-contained house the council charges rents ranging from Shs. 350/- to Shs. 600/- per month. Our so-called true citizens are charging at the rate of 450/- for just a single cube. What a dirty business?

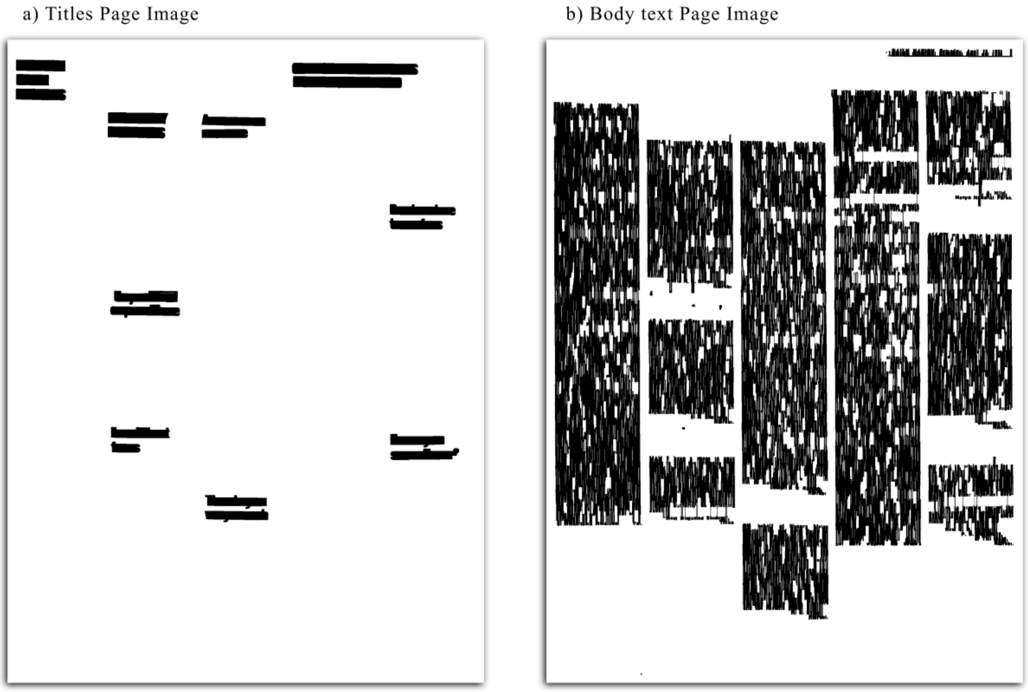
Through my personal observations the question of inflation to some extent appears to have been promoted by certain societies who are running for riches at the expense of the poor wananchi.

David K. A. Chepkwony.
Nairobi.

Figure 4.8: RLSA algorithm applied on a letter snippet in (b) vertical direction, (c) horizontal direction and (d) both vertical and horizontal directions.

Ferilli (2011) discusses at length the hierarchy of logical operations and limitations of RLSA that make the combined approach shown in Figure 4.8d less reliable (pp. 169-171). This was evident in this corpus, there was a high probability of false blocks in cases where column margins are skewed or slanted resulting in nearly pixels to merge incorrectly onto adjacent blocks.

Through control of the orientation (vertical or horizontal) based on Python language solution by Reddy (2019), the result of this *smearing* process as shown in Figure 4.9 for the most part was coherent with the structure of content in the page image, and allowed for a more robust strategy to group regions of interest. The next step sought to reassemble each letter, by matching each body text with its title. This time, contours were blocks of texts rather than absolute characters (as a result of smearing processes). Given the new state of contours, three emerging approaches were considered for logical segmentation of titles and body texts, based top-down left-right strategies.



a) Titles Page Image

b) Body text Page Image

Figure 4.9: Two images represent "titles" and "body texts" on separate masks. Applying horizontal smearing on contours with taller contours and vertical smearing on contours with average height establishes distinct blocks.

This approach of unifying two sets of texts concurred with the layout and organization of content on the page. The final method implemented to combine the texts can be summarized in the following pseudo-code, implemented recursively:

1. Sort the title and body text contours in a top-down, left-right fashion.
2. Collect all blocks of body text occurring sequentially before the next title and below the current processing title.
3. Where body text spans multiple columns, consider title width as the boundary guide and the next title in sequence as the limit for content to collect.
4. Use geometric properties of the contours to determine coordinates to clip on the original page image.

The geometric properties as shown in Figure 4.10a were coarse segments with irregular height used to extract regions of interest from the original image (see Figure 4.10b). Mimicking the top-to-down, left-to-right order of English reading to computationally attach content to respective titles.

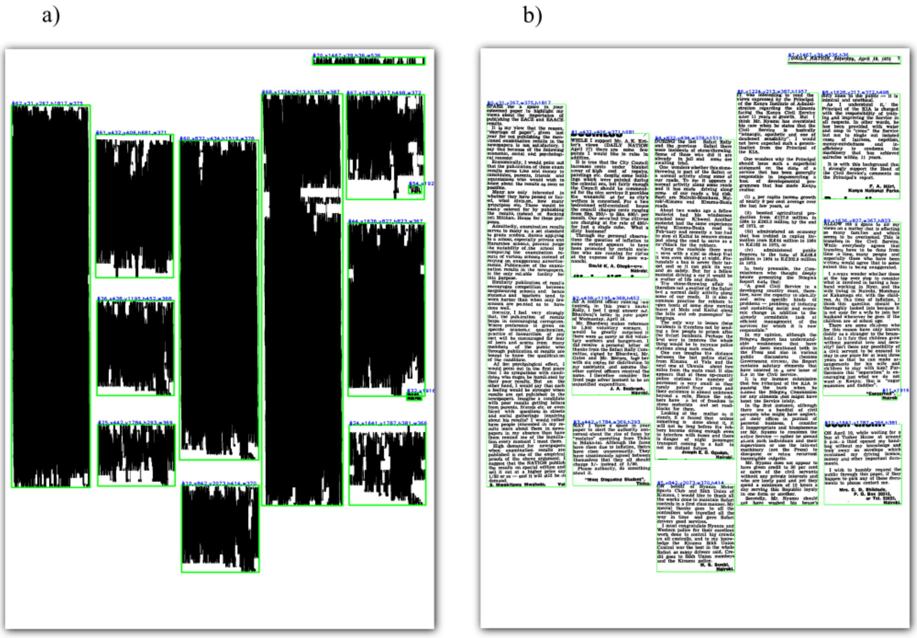


Figure 4.10: RLSA applied to body text mask (a) with more aggressive vertical smearing to demarcate individual letter regions. Extracted texts from the original image (b). Contour number, coordinates and measurements of the contour are labeled on each bounding box.

From the recursive process, the result were blocks of body text with spaces in between for titles. A similar process was performed to extract a title. At any rate, the goal was to collect individual letters and thus the final step in this image analysis phase was to assemble the two logical blocks of title and body text. Each letter was thus reconstructed as a new image, which was input for automatic transcription process that is OCR-ing.

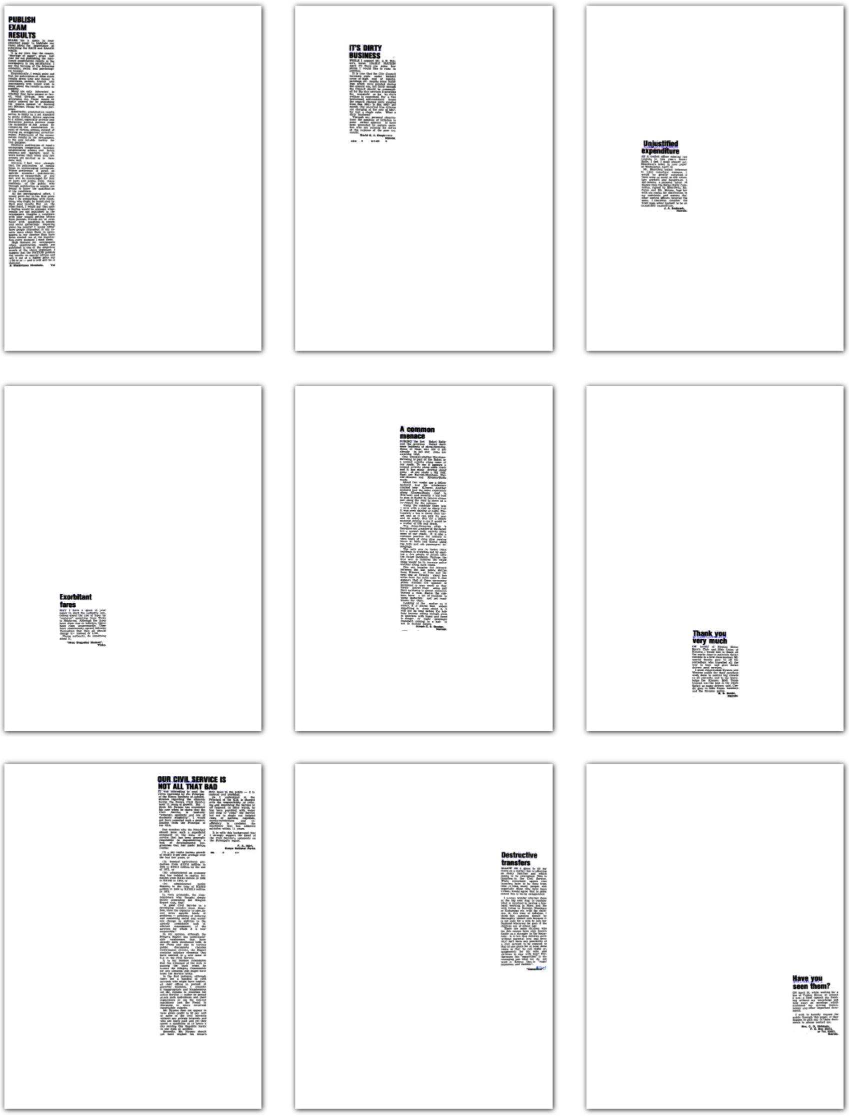


Figure 4.11: Individual letters as separate page images after the segmentation process and subsequent reassembly.

4.4.1 Alternative Methods

As mentioned earlier, there were two prior approaches implemented for the final step to reassemble contents logically. Although they significantly influenced the final method, they overreached in trying to address the subtle intricacies of merging contours. It is useful to shed light on the pitfalls of those attempts before proceeding to the methods of textual analysis.

The first top-down approach mimicked the original image columns. Title contours controlled the sequence of steps to determine the beginning and end of a letter. By drawing vertical “virtual lines” running along the image, body texts were collected recursively unless: a) the *following* title in sequence was encountered; or b) the next line was greater than title width. While this strategy was seemingly sound at first, it proved unreliable in practice. Straight lines were at any given time uncoordinated with the position of both titles and body text contours. Title and body text were likely misaligned or overlapped by a few pixels, a scenario that was noticeable in initial sample images. It became clear a less rigid and rather unsystematic approach was needed to attend to such inconsistent patterns.

The second strategy was more rudimentary and was based on several assumptions. Still controlled by titles, it was based on two principles obtained from close observation: that every subsequent title block (see Figure 4.10a) is located below the current title, or in the next column. What it failed to consider was the different permutations available for body text arrangement. As the processing moved from left-to-right and top-to-bottom, the rules revealed complexities of extracting multi-column letters, variable lengths, and handling of minute misalignments of title and body text components. The goal of rule-based approach was to maintain a reasonable degree of rules; therefore, as more rules became necessary it also signaled the need for a new simplified strategy – one that was less wieldy and more comprehensible. One advantage of this approach is that its principles, with a few modifications, became foundational to the final method.

4.5 Methods of Text Analysis

While research on page layout analysis in OCR has for long devised solutions for heterogeneous documents (Chen, Yin, & Liu, 2013; Hung-Ming Sun, 2005; Shafait & Smith, 2010), in practice there is still a gap between detecting the layout and extracting individual texts. Moreover the application of column layout analysis has not always translated well in application for text retrieval. By this, I mean the logical layout analysis within heterogeneous text regions in order to find and retrieve articles–or letters. As Ray Smith (2009) observed, physical page layout analysis is essential to how OCR engines such as Tesseract–a popular open-source library for OCR–process images of arbitrary texts and non-texts. The preceding methods of segmentation using OpenCV can be said to have addressed not only the physical layout, but also the logical layout of content.

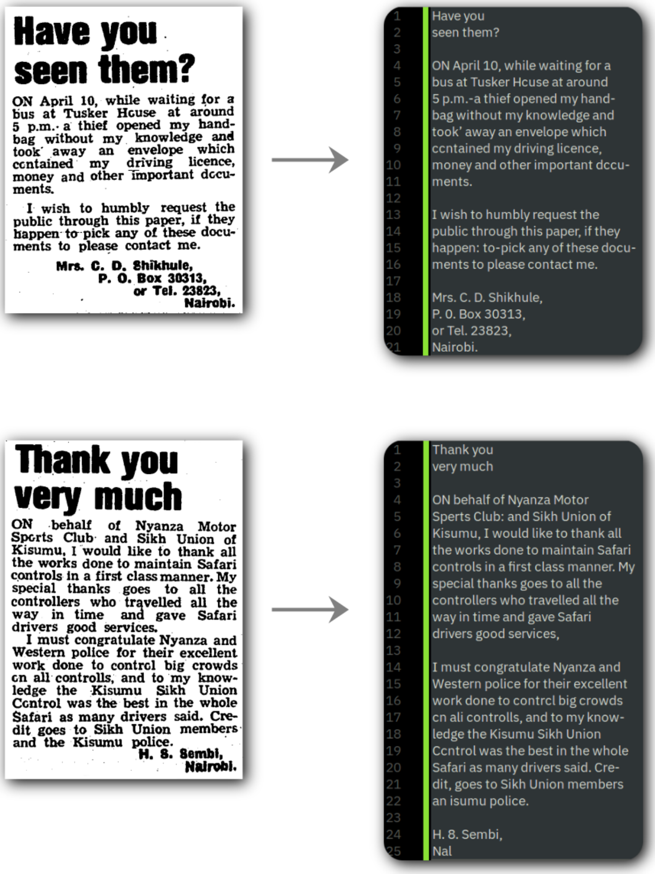
Logical layout analysis or article segmentation was also lacking in Tesseract and could not adequately attend to letter components: title, body text and byline. While the version of Tesseract[23] at the time of this research used neural networks and provided deep-learning models, it remained inadequate at logical segmentation of texts from the results obtained. Further post-processing was required on the OCR texts to reestablish titles, body texts and other meaningful components such as dates, bylines, and inter-referencing. The final script[24] combined both image analysis and OCR-ing into a single pipeline whose output was both a plain text file and an XML (eXtensible Markup Language) file, with structural metadata obtained in the segmentation. Though the OCR-generated errors in the output were for the most part attributable to the quality of the original images, the plain text output retained the textual layout. Such as line breaks and paragraphs.

However, not all textual features are persisted by the OCR. For example, the large font-size and bold typeface that are readily identified were lost, posing the challenges anew, only now in different contexts. Shown in Figure 4.12, this meant that assumptions similar to those used in the segmentation process were necessary. I relied on the paragraphs as a discriminant between title and body text. Text

[23] Tesseract 4.0 with LSTM https://github.com/tesseract-ocr/tesseract/wiki/4.0-with-LSTM

[24] https://github.com/ooduor/lettersiterate/blob/master/main_steps.py

occurring before the first paragraph would be considered as title. Text following the last newline and empty line would be considered the byline component of the letter.



Have you seen them?

ON April 10, while waiting for a bus at Tusker Hcuse at around 5 p.m.·a thief opened my handbag without my knowledge and took' away an envelope which ccntained my driving licence, money and other important dccuments.

I wish to humbly request the public through this paper, if they happen to pick any of these documents to please contact me.

Mrs. C. D. Shikhule,
P. O. Box 30313,
or Tel. 23823,
Nairobi.

```
Have you
seen them?

ON April 10, while waiting for a
bus at Tusker Hcuse at around
5 p.m.-a thief opened my hand-
bag without my knowledge and
took' away an envelope which
ccntained my driving licence,
money and other important dccu-
ments.

I wish to humbly request the
public through this paper, if they
happen: to-pick any of these docu-
ments to please contact me.

Mrs. C. D. Shikhule,
P. 0. Box 30313,
or Tel. 23823,
Nairobi.
```

Thank you very much

ON behalf of Nyanza Motor Spcrts Club and Sikh Union of Kisumu, I would like to thank all the works done to maintain Safari controls in a first class manner. My special thanks goes to all the controllers who travelled all the way in time and gave Safari drivers good services.

I must congratulate Nyanza and Western police for their excellent work done to contrcl big crowds cn all controlls, and to my knowledge the Kisumu Sikh Union Ccntrol was the best in the whole Safari as many drivers said. Credit goes to Sikh Union members and the Kisumu police.

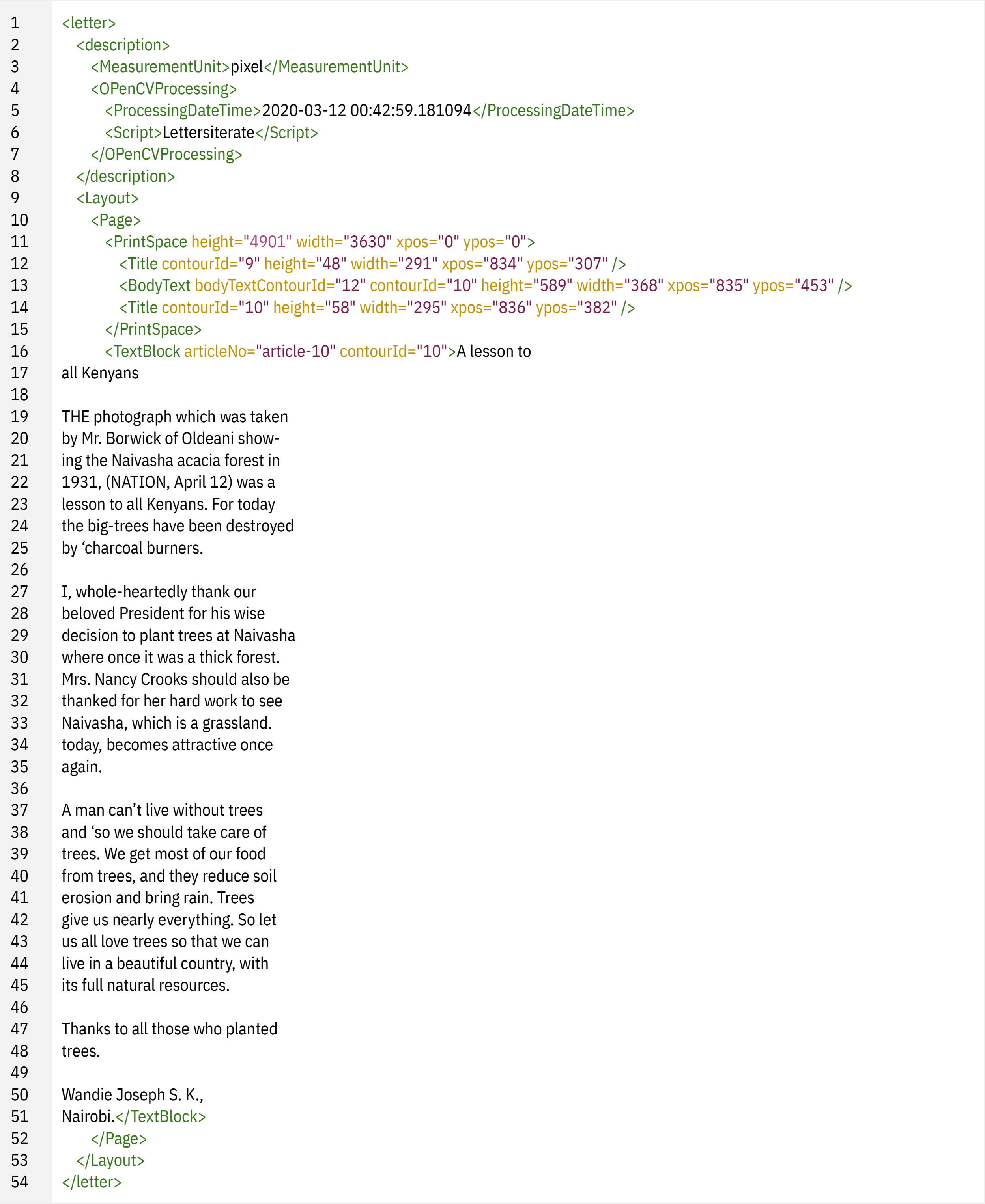
H. S. Sembi,
Nairobi.

```
Thank you
very much

ON behalf of Nyanza Motor
Sperts Club: and Sikh Union of
Kisumu, I would like to thank all
the works done to maintain Safari
controls in a first class manner. My
special thanks goes to all the
controllers who travelled all the
way in time and gave Safari
drivers good services,

I must congratulate Nyanza and
Western police for their excellent
work done to contrcl big crowds
cn ali controlls, and to my know-
ledge the Kisumu Sikh Union
Ccntrol was the best in the whole
Safari as many drivers said. Cre-
dit, goes to Sikh Union members
an isumu police.

H. 8. Sembi,
Nal
```

Figure 4.12: Clipped letters alongside OCR output in a text editor. As the textual topology shows, only line breaks and paragraph structures were persisted by the OCR process.

As mentioned earlier, two text files were generated for each letter. One was plain text file as shown in Figure 4.12, and the other was structured as a custom XML document. The rationale for serializing in two formats was because plain text files are not only easy to write and read compared to XML, their contents are also quicker to parse and therefore attract lower processing costs with respect to the volume of documents at hand. As such, the bulk of subsequent text analysis was done on the plain text version of the letters. The XML version was crucial insofar as it stored geometric properties during the segmentation stage that would be required to refigure a letter along with the page image on an interface. A custom XML was adopted since the desired components extracted did not fully comply with the specifications of current standards such as ALTO XML or METS.[25] The letters serialized in plain text format had file extension *.txt* and *.xml* for the custom XML format. Figure 4.13 shows a number of elements and attributes stored in the custom XML format in addition to the actual letter content.

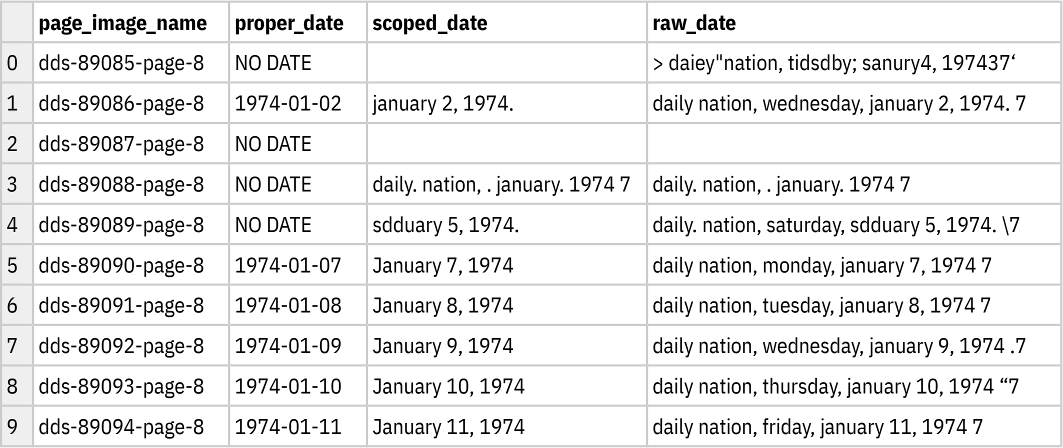
[25] Analyzed Layout and Text Object (ALTO) is an XML format describing layout and content on digitized resources such as newspapers. ALTO XML as a schema is used to describe technical metadata but is often paired with Metadata Encoding and Transmission Schema (METS), another XML schema which is concerned with the management, administration, and transmission of digital objects between repositories.



```
1  <letter>
2    <description>
3      <MeasurementUnit>pixel</MeasurementUnit>
4      <OPenCVProcessing>
5        <ProcessingDateTime>2020-03-12 00:42:59.181094</ProcessingDateTime>
6        <Script>Lettersiterate</Script>
7      </OPenCVProcessing>
8    </description>
9    <Layout>
10     <Page>
11       <PrintSpace height="4901" width="3630" xpos="0" ypos="0">
12         <Title contourId="9" height="48" width="291" xpos="834" ypos="307" />
13         <BodyText bodyTextContourId="12" contourId="10" height="589" width="368" xpos="835" ypos="453" />
14         <Title contourId="10" height="58" width="295" xpos="836" ypos="382" />
15       </PrintSpace>
16       <TextBlock articleNo="article-10" contourId="10">A lesson to
17 all Kenyans
18 
19 THE photograph which was taken
20 by Mr. Borwick of Oldeani show-
21 ing the Naivasha acacia forest in
22 1931, (NATION, April 12) was a
23 lesson to all Kenyans. For today
24 the big-trees have been destroyed
25 by 'charcoal burners.
26 
27 I, whole-heartedly thank our
28 beloved President for his wise
29 decision to plant trees at Naivasha
30 where once it was a thick forest.
31 Mrs. Nancy Crooks should also be
32 thanked for her hard work to see
33 Naivasha, which is a grassland.
34 today, becomes attractive once
35 again.
36 
37 A man can't live without trees
38 and 'so we should take care of
39 trees. We get most of our food
40 from trees, and they reduce soil
41 erosion and bring rain. Trees
42 give us nearly everything. So let
43 us all love trees so that we can
44 live in a beautiful country, with
45 its full natural resources.
46 
47 Thanks to all those who planted
48 trees.
49 
50 Wandie Joseph S. K.,
51 Nairobi.</TextBlock>
52     </Page>
53   </Layout>
54 </letter>
```

Figure 4.13: Custom XML document. Showing coordinates and geometric parameters of body text and title, on lines 12 - 14.

Both files shared the same naming format which was derived from the contour number during image processing. For example, *dds-89807-page-8-article-10.txt* refers to an extracted letter number 10. The other parts to the left of the file name originated from the actual digitized file names. Together they formed a unique identifier for each letter on the subsequent stages of analysis. The original file names were also helpful in determining the date of each issue in the same manner of processing as the letters themselves. This was a separate Python script[26] to read only in the top-right corner of the page image to extract the raw date. Once OCR was performed, it was associated with the file name and parsed into a standard date format. A tab separate values (TSV) file was generated for each year under processing. For the most part, these values were automatically generated with a few cases requiring correction by hand. As Table 4.1 shows, heuristic approaches informed by the state of extracted texts helped design the algorithms to scope the relevant date characters.



	page_image_name	proper_date	scoped_date	raw_date
0	dds-89085-page-8	NO DATE		> daiey"nation, tidsdby; sanury4, 197437‘
1	dds-89086-page-8	1974-01-02	january 2, 1974.	daily nation, wednesday, january 2, 1974. 7
2	dds-89087-page-8	NO DATE		
3	dds-89088-page-8	NO DATE	daily. nation, . january. 1974 7	daily. nation, . january. 1974 7
4	dds-89089-page-8	NO DATE	sdduary 5, 1974.	daily. nation, saturday, sdduary 5, 1974. \7
5	dds-89090-page-8	1974-01-07	January 7, 1974	daily nation, monday, january 7, 1974 7
6	dds-89091-page-8	1974-01-08	January 8, 1974	daily nation, tuesday, january 8, 1974 7
7	dds-89092-page-8	1974-01-09	January 9, 1974	daily nation, wednesday, january 9, 1974 .7
8	dds-89093-page-8	1974-01-10	January 10, 1974	daily nation, thursday, january 10, 1974 “7
9	dds-89094-page-8	1974-01-11	January 11, 1974	daily nation, friday, january 11, 1974 7

Table 4.1: A section of 1974 TSV file showing the page image processed in the first column and in the second column, the proper date where parsable. The third and fourth columns show intermediate stages of processing the OCR output.

Through this step, and the naming format used on page images, it was possible to ascertain the date each letter was published in the context of dated references across the corpus.

As foregrounded in the methodology section, the OCR output required post-processing steps in an attempt to recreate the logical structure that is lost in transcription. Each text file was thus parsed to distinguish between the title, body text and byline components in the file.

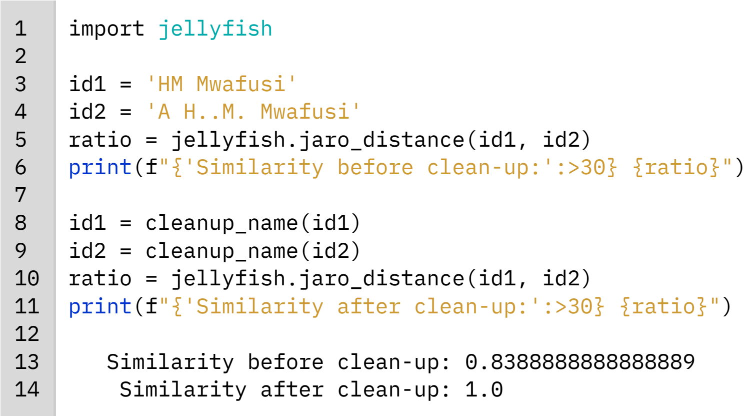
[26] Python script this process can be found at: https://github.com/ooduor/lettersiterate/blob/master/column_dates.py (Antony, “Extract Date of Publication from a Newspaper Page Image”).

Coding had to rely on the first paragraph to obtain the title, and on the last paragraph, byline. Similar stylistic approaches in automated abstraction have been discussed by (Cleveland & Cleveland, 2001, p. 214). However, although extracting titles in this manner was less complicated, bylines were registered in several formats: from multi- to single-line paragraphs.

On one end, bylines included the author's proper name, professional capacity, organization and location, and on the other end, simply a given name or identity constituted a byline. Exceptions were made in the algorithm to handle cases with a limited number of paragraphs or bylines that were longer than was expected. Such exceptions served a couple of important purposes. One, they were evaluative of the process thus far, which, depended on successful determination of components. Secondly, they were also indicative of the general state of the digitized collection and OCR state-of-the-art. In other words, the final number of titles and bylines collected pointed to the condition of page images, OCR used, or both. This would in turn impact the volume of interpretable documents as well as have implications on access strategies.

Tab separated records stored the extracted titles for each letter, categorized by year. In the case of bylines, a further attempt was made to distinguish the various components with the goal of establishing the name of the author, and where permissible, location and organization. The elusive challenge of separating byline pieces beyond stylistic elements; that is, commas and number of lines, was more pronounced by the need to accurately recognize contributor names. The three main reasons were: 1) a significant number of named entities were misspelled; 2) the localized nature of terminology used by correspondents, and 3) literal namelessness. The last observation can be attributed to editorial powers of ascribing pseudonymous and anonymous tags in place of proper names. The column, as a result featured nameless letters from “Disgusted Phone-User”, “Historian”, “Concerned teacher”, “Petrol Retailer” as proof of the editor's burden to describe the writer / author with the authorial topic. Of course, one speculation is that some contributors would wish to have their letters published anonymously for whatever reason.

In the end, the ability to collect letters from named authors would rely more on approximate and phonetic similarities of named strings than exactly defined. In order to optimize the Python library used to approximate names; dot and space characters, in addition to single-letter words were stripped as part of the text-cleaning process. Figure 4.14 shows the results of this extra step in lines 13 and 14. The library was based on a scale of 0.00 (completely dissimilar) to 1.00 (strong similarity) and in this case, the two identity strings, while already similar, were absolutely identical strings after additional cleaning. At the corpus level, a threshold of 0.84 was chosen to collate similar identities.[27] Due to the exploratory nature of the research, collated identities would themselves be considered alternative labels / terms, as a way of establishing textual relationships by use of bylines. See Appendix C for a visualization of one such identity.



```
1   import jellyfish
2
3   id1 = 'HM Mwafusi'
4   id2 = 'A H..M. Mwafusi'
5   ratio = jellyfish.jaro_distance(id1, id2)
6   print(f"{'Similarity before clean-up:':>30} {ratio}")
7
8   id1 = cleanup_name(id1)
9   id2 = cleanup_name(id2)
10  ratio = jellyfish.jaro_distance(id1, id2)
11  print(f"{'Similarity after clean-up:':>30} {ratio}")
12
13     Similarity before clean-up: 0.8388888888888889
14      Similarity after clean-up: 1.0
```

Figure 4.14: Code snippet comparing identities from two different letters (lines 3 and 4) and how similarity was optimized. Used *jellyfish*, a Python library with a variety of string matching and approximation algorithms.

The body texts were a trove of dates that often referenced particular events, editorials, or letters which were interpreted as a trail of correspondence. The algorithm for finding date-like pieces from body text was remarkably stable despite a lack of structure or convention seen earlier when parsing column dates. This was done through the use of regular expressions

[27] The Jupyter Notebook for this process can be found at: https://github.com/ooduor/lettersiterate/blob/master/notebooks/cluster_similar_reader_names_jellyfish.ipynb (Antony, "Byline Similarities of names using jellyfish").

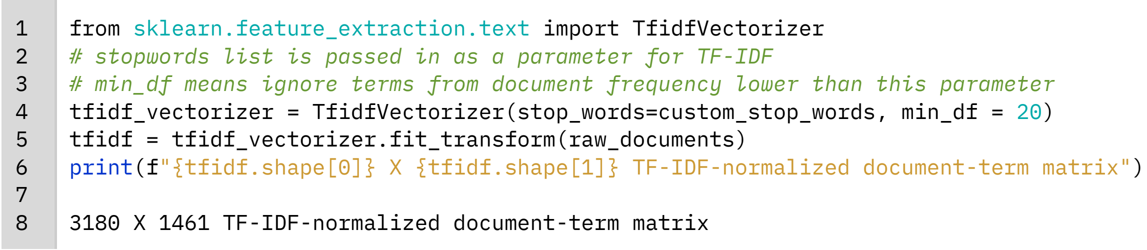
(a.k.a regex) to find date-like strings and parse them as such. In addition, having previously established dates each column was published meant that incomplete dates found in the letter could be automatically inferred, especially when “year” was missing. An attempt was made to determine whether the extracted date was likely corresponding to an earlier letter. This was achieved by searching a specific set of keywords in the letter, a higher occurrence meant it was likely a correspondence than not. A date found together with any instance of “correspond”, “comment”, “comments in”, “column”, “daily nation reader”, “letter by”, “letter headed”, “mailbox”, “reply”, “referring to”, “support”, “suggested by”, and “written by” was deemed a correspondence. Four other auxiliary categories were defined besides correspondence; these were editor, editorial, article and miscellaneous.[28] Editor and editorial categories referred to letters addressing the editor and responding to an editorial, respectively. Dates extracted in the article category made reference to “article” or “reporter.” Those marked as miscellaneous likely referred to events or narratives not covered in the newspaper. See Appendix C for a full list of categories and keywords.

4.6 Methods of Subject Indexing

Collectively, the previous methods of text wrangling and analyses largely formed the basis of collecting metadata as well as connecting otherwise unrelated documents. As short text passages, finding topic(s) of essence for each document letter was based on the thematic structure of keywords that often co-occur in the corpus. Known as topic modeling, it is based on the discovery of latent themes recurring in a large collection of texts (Blei, 2012; Ignatow & Mihalcea, 2017). Using this model, a topic was described by the set of meaningful and coherent terms found across the corpus. Unlike subject index terms that comprise a controlled vocabulary, topic terms are derived from the collection under analysis by statistical models, and therefore elusive to forms of organization and subsequent retrieval of documents.

[28] The Jupyter Notebook for this process can be found at: https://github.com/ooduor/lettersiterate/blob/master/notebooks/trace_dates_guided_by_keywords.ipynb (Antony, “Trace Dates after Narrowing down on keywords”).

While topic modeling by definition was useful in discovering the hidden thematic structure in a corpus, it was inversely capable of finding individual documents in the corpus that relate to particular topics. However, arriving at topical terms and potentially coherent topics involved a number of steps. A list of stop-words was created. This custom list consisted of common English words, and additional words such as “letter”, “space”, “kenya”, “nation” and many others which were overused in the column. This list was an additional input parameter for TF-IDF implementation by Pedregosa et al. (2011) on Scikit-learn[29] machine learning library. TF-IDF technique converted raw document texts into a document-term matrix. Because of the bag-of-words representation, the resulting document-term matrix consisted of rows of documents against columns of unique terms with each cell denoting the number of occurrences registered.



```
1  from sklearn.feature_extraction.text import TfidfVectorizer
2  # stopwords list is passed in as a parameter for TF-IDF
3  # min_df means ignore terms from document frequency lower than this parameter
4  tfidf_vectorizer = TfidfVectorizer(stop_words=custom_stop_words, min_df = 20)
5  tfidf = tfidf_vectorizer.fit_transform(raw_documents)
6  print(f"{tfidf.shape[0]} X {tfidf.shape[1]} TF-IDF-normalized document-term matrix")
7
8  3180 X 1461 TF-IDF-normalized document-term matrix
```

Figure 4.15: Code snippet showing summary results of TF-IDF implementation. The *TfidfVectorizer* function generates a matrix of 3180 documents against 1461 terms (line 8).

With the generated document-term matrix, the next step not only worked with the top terms after filtering out weighting and relative frequencies, it also optimized the coherence degree of each topic. One of the foremost approaches of topic modeling works by decomposing the document-term matrix further into two matrices. The implication of this decomposition, without delving into the formal mathematical complexities, is that documents and terms can be analyzed separately against topics. Document-topics matrix represented number of topics per document. Topic-terms matrix, however, represented the weight of terms distributed across the topics. This branched matrix approach is known as Non-negative Matrix

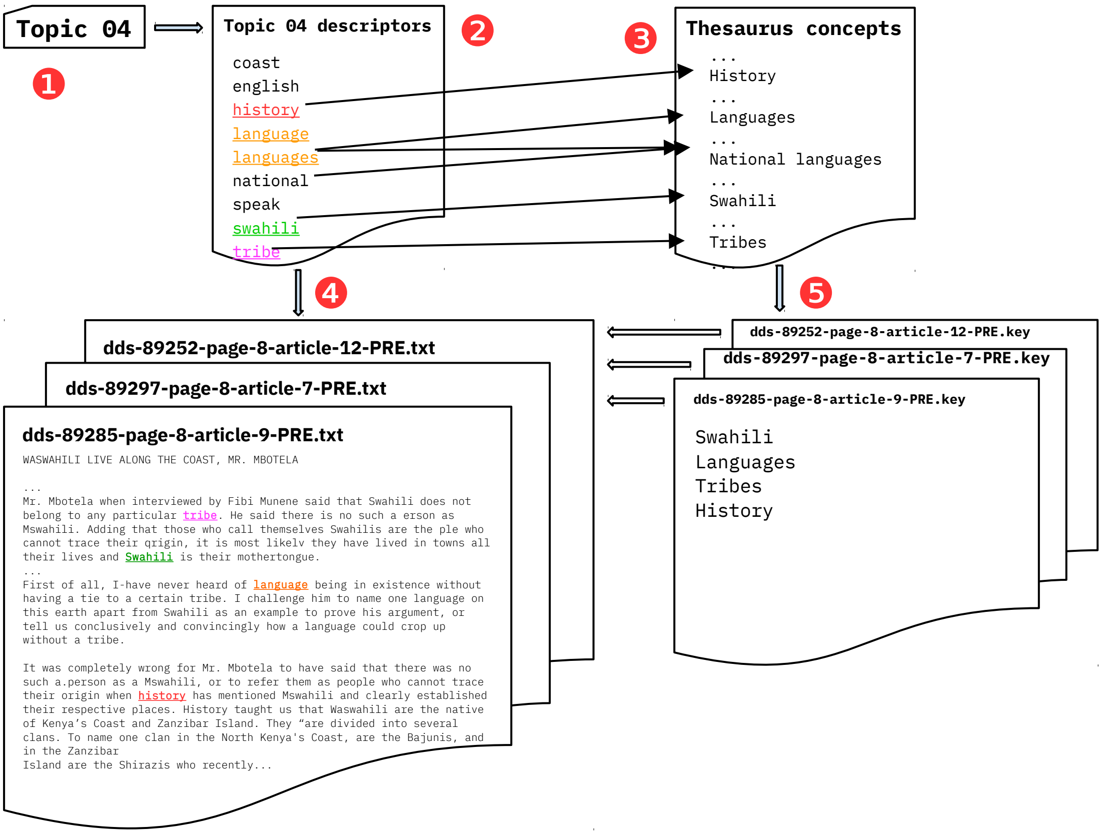
[29] Scikit-learn is a Python language package that implements a suite of machine-learning algorithms.

Factorization (NMF). Besides having "semantic interpretability" (O'Callaghan et al., 2015, para. 1) and more coherent topics compared to other topic modeling approaches, NMF was better suited for learning topics from short texts (Chen et al., 2019) and non-mainstream domains (O'Callaghan et al., 2015).

Even though topic descriptors were derived directly from documents, maintaining topic coherence required limiting the number of topics and descriptors. Therefore only the salient and coherent topics were matched with corresponding terms in the controlled vocabulary. For each topic, the set of descriptors that mapped with the vocabulary was further filtered against documents associated with the topic to narrow the literal term specificity. Each plain text file (a.k.a *.txt*) with salient topic(s) had a corresponding *.key* file which contained vocabulary terms which best described the subject of its content.[30] In Figure 4.16 below, "Topic 04" – which may be interpreted as discussions on the origins of Swahili language and its status as a national language – had descriptors coinciding with thesaurus terms. Therefore, each document associated with this topic was not simply assigned all relevant thesaurus terms, but only specific terms. The paired sets formed the training and evaluation dataset on which other documents, including those with insufficient optimum topic coherence, would be assigned terms.

The final step involved training the dataset using Annif. Training and testing datasets were divided by year. Since the above steps were performed in yearly batches, odd years (1975, 1977) were used for training and even years (1974, 1976, 1978) as both evaluation and testing datasets. The downside for this strategy was that since the nature of topics shifted over the years, there was potential of the evaluation and test dataset being unrelated and therefore mismatched. Still, a significant number of topics recurred over the years.

[30] The Jupyter Notebook for this process can be found at: https://github.com/ooduor/lettersiterate/blob/master/notebooks/topic_modeling_nmf.ipynb (Antony, "Topic Model using NMF").


Topic 04
1
Topic 04 descriptors
2
coast
english
history
language
languages
national
speak
swahili
tribe
3
Thesaurus concepts
...
History
...
Languages
...
National languages
...
Swahili
...
Tribes
...
4
5
dds-89252-page-8-article-12-PRE.txt
dds-89297-page-8-article-7-PRE.txt
dds-89285-page-8-article-9-PRE.txt
dds-89252-page-8-article-12-PRE.key
dds-89297-page-8-article-7-PRE.key
dds-89285-page-8-article-9-PRE.key
Swahili
Languages
Tribes
History

```
WASWAHILI LIVE ALONG THE COAST, MR. MBOTELA

...
Mr. Mbotela when interviewed by Fibi Munene said that Swahili does not
belong to any particular tribe. He said there is no such a erson as
Mswahili. Adding that those who call themselves Swahilis are the ple who
cannot trace their qrigin, it is most likelv they have lived in towns all
their lives and Swahili is their mothertongue.
...
First of all, I-have never heard of language being in existence without
having a tie to a certain tribe. I challenge him to name one language on
this earth apart from Swahili as an example to prove his argument, or
tell us conclusively and convincingly how a language could crop up
without a tribe.

It was completely wrong for Mr. Mbotela to have said that there was no
such a.person as a Mswahili, or to refer them as people who cannot trace
their origin when history has mentioned Mswahili and clearly established
their respective places. History taught us that Waswahili are the native
of Kenya’s Coast and Zanzibar Island. They “are divided into several
clans. To name one clan in the North Kenya's Coast, are the Bajunis, and
in the Zanzibar
Island are the Shirazis who recently...
```

Figure 4.16: Flowchart showing how topic descriptors are mapped onto thesaurus terms. Instead of assigning thesaurus terms to all documents under this topic, only those most relevant to the document are considered for the training phase.

Two important text normalization steps were omitted. One was stemming, which crudely shortened words at the *root* of inflection. Therefore words like “universe” and “university” would share the stem “univers” – often not a valid root. The second was lemmatization, which unlike stemming, considered the meaning of the word. Thus, occurrences of “universe” or “university” would remain the same in their lemma form. Both steps were bypassed because stemmed terms like “univers” would be ambiguously interpreted against the controlled terms when “university students”, for instance, was the appropriate concept in the thesaurus, instead of “universe”. The benefit of omitting these steps was the ability to generate bigrams from the topic descriptors, which in this context represented co-occurring terms such as “civil servants”, “student teachers”, “national language”, among others. Even more useful is how this process combined topic descriptors into possible composite subject terms for purposes of coordinate indexing when mapping to thesaurus concepts.

4.7 Conclusion

This chapter began by introducing the multifaceted enterprise of exploring digitized collections as documents of interest. Working with the digitized documents meant acknowledging the fraught state of locating and collecting source data at the onset of this research. I outlined the overarching methodology in three stages. The image analysis stage performed physical segmentation of the *Mailbox* column to extract individual letters sent to the editor. Next, image transcription not only involved OCR-ing the segmented images, it also analyzed the logical layout to obtain meaningful components of each letter. Metadata with respect to locating the letter on the physical layout was also collected. In the last stage, the discussion focused on implementation of subject indexing based on topic modeling.

Progressively, each stage sought to address the research questions. And in their respective method sections; I engaged page layout analysis techniques, which demanded clarity of instructions to reconcile the human–computer perception of images. This was followed image by segmentation whose output was used in the OCR process thereafter. Finally, the extracted topics formed the basis on which text documents were assigned thesaurus terms in order to index the larger collection. The chapter laid the foundation for locating digitized texts and more, lent itself fully to indexing practices. Throughout the chapter, attempts were made to articulate the challenges of remediating the digital archive.

Chapter 5

5 Results and Application

5.1 Introduction

In most cases the OCR results were truncated texts due to failed segmentation or limitations of the OCR processes. However, there was also a notable result set that I focus on in the following sections as I examine the outcome of Chapter 4.

This chapter has three sections which discuss the application of results reimagined under bibliographic forms of control and organization. It highlights characteristics of indexing and abstracting in the literature and how they are applied on the letters column. While (Mulvany, 1994) has argued (regarding conventional indexing) that chapter and book indexes are largely catalogued to be specific and exhaustive for retrieval purposes (p. 49), I demonstrate that for letters as a special format, the access needs and used terminology proffers these qualities for automated indexing. A closer look at some of the results showed that such assigned terms not only reflected the topic of the letter at surface level, but added breadth on the subject. This served to highlight broader topics of discussion in the letter as well as become a link to other letters–creating an affordance for intra-column references. Beyond the inherently small size of letters compared to chapters and books, the application of inter-referencing, surrogation and what constitutes indexable matter is also demonstrated from the results.

5.2 Application of Results

Individual results from the methodology and methods section were fragmented datasets; representing topics in the form of thesaurus subjects, authors as byline identities, quoted phrases and date references as gateways to topical narratives over time. They were for the most part stored in semi-structured files. Pandas[31], a Python library, was used to query the TSV files albeit in a rudimentary

[31] https://pandas.pydata.org/

manner due to inability to establish relationships between files. The application interfaces were built on Jupyter Notebooks[32] environment and while they can be rendered fully with hyperlinks, they cannot fully relay the web of linkages as they would on a web server. Besides the ability to analyze and prototype with such datasets, Notebooks inter alia, are unwieldy to effectively work with conventional databases or interlink resources internally to navigate the collection. But they still allow sufficient interaction to explore the level of digitization I have argued for. Though these limitations could potentially understate the results, Notebooks in this case create the possibility to prototype with basic results as well as share analysis and documentation without copyright infringement.

Other generated formats previously mentioned in the methodology (plain text and XML) are also part of the Notebooks applications in addition to the page image itself. It is worth mentioning that each letter in the corpus was assigned an optimum number of four vocabulary terms from the thesaurus, which were embedded in the XML document. Accordingly, during the same indexing process each thesaurus term had two attributes; a score value and a URI[33] to indicate the degree of relevance to the topic and point to the concept in the thesaurus, respectively. A number of interfaces were developed in the Notebooks environment to simulate user queries via subject vocabularies, author names, and date references.

Page images feature across these hypermediated interfaces both as respectful remediation of texts and as a reminder of the medium. *Respectful* here follows from Bolter and Grusin’s (1999) notion of respectful remediation where “venerable” media includes printed books, painting and photographs are remediated without critique of their original form. For instance, they argue the purpose of Project Gutenberg[34] “is to collect pure verbal versions of the ‘classic’ texts” (p. 201). Conversely, not acknowledging the medium being remediated is a form of radical remediation. The interfaces in the following sections exercise the respectful attitude only outside the limits of “classic” texts, they are

[32] https://jupyter.org/

[33] URI is a sequence of characters in a defined scheme to locate a resource on the internet. URI stands for Uniform Resource Identifier, in this case the URIs were in the form of hyperlinks back to the online UNESCO Thesaurus.

[34] Project Gutenberg is a digital library and repository of digitized public domain books. As the preeminent digital library, its collection primarily consists of classical Western literature that has been digitized and available freely not only as full text but in a number of open publication formats. See https://www.gutenberg.org/.

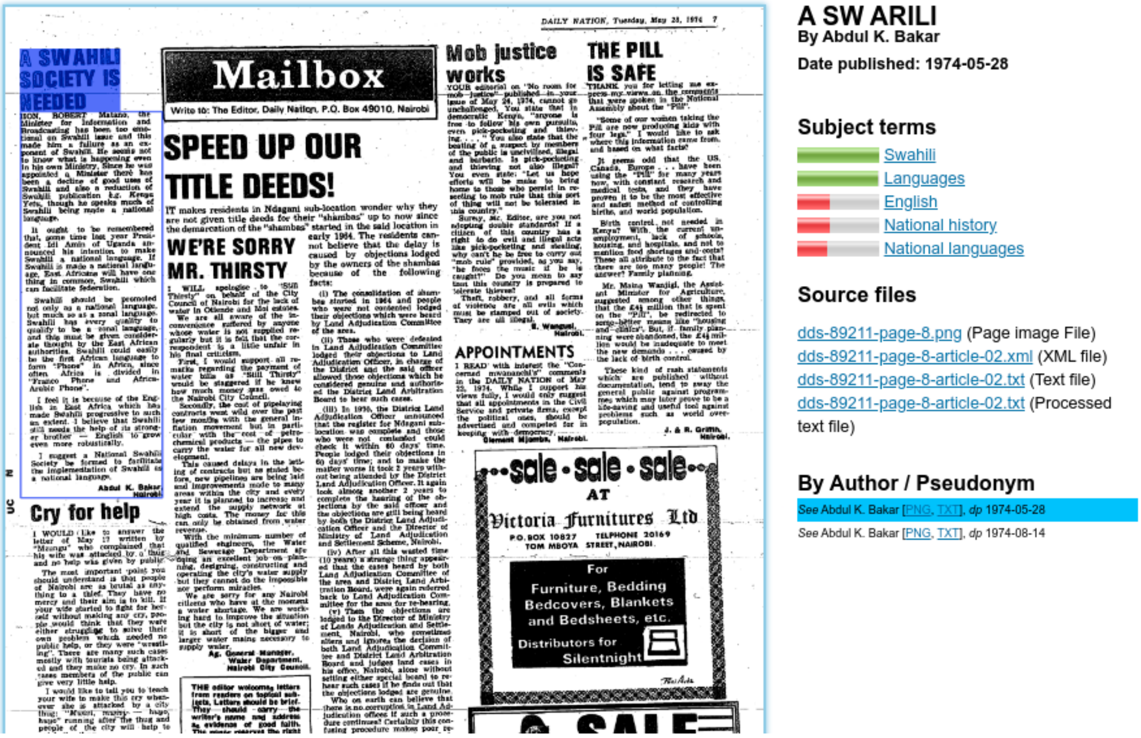
thus respectful in attempting to reproduce the text more so in multiple formats (plain text, XML, PNG) and represent the page image as a claim to immediacy. The interfaces illustrate documents that are located contextually within a page as well as authoritatively through the use of thesaurus vocabulary among other access points.

5.2.1 Subject Vocabularies

In the first approach, the Notebook prompts the user to enter a thesaurus term of interest for inquiry. Based on term assignments and scoring, the page image is returned accompanied by metadata. Each *Mailbox* letter therefore is accessible not only by the assigned terms but also by their degree of relevance to the subject of interest. Examples of these representations are shown in Figures 5.1 – 5.3 with the Jupyter Notebook that generated them included in Appendix B. Other configurable parameters allow filtering by year of publication, and a threshold which can be set before performing the inquiry. I chose to search by one of the most engaging topic term generated over the years: "Swahili". In this instance, there were 33 *Mailbox* letters published in 1974 which had Swahili language as the main topic of discussion. Though the full-page image is rendered, the letter of interest is highlighted around the dimensions of its title and body text. Beside the image is a side panel which refigures content as a bibliographic object. First, the title, name of the author and date of publication are made legible for attribution. Secondly, multiple specific entries of subject terms describe the object using the UNESCO Thesaurus. The scores assigned to each term are visualized on a color-coded meter showing the degree of specificity to the *Mailbox* letter.

Though not a claim exhaustiveness, the list of assigned thesaurus terms on the side panel is a good indicator of the topics discussed, however, the claim to specificity is exerted through the use of a scoring meter where each term's degree of generality can be considered. In addition, the hyperlinked subject terms point to similar letters in the *Mailbox* column as well as informs the user of assigned concepts from the thesaurus.

```
Enter a term in the thesaurus:Swahili
Searching for term starting with: Swahili
```



A SWAHILI SOCIETY IS NEEDED

HON. ROBERT Matano, the Minister for Information and Broadcasting has been too emotional on Swahili issue and this made him a failure as an exponent of Swahili. He seems not to know what is happening even in his own Ministry. Since he was appointed a Minister there has been a decline of good uses of Swahili and also a reduction of Swahili publication e.g. Kenya Yetu, though he speaks much of Swahili being made a national language.

It ought to be remembered that, some time last year President Idi Amin of Uganda announced his intention to make Swahili a national language. If Swahili is made a national language, East Africans will have one thing in common, Swahili which can facilitate federation.

Swahili should be promoted not only as a national language, but much so as a zonal language. Swahili has every quality to qualify to be a zonal language, and this must be given considerable thought by the East African authorities. Swahili could easily be the first African language to form "Phone" in Africa, since often Africa is divided in "Franco Phone and Africa-Arabic Phone".

I feel it is because of the English in East Africa which has made Swahili progressive to such an extent. I believe that Swahili still needs the help of its stronger brother — English to grow even more robustically.

I suggest a National Swahili Society be formed to facilitate the implementation of Swahili as a national language.

Abdul K. Bakar,
Nairobi.

A SW ARILI
By Abdul K. Bakar
Date published: 1974-05-28

Subject terms
- Swahili
- Languages
- English
- National history
- National languages

Source files

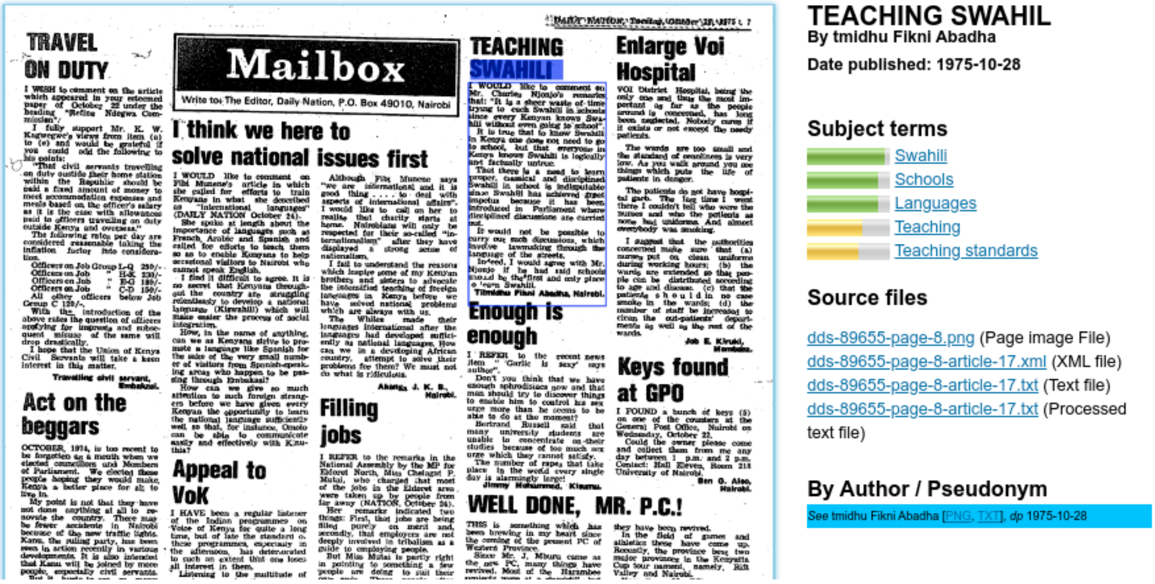
dds-89211-page-8.png (Page image File)
dds-89211-page-8-article-02.xml (XML file)
dds-89211-page-8-article-02.txt (Text file)
dds-89211-page-8-article-02.txt (Processed text file)

By Author / Pseudonym

See Abdul K. Bakar [PNG, TXT], dp 1974-05-28
See Abdul K. Bakar [PNG, TXT], dp 1974-08-14

Figure 5.1: Trimmed screenshot of a single page image result for assigned subject term "Swahili". The matched letter (*Daily NATION*, 28 May 1974) is highlighted in the image itself with an outline. Besides the image, the extracted title, subject terms associated with the topic, and links to the primary source derivatives as well as the author are also displayed alongside.

Beyond the utility of capturing the letter's aboutness at a glance, each term allows the user to access adjacent topics. For example, the subject term "Languages" extends the discussion into other languages in general as it happened within this column. In another result from the same query Figure 5.2, other high-scoring assigned terms "Schools", "Languages" and "Teaching" reveal other aspects of the discussion. Opinions and correspondences from contributors as secondary sources are made accessible through indexing. As the date references section will discuss, majority of the column letters also served to highlight primary sources that can be referenced outside the column.



TEACHING SWAHIL
By tmidhu Fikni Abadha
Date published: 1975-10-28

Subject terms

🟩 Swahili
🟩 Schools
🟩 Languages
🟨 Teaching
🟨 Teaching standards

Source files

dds-89655-page-8.png (Page image File)
dds-89655-page-8-article-17.xml (XML file)
dds-89655-page-8-article-17.txt (Text file)
dds-89655-page-8-article-17.txt (Processed text file)

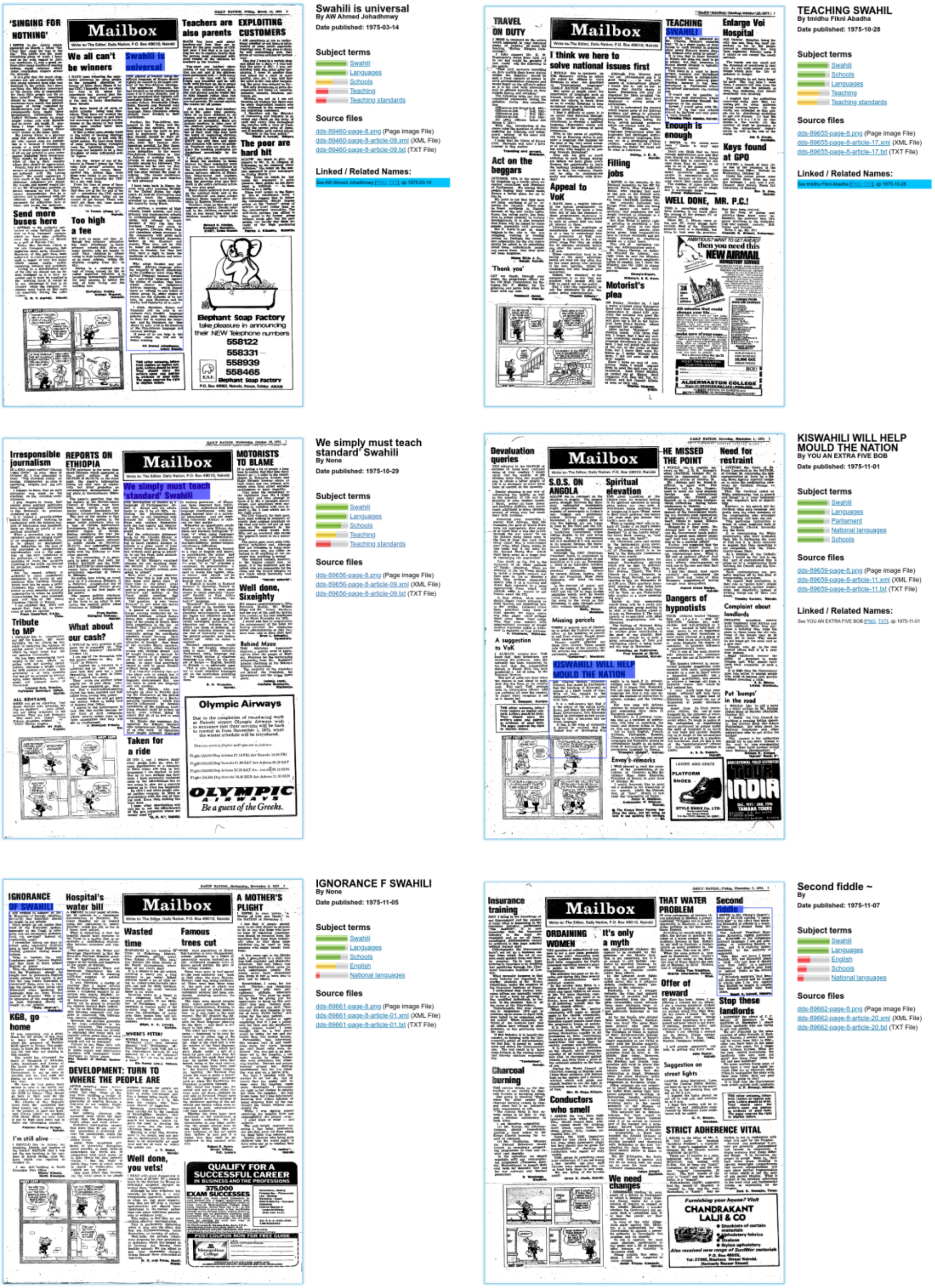
By Author / Pseudonym

See tmidhu Fikni Abadha [PNG, TXT], *dp* 1975-10-28

Figure 5.2: Trimmed screenshot of another single result of the subject term “Swahili” (*Daily NATION*, 28 October 1975).

A number of formats are generated and made accessible through the side panel. The particularly useful formats are plain text and XML files. The former allows a user to conduct textual analysis similar to that described in the methodology as well as regular full-text search. XML files are enriched further with subject terms to enable retrieval, hyperlinking, and sophisticated queries. Adding the relevance scores on each assigned term allows the user to filter out the generic level of results to retrieve. Having different file formats not only provides flexibility of handling located texts, they also allow human intervention, for example, as an indexing specialist. Intellectually, contents in the serialized XML file can be interpreted as suggestions to reject or accept as descriptors and potentially improve on term assignments. Automated indexing results in specific entries and therefore can also benefit from a more nuanced determination of descriptors. For example, “Swahili and Languages” would be an apposite coextensive entry than separate entries of “Swahili” and “Languages” as descriptors for opinions around the adoption of Swahili as a national language – a recurring subject within the column.

Having hyperlinks on the thesaurus subjects is intended to link the user to other letters in the collection that share the subject. The subjects are ordered and visualized to indicate their relevance to each assigned subject. Highly relevant subjects are listed first at the top, they are green and have a score of 0.75 or higher (based on a scale of 0.0 to 1.0 that is broken into four quarters of 0.25 each). Because the subjects and their scoring is stored in XML, it is possible to interact with the letters in a systematic way to obtain *Mailbox* letters that meet a user's threshold for relevance. Use of a thesaurus also allows linking users to the vocabulary especially since the hierarchical structure in the UNESCO Thesaurus is not part of the interface, or lost.



Swahili is universal
By AW Ahmed Johadhmwy
Date published: 1975-03-14

Subject terms
🟩 Swahili
🟩 Languages
🟨 Schools
🟥 Teaching
🟥 Teaching standards

Source files
dds-89460-page-8.png (Page image File)
dds-89460-page-8-article-09.xml (XML File)
dds-89460-page-8-article-09.txt (TXT File)

Linked / Related Names:
See AW Ahmed Johadhmwy [PNG, TXT], dp 1975-03-14

TEACHING SWAHIL
By tmidhu Fikni Abadha
Date published: 1975-10-28

Subject terms
🟩 Swahili
🟩 Schools
🟩 Languages
🟨 Teaching
🟨 Teaching standards

Source files
dds-89655-page-8.png (Page image File)
dds-89655-page-8-article-17.xml (XML File)
dds-89655-page-8-article-17.txt (TXT File)

Linked / Related Names:
See tmidhu Fikni Abadha [PNG, TXT], dp 1975-10-28

We simply must teach standard' Swahili
By None
Date published: 1975-10-29

Subject terms
🟩 Swahili
🟩 Languages
🟩 Schools
🟨 Teaching
🟥 Teaching standards

Source files
dds-89656-page-8.png (Page image File)
dds-89656-page-8-article-09.xml (XML File)
dds-89656-page-8-article-09.txt (TXT File)

KISWAHILI WILL HELP MOULD THE NATION
By YOU AN EXTRA FIVE BOB
Date published: 1975-11-01

Subject terms
🟩 Swahili
🟩 Languages
🟩 Parliament
🟩 National languages
🟩 Schools

Source files
dds-89659-page-8.png (Page image File)
dds-89659-page-8-article-11.xml (XML File)
dds-89659-page-8-article-11.txt (TXT File)

Linked / Related Names:
See YOU AN EXTRA FIVE BOB [PNG, TXT], dp 1975-11-01

IGNORANCE F SWAHILI
By None
Date published: 1975-11-05

Subject terms
🟩 Swahili
🟩 Languages
🟩 Schools
🟨 English
🟥 National languages

Source files
dds-89661-page-8.png (Page image File)
dds-89661-page-8-article-01.xml (XML File)
dds-89661-page-8-article-01.txt (TXT File)

Second fiddle ~
By
Date published: 1975-11-07

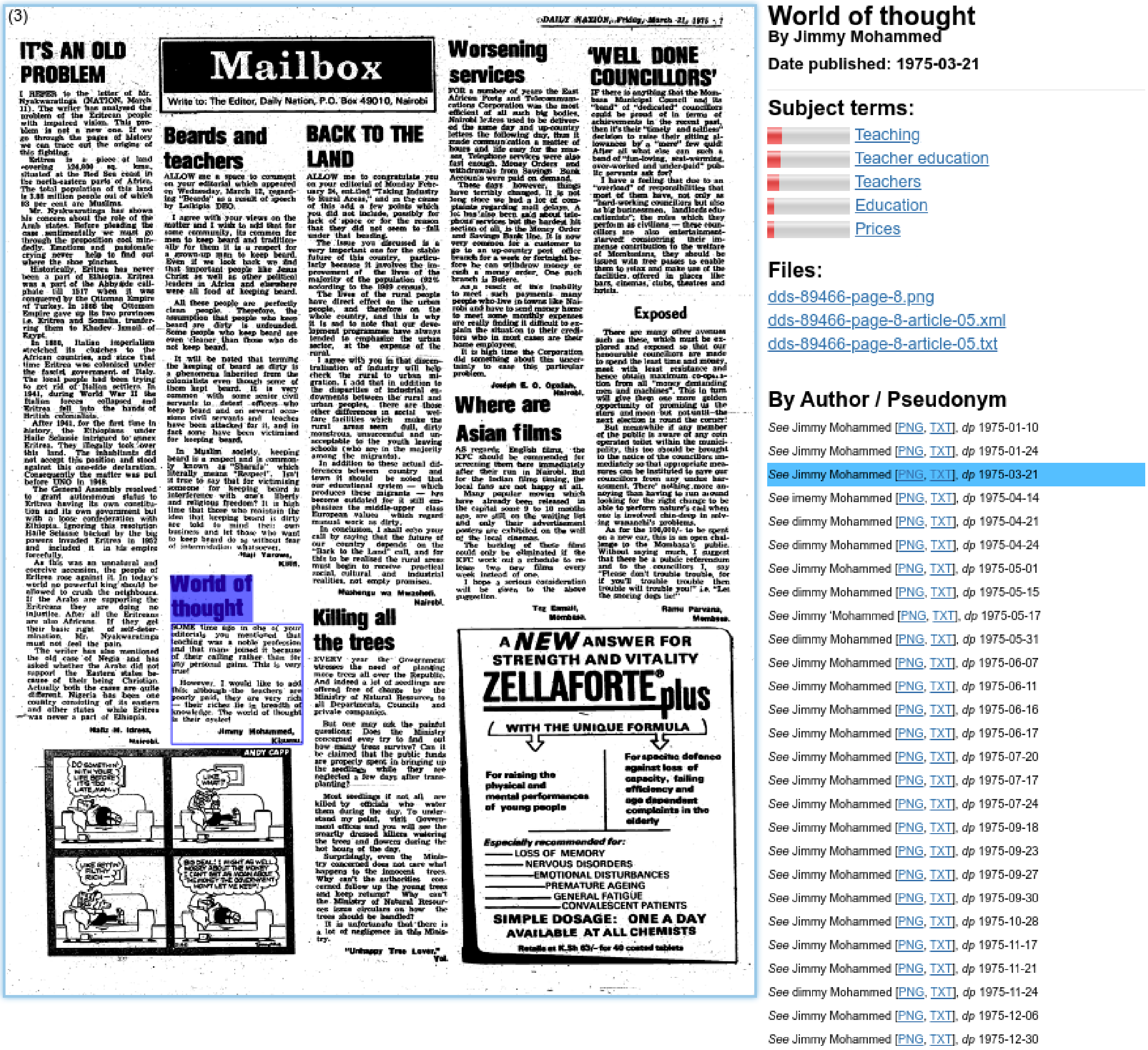
Subject terms
🟩 Swahili
🟩 Languages
🟥 English
🟥 Schools
🟥 National languages

Source files
dds-89662-page-8.png (Page image File)
dds-89662-page-8-article-20.xml (XML File)
dds-89662-page-8-article-20.txt (TXT File)

Figure 5.3: Screenshot showing a subset of six out of 33 results for the subject term "Swahili" found across the *Mailbox* column in the year 1975.

5.2.2 Contributor’s Corpus

The side panel on this interface has “By Author / Pseudonym” section which is a listing of all matched byline names. Shown in Figure 5.4, the current *Mailbox* letter is highlighted; date published informs the user of the letter’s situation and chronology of the author’s contributions. Being able to interconnect letters by an author or related identity is a powerful feature of this interface because it collects all contributions by the author and gives the user quick access to their corpus through hyperlinks. In this hyperlinked state, contributors and their audiences are more than a generalized collection of text, rather, their utterances continuously reveal the otherwise elusive context to the user. Though the majority of letters were published anonymously or using pseudonyms, this interface revealed several prolific contributors whose identities remained constant throughout. For example, in 1975 there were approximately 27 letters published on the column by one “Jimmy Mohammed” (see Figure 5.4). This author corpus can be generated reliably with a robust algorithm. To solely depend on this approach, however, has potential to mismatch in instances where pseudonyms are used. For example, anonymous tags commonly assigned by editors such as “Concerned parent”, “Concerned resident”, “Concerned Mother” or even “Concerned teacher” as Figure 5.5 shows, are contributions by the same pseudonym. However, it is apparent by their naming that these contributions address separate concerns just as they are likely written by unrelated identities. The complaints in this instance only share the lexical similarity but in this case are inaccurately organized as a corpus.



(3)

DAILY NATION, Friday, March 21, 1975

IT'S AN OLD PROBLEM

Mailbox

Write to: The Editor, Daily Nation, P.O. Box 49010, Nairobi

Beards and teachers

BACK TO THE LAND

Worsening services

'WELL DONE COUNCILLORS'

Exposed

Where are Asian films

World of thought

SOME time ago in one of your editorials you mentioned that teaching was a noble profession and that many joined it because of their calling rather than for any personal gains. This is very true!

However, I would like to add this: although the teachers are poorly paid, they are very rich — their riches lie in breadth of knowledge. The world of thought is their oyster!

Jimmy Mohammed,
Kisumu.

Killing all the trees

A NEW ANSWER FOR STRENGTH AND VITALITY ZELLAFORTE plus WITH THE UNIQUE FORMULA

World of thought

By Jimmy Mohammed

Date published: 1975-03-21

Subject terms:

Teaching
Teacher education
Teachers
Education
Prices

Files:

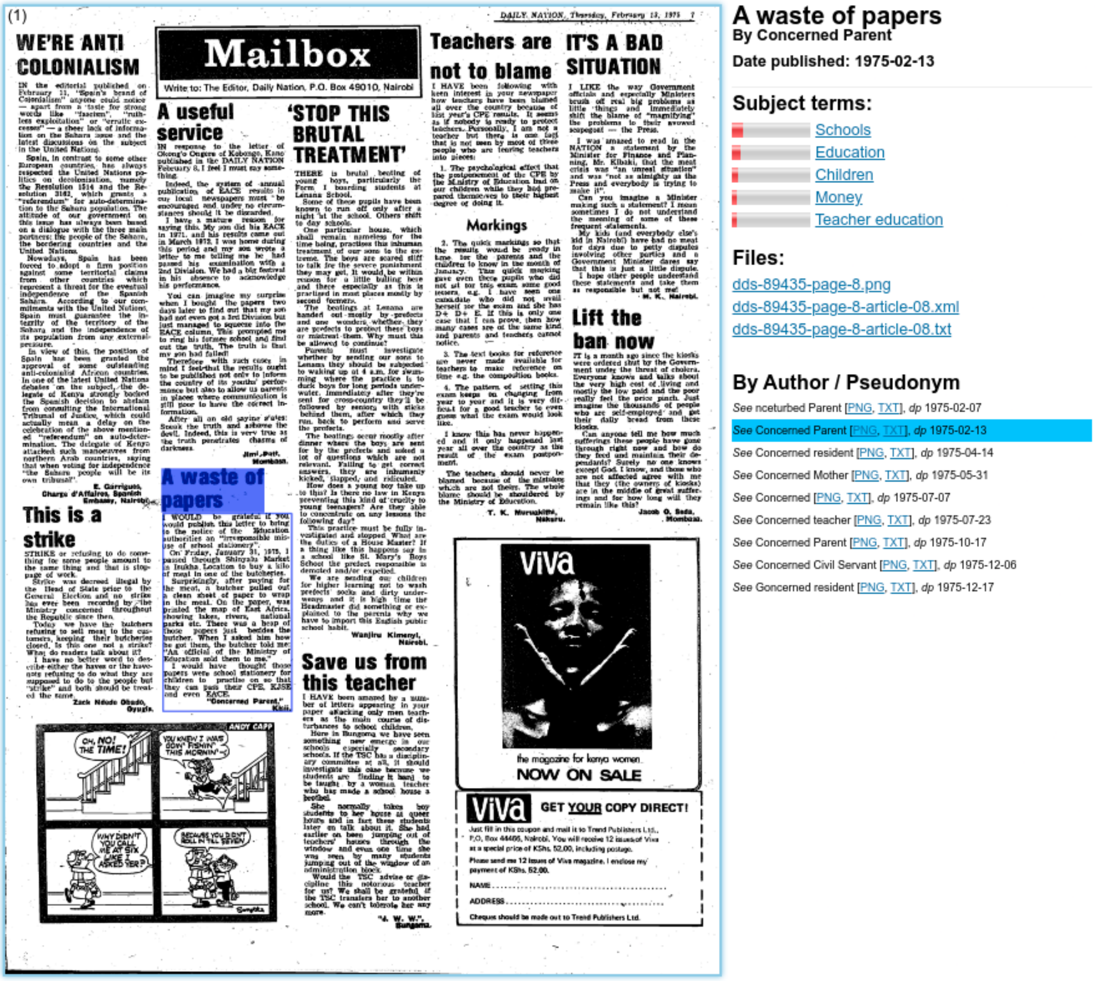
dds-89466-page-8.png
dds-89466-page-8-article-05.xml
dds-89466-page-8-article-05.txt

By Author / Pseudonym

See Jimmy Mohammed [PNG, TXT], *dp* 1975-01-10
See Jimmy Mohammed [PNG, TXT], *dp* 1975-01-24
See Jimmy Mohammed [PNG, TXT], *dp* 1975-03-21
See imemy Mohammed [PNG, TXT], *dp* 1975-04-14
See dimmy Mohammed [PNG, TXT], *dp* 1975-04-21
See dimmy Mohammed [PNG, TXT], *dp* 1975-04-24
See Jimmy Mohammed [PNG, TXT], *dp* 1975-05-01
See dimmy Mohammed [PNG, TXT], *dp* 1975-05-15
See Jimmy 'Mohammed [PNG, TXT], *dp* 1975-05-17
See dimmy Mohammed [PNG, TXT], *dp* 1975-05-31
See Jimmy Mohammed [PNG, TXT], *dp* 1975-06-07
See Jimmy Mohammed [PNG, TXT], *dp* 1975-06-11
See Jimmy Mohammed [PNG, TXT], *dp* 1975-06-16
See Jimmy Mohammed [PNG, TXT], *dp* 1975-06-17
See dimmy Mohammed [PNG, TXT], *dp* 1975-07-20
See Jimmy Mohammed [PNG, TXT], *dp* 1975-07-17
See Jimmy Mohammed [PNG, TXT], *dp* 1975-07-24
See Jimmy Mohammed [PNG, TXT], *dp* 1975-09-18
See Jimmy Mohammed [PNG, TXT], *dp* 1975-09-23
See Jimmy Mohammed [PNG, TXT], *dp* 1975-09-27
See Jimmy Mohammed [PNG, TXT], *dp* 1975-09-30
See Jimmy Mohammed [PNG, TXT], *dp* 1975-10-28
See Jimmy Mohammed [PNG, TXT], *dp* 1975-11-17
See Jimmy Mohammed [PNG, TXT], *dp* 1975-11-21
See dimmy Mohammed [PNG, TXT], *dp* 1975-11-24
See Jimmy Mohammed [PNG, TXT], *dp* 1975-12-06
See Jimmy Mohammed [PNG, TXT], *dp* 1975-12-30

Figure 5.4: Notebook screenshot showing result by authorship. The side panel highlights the current letter (*Daily NATION*, 21 March 1975) as well as previous and subsequent contributions by the author. In addition, hyperlinks to page image and plain text file for quick navigation.

To briefly analyze this from a historical perspective, such tags also highlight the gatekeeping role editors played by accepting and rejecting letters to generate and sustain commentary within their readerships. If anything, published letters reflected editorial subject position than they reflected the general sentiment of the society at that time. By marking the absence of authorial identity, the editor thus reified the necessity of attribution in the column. Therefore, examining this subset of documents within newspapers also intrudes on the editors' processes of publishing. It is the ability to organize texts along certain structures that allow us to collect evidence to support such arguments.



(1)

DAILY NATION, Thursday, February 13, 1975

WE'RE ANTI COLONIALISM

Mailbox

Write to: The Editor, Daily Nation, P.O. Box 49010, Nairobi

Teachers are not to blame

IT'S A BAD SITUATION

A useful service

'STOP THIS BRUTAL TREATMENT'

Markings

Lift the ban now

This is a strike

A waste of papers

I WOULD be grateful if you would publish this letter to bring to the notice of the Education authorities an "irresponsible misuse of school stationery".

On Friday, January 31, 1975, I passed through Shingalu Market in Isukha Location to buy a kilo of meat in one of the butcheries.

Surprisingly, after paying for the meat, a butcher pulled out a clean sheet of paper to wrap in the meal. On the paper, was printed the map of East Africa, showing lakes, rivers, national parks etc. There was a heap of those papers just besides the butcher. When I asked him how he got them, the butcher told me: "An official of the Ministry of Education sold them to me."

I would have thought those papers were school stationery for children to practise on so that they can pass their CPE, KJSE and even EACE.

"Concerned Parent,"
Kkii.

Save us from this teacher

ANDY CAPP
OH, NO! THE TIME!
YOU KNEW I WAS GOIN' FISHIN' THIS MORNIN'
WHY DIDN'T YOU CALL ME AT SIX LIKE I ASKED YER?
BECAUSE YOU DIDN'T ROLL IN TILL SEVEN!

A waste of papers
By Concerned Parent

Date published: 1975-02-13

Subject terms:

- Schools
- Education
- Children
- Money
- Teacher education

Files:

dds-89435-page-8.png
dds-89435-page-8-article-08.xml
dds-89435-page-8-article-08.txt

By Author / Pseudonym

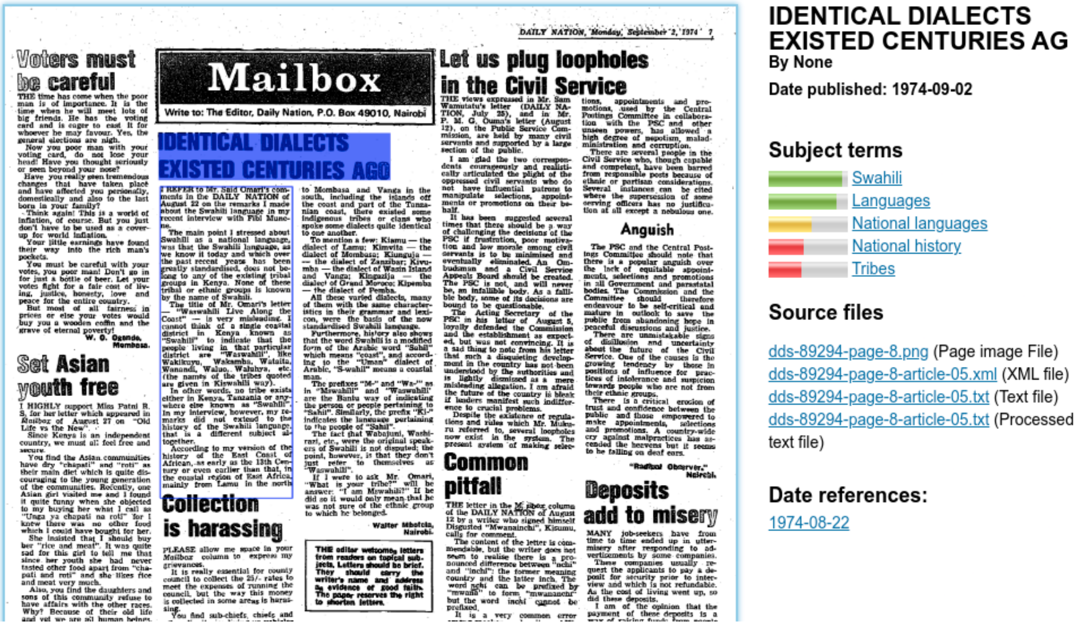
See nceturbed Parent [PNG, TXT], *dp* 1975-02-07
See Concerned Parent [PNG, TXT], *dp* 1975-02-13
See Concerned resident [PNG, TXT], *dp* 1975-04-14
See Concerned Mother [PNG, TXT], *dp* 1975-05-31
See Concerned [PNG, TXT], *dp* 1975-07-07
See Concerned teacher [PNG, TXT], *dp* 1975-07-23
See Concerned Parent [PNG, TXT], *dp* 1975-10-17
See Concerned Civil Servant [PNG, TXT], *dp* 1975-12-06
See Goncerned resident [PNG, TXT], *dp* 1975-12-17

Figure 5.5: Another Notebook screenshot showing authorship side panel with a list of pseudonyms related to the current letter (*Daily NATION*, 13 February 1975).

5.2.3 Date-references

It was common for correspondences and opinions to refer to primary sources published within the newspaper; be they editorials, advertisements, political and social news, stories, or even earlier *Mailbox* letters. Therefore, all occurrences of a date or date-like facet in the *Mailbox* column are a potential thread to an ongoing conversation especially when the date occurs in the past. This approach is helpful to glean from because of the way it creates connections to page images of the referenced date and gives users direct access (see Figure 5.6), with an emphasis only on correspondences addressing an ongoing debate in the column. More often than not, it provides context to the discussion or point to previous correspondences. Consider this short narrative obtained from a letter titled “Identical Dialects Existed Centuries Ago” (*Daily NATION*, 19 September 1974) by Walter Mbotela who invokes external sources in addition to a named and date-referenced *Mailbox* letter (see Figure 5.7a). As a hyperlink, it points the user one week earlier to a letter titled “WASWAHILI LIVE ALONG THE COAST, MR. MBOTELA” (*Daily NATION*, 22 August 1974) by Said Omar which incidentally refers to an earlier post by Walter Mbotela (see Figure 5.7b). Even though such threads in themselves are not indexes, to special formats such as newspaper correspondences it enables access to what is pertinent to the user.

Trails of discussion such as the one above primarily rely on extracted dates and therefore limited where none is provided. And particularly evident in this example where Said Omar while replying to the earlier post by Walter Mbotela is not referenced directly in any way. Because the date reference narrows down the scope of where and when to locate the source, lack of date, however, prevents drawing subsequent linkages. Another potential limitation is the lack of direct connections to the particular letter but only to the page image in general. This is due to the fact that date references alone cannot reliably point to the subject letter, references must be paired with additional metadata such as cited name, quotation or even geometric properties obtained during segmentation.



DAILY NATION, Monday, September 2, 1974 7

Voters must be careful

Mailbox

Write to: The Editor, Daily Nation, P.O. Box 49010, Nairobi

IDENTICAL DIALECTS EXISTED CENTURIES AGO

Let us plug loopholes in the Civil Service

Anguish

Set Asian youth free

Collection is harassing

Common pitfall

Deposits add to misery

IDENTICAL DIALECTS EXISTED CENTURIES AG

By None

Date published: 1974-09-02

Subject terms

- Swahili
- Languages
- National languages
- National history
- Tribes

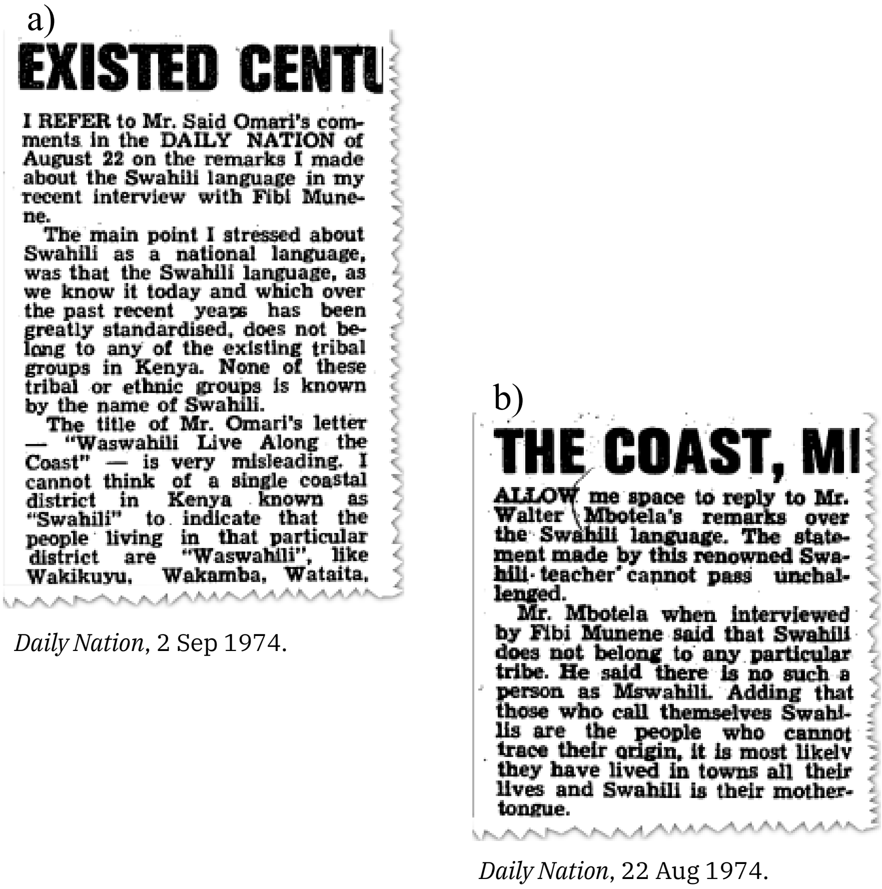
Source files

dds-89294-page-8.png (Page image File)
dds-89294-page-8-article-05.xml (XML file)
dds-89294-page-8-article-05.txt (Text file)
dds-89294-page-8-article-05.txt (Processed text file)

Date references:

1974-08-22

Figure 5.6: Trimmed screenshot of a letter correspondence with a single date reference.



a)

EXISTED CENTU

I REFER to Mr. Said Omari's comments in the DAILY NATION of August 22 on the remarks I made about the Swahili language in my recent interview with Fibi Munene.

The main point I stressed about Swahili as a national language, was that the Swahili language, as we know it today and which over the past recent years has been greatly standardised, does not belong to any of the existing tribal groups in Kenya. None of these tribal or ethnic groups is known by the name of Swahili.

The title of Mr. Omari's letter — "Waswahili Live Along the Coast" — is very misleading. I cannot think of a single coastal district in Kenya known as "Swahili" to indicate that the people living in that particular district are "Waswahili", like Wakikuyu, Wakamba, Wataita,

Daily Nation, 2 Sep 1974.

b)

THE COAST, MI

ALLOW me space to reply to Mr. Walter Mbotela's remarks over the Swahili language. The statement made by this renowned Swahili teacher cannot pass unchallenged.

Mr. Mbotela when interviewed by Fibi Munene said that Swahili does not belong to any particular tribe. He said there is no such a person as Mswahili. Adding that those who call themselves Swahilis are the people who cannot trace their origin, it is most likelv they have lived in towns all their lives and Swahili is their mother-tongue.

Daily Nation, 22 Aug 1974.

Figure 5.7: Clippings of corresponding *Mailbox* letters.

Nonetheless, gleaning into these narratives reveals a topic trail that in turn uncovers information on the nature of discourse, articulation, and the class of readers that contributed to this column. On the surface, these interface applications place the user within access of the image and other metadata as multidimensional texts, yet fundamentally they attempt to address the need for organization, indexing of what one would otherwise consider unclassifiable texts. Navigating the collection thus is not merely via full-text search but mediated through named entities and characteristics extracted from individual letters and connected together.

5.3 Conclusion

Though the focus of this chapter was on the application of results of the preceding chapter, presenting results required that we also acknowledge the state of results. Following a brief summary of quantitative or empirical results to that end, this chapter has presented a number of interfaces that seek to mobilize the *Mailbox* letters along bibliographic forms. In particular the use of subject vocabulary and byline to show a refigured page image interface that is accessible and also reveals related topics, subjects and other components such as date references. As exemplars, the application interfaces aimed to correspond further with the last question of this thesis research, which is concerned with how to organize access to the *Mailbox* column of the newspaper through the multiple dimensions of bibliographic access.

Chapter 6

6 Conclusion

6.1 Summary

This final chapter reflects on the work as well as contributions this thesis makes, before concluding with areas of potential future research. In several ways, it brings an array of interdisciplinary elements with the goal of extending access to the vast archives of digitized newspapers. In its theorizing, the thesis has argued that the challenges of locating content of particular interest within digitized newspapers are not just impeded by the vastness of its text, but also entangled in the processes it seeks to adopt. The goal was to examine the principles which guide the practices of identifying, describing texts with the aim of facilitating retrieval. This was done by focusing on the characteristics that favor the text at hand; that is, letters to the editor and the method of identifying and describing useful texts. Based on the existing research, the task of constructing automated indexes is both complicated and constantly evolving with technology and human techniques. Digital scanning of historical newspapers serve the first step. However, it only creates an underutilized archive because the apparatuses which mediate access to research texts remain passive on presentation and usability aspects. In this case, I have paid more attention to letters to the editor column in newspapers because they represent an inherently topical, and hold historical merit.

Situating this argument outside the traditional field of library and information sciences was an attempt at ameliorating this form of inaccessibility inside an interdisciplinary framework. The foremost disciplines being media studies and the self-named practices of the digital humanities (Gold & Klein, 2016, p. xiv). Media shifting from one form of media to another has to be refigured to imitate characteristics of its new format and environment, while also trying to foreground what is embedded in the text. While seemingly a metaphor, it seeks to realize the notional immediacy of research where subjects or topics of interest found in text are no longer disconnected, but as text, they are reenacted as narratives (correspondence, debates, opinions) that users encounter navigating the archive. Despite that, to attenuate the distance between archive and the user in this framing, there is an interplay between formalized access strategies, state of the art and techniques.

On this foundation, I have explored how this was realized using a sizable collection of the *Daily NATION* newspapers covering letters sent to the editor between 1974 and 1978. The focus of the methodology, beyond critique and examination of the techniques, was on the challenges involved in this exploratory process. First, it is an interactive process with limited scope for nuance, particularly when dealing with physical and logical image segmentation. Secondly, cleaning text extracted from OCR foreclosed the benefit of working with localized names and locales as entities. Because in order to make *Mailbox* letters legible for any effective indexing, using a spell checker meant that these entities would be incorrectly considered misspellings and lost in the process. Finally, the heuristic approach on which this research is designed does not offer complete solution; rather it presents the state of the art and bricolage of tools as an effective framework. Tools useful to tackle challenges of exploring knowledge resources in the growing digitized archive. In the end, the outcome represents another reference point onto which print materials were included into the fold of knowledge systems. These explorations only attempt to realize what digitization that attends to a special form of document entails.

Although the illustrated application interfaces were prototypical, they (re)construct the contemporaneous relationships, identities and subjects as the ideal way of navigating the archive. In the process, this thesis sheds new light on the challenges other researchers and practitioners have encountered, and brings to fore the daunting task ahead. Furthermore, it also presents a compelling instance of an oft-obscured text with new potential ways of imagining how users will better locate and access historical texts with the digital environment.

6.2 Contributions

At the onset of this research, I was motivated by the idea that navigating digitized newspaper collections should not be laborious. But the collections lacked affordances to trace subjects or mobilize a corpus by particular subjects. On the other hand, the rich experience elsewhere on born-digital collections particularly in academic journal archives and online news articles connected me to "related items", works by the same author, and even recommended similar items of interest. While both digitized newspapers and born-digital counterparts now exist in the digital plane, my interaction with the former which only embodies an image form brought me to a conclusion that beside the digitized embodiment, there exists a number of theoretical, disciplinary and methodological reasons which limit the reach of the *digital* experience. Sure, the presuppositions of engaging with a page image often means that its contents are inherently not hyperlinked for interactions; but still, the embedded texts exist digitally to benefit research, particularly for historians, cultural researchers and humanities scholars.

Toward this goal, and through the thesis I have promoted the narrative that embodiment of a text in itself does not limit the very text but actually points to the limits of knowledge systems. First, I have presented the methodological approaches that center obscured texts as basic units for textual and bibliographical analyses. Secondly, and related to the first, the resulting datasets contribute to the expansion of subjects of historical research, be they personalities or topics that emerge from such approaches and the attempt to achieve bibliographic control within these collections. Finally, this thesis has also illuminated on interdisciplinarity as another means by which we can permeate disciplinary limits. Through the theorization and techniques of adjacent fields, I have demonstrated possibilities of giving full treatment to the vast collection that is digitized newspapers. However, as the results have shown, they also highlight epistemological and technical challenges. The framework used and to a greater extent the computer tools methodologies espouse a discreteness and binarism that pervades interaction with computational tools; which while inadequate to deal with ambiguity found in handling data was only sufficient for this research. Perhaps it would be refreshing for future research to explore fuzzy epistemologies and logics as alternatives to the twin logic of remediation and binary logic of computations.[35] At any rate, there is room to improve on the current methodology.

This thesis has also laid the groundwork for broader empirical research that attends to specificity of subjects and topics. In the conclusion to *The power to name*, Newell (2013d) suggests among other things, the need for a broad framework for comparative studies between historical readerships across Africa and the world (pp. 180 - 181). By building interfaces which make intertextuality apparent and allow users to interact with embedded print relationships, I believe, illustrates the practical implications toward this end.

6.3 Limitation and Future Research

Regarding tools and methods, further research needs to be done to improve working with localized entities. For instance, a significant number of local terminologies and named contributors were miscorrected by automated spell checker. While such tools are needed to improve digitization and generally perform well to correct OCR output, additional work is often required to improve the embedded dictionaries. This requires some form of training data, and newspaper columns proffer a robust resource to build a meaningful collection of people, events, organizations, locations and other entities not readily found in spell checkers when indexing for localized needs.

Although the scope to refigure the entire newspaper run was inconceivable for this thesis, future research could see newspaper archives and holdings refigured as rich collections organized by specific genre, thesauri vocabulary, among many other dimensions of access. These can provide new opportunities for cross-reference and additional headings, techniques can be helpful to collection users to retrieve within and across disparate columns. On the methodological side, there is potential to expand the heuristic approaches to benefit other columns found in the archive. These include poems, local events, notices, political reportage and other columns that are determined useful for users.

[35] An exploration of non-Western epistemologies and their implication in humanistic and social science disciplines. It notes how binary logic of computing, on the other hand, is inadequate to handle fuzzy or complex realities disciplines engage with. Reiter, B. (2020). Fuzzy epistemology: Decolonizing the social sciences. Journal for the Theory of Social Behaviour, 50(1), 103–118. https://doi.org/10.1111/jtsb.12229

Future research would cover multiple newspapers and an extended timeline to include the earliest days of the newsprint. A considerable challenge no doubt but hopefully it will advance knowledge and practice on how to compile material from different locales, find utility to users, and make these automated methods more effective and accessible to library and information workers.

REFERENCES

Andersen, J. (2008). Information Criticism: Where is it? In A. M. Lewis (Ed.), *Questioning library neutrality: Essays from Progressive librarian* (pp. 97–108).

Andersen, P. B. (2003). Acting Machines. In G. Liestøl, A. Morrison, & T. Rasmussen (Eds.), *Digital Media Revisited* (pp. 183–213). MIT Press. http://dl.acm.org/citation.cfm?id=778062.778070

Anderson, J. D. (1985). Indexing systems: Extensions of the mind's organizing power. *Information and Behavior*, *1*, 287–323.

Anderson, J. D., & Pérez-Carballo, J. (2001). The nature of indexing: How humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing. *Information Processing & Management*, *37*(2), 231–254. https://doi.org/10.1016/S0306-4573(00)00026-1

Atkinson, S. D., & Hudson, J. (1990). *Women Online: Research in Women's Studies Using Online Databases*. Psychology Press.

Baird, H. S. (1992). Background Structure In Document Images. *In Advances in Structural and Syntactic Pattern Recognition*, 17–34.

Balahur, A., & Steinberger, R. (n.d.). *Rethinking Sentiment Analysis in the News: From Theory to Practice and back*. 12.

Barber, K. (2007, December). *The Anthropology of Texts, Persons and Publics by Karin Barber*. Cambridge Core. https://doi.org/10.1017/CBO9780511619656

Bingham, A. (2010). 'The Digitization of Newspaper Archives: Opportunities and Challenges for Historians.' *Twentieth Century British History*, *21*(2), 225–231. https://doi.org/10.1093/tcbh/hwq007

Blei, D. M. (2012). Topic Modeling and Digital Humanities Journal of Digital Humanities. *Journal of Digital Humanities*, *2*(1). http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/

Bolter, J. D. (2001). *Writing space: Computers, hypertext, and the remediation of print /* (2nd ed.). Lawrence Erlbaum Associates,.

Bolter, J. D., & Grusin, R. (1999). *Remediation: Understanding New Media*.

Bond, R. P. (1969). *Growth & change in the early English press*. University of Kansas Libraries,.

Bradski, G. (2000). The OpenCV library. *Dr. Dobb's Journal*, *25*(11), 120–125. Business Premium Collection.

Campbell, G. (2013). Aboutness and Meaning: How a Paradigm of Subject Analysis Can Illuminate Queer Theory in Literary Studies. *Proceedings of the Annual Conference of CAIS / Actes Du Congrès Annuel de l'ACSI*. https://doi.org/10.29173/cais9

Canny, J. (1986). A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-8*(6), 679–698. https://doi.org/10.1109/TPAMI.1986.4767851

Cao, H., Prasad, R., Natarajan, P., & MacRostie, E. (2007). Robust Page Segmentation Based on Smearing and Error Correction Unifying Top-down and Bottom-up Approaches. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, *1*, 392–396. https://doi.org/10.1109/ICDAR.2007.4378738

Chen, K., Yin, F., & Liu, C. (2013). Hybrid Page Segmentation with Efficient Whitespace Rectangles Extraction and Grouping. *2013 12th International Conference on Document Analysis and Recognition*, 958–962. https://doi.org/10.1109/ICDAR.2013.194

Chen, Y., Zhang, H., Liu, R., Ye, Z., & Lin, J. (2019). Experimental explorations on short text topic mining between LDA and NMF based Schemes. *Knowledge-Based Systems*, *163*, 1–13. https://doi.org/10.1016/j.knosys.2018.08.011

Cleveland, D. B., & Cleveland, A. D. (2001). *Introduction to indexing and abstracting*. Libraries Unlimited.

Cohen, M. (2009). Narratology in the Archive of Literature. *Representations*, *108*(1), 51–75. https://doi.org/10.1525/rep.2009.108.1.51

Cordell, R. (2017). "Q i-jtb the Raven": Taking Dirty OCR Seriously. *Book History*, *20*(1), 188–225. https://doi.org/10.1353/bh.2017.0006

Day, R. E., Buckland, M., Furner, J., & Krajewski, M. (2014). *Indexing It All: The Subject in the Age of Documentation, Information, and Data*. MIT Press.

Dixon, D. (2012). Analysis Tool or Research Methodology: Is There an Epistemology for Patterns? In D. M. Berry (Ed.), *Understanding Digital Humanities* (pp. 191–209). Palgrave Macmillan UK. https://doi.org/10.1057/9780230371934_11

Esposito, F., Ferilli, S., Basile, T. M. A., & Di Mauro, N. (2008). Machine Learning for Digital Document Processing: From Layout Analysis to Metadata Extraction. In S. Marinai & H. Fujisawa (Eds.), *Machine Learning in Document Analysis and Recognition* (pp. 105–138). Springer. https://doi.org/10.1007/978-3-540-76280-5_5

Fagerjord, A. (2003). Rhetorical Convergence. In T. Rasmussen, A. Morrison, & G. Liestøl, *Digital Media Revisited: Theoretical and Conceptual Innovations in Digital Domains* (pp. 293–325). The MIT Press; e000xna.

Ferilli, S. (2011). *Automatic digital document processing and management: Problems, algorithms and techniques*. Springer.

Ferilli, S., Esposito, F., & Redavid, D. (2017). A Study on the Classification of Layout Components for Newspapers. In M. Agosti, M. Bertini, S. Ferilli, S. Marinai, & N. Orio (Eds.), *Digital Libraries and Multimedia Archives* (Vol. 701, pp. 166–178). Springer International Publishing. https://doi.org/10.1007/978-3-319-56300-8_15

Fidel, R. (1994). User-centered indexing. *Journal of the American Society for Information Science*, *45*(8), 572–576. https://doi.org/10.1002/(SICI)1097-4571(199409)45:8<572::AID-ASI11>3.0.CO;2-X

Gadsden, F. (1980). The African Press in Kenya, 1945–1952. *The Journal of African History*, *21*(04), 515. https://doi.org/10.1017/S0021853700018727

Gardiner, E., & Musto, R. G. (Eds.). (2015). The Elements of Digital Humanities: Text and Document. In *The Digital Humanities: A Primer for Students and Scholars* (pp. 31–42). Cambridge University Press. https://doi.org/10.1017/CBO9781139003865.004

Garrod, P. (2000). Use of the UNESCO Thesaurus for Archival Subject Indexing at UK NDAD. *Journal of the Society of Archivists*, *21*(1), 37–54. https://doi.org/10.1080/00379810050006902

Gold, M. K., & Klein, L. F. (Eds.). (2016). *Debates in the digital humanities: 2016*. University of Minnesota Press.

Hayles, N. K. (2004). *Print Is Flat, Code Is Deep: The Importance of Media-Specific Analysis*. 14.

Hebert, D., Palfray, T., Nicolas, S., Tranouez, P., & Paquet, T. (2014). Automatic Article Extraction in Old Newspapers Digitized Collections. *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 3–8. https://doi.org/10.1145/2595188.2595195

Hiley, N. (2006). Indexing cartoons. *The Indexer: The International Journal of Indexing*, *25*(2), 100–104. https://doi.org/10.3828/indexer.2006.27

Hjørland, B., birger. hjorland@hum. ku. dk. (2017). Subject (of Documents). *Knowledge Organization*, *44*(1), 55–64. lls.

Howard-Reguindin, P. (2008). Initiatives in Kenya for Digitizing, Indexing and Preserving Newspapers. In H. Walravens (Ed.), *IFLA Publications*. Walter de Gruyter – K. G. Saur. https://doi.org/10.1515/9783598441011.357

Hung-Ming Sun. (2005). Page segmentation for Manhattan and non-Manhattan layout documents via selective CRLA. *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, 116-120 Vol. 1. https://doi.org/10.1109/ICDAR.2005.185

Hutchins, W. J. (1977). *The concept of "aboutness" in subject indexing*.

Ignatow, G., & Mihalcea, R. (2017). *Text Mining: A Guidebook for the Social Sciences*. SAGE Publications, Inc. https://doi.org/10.4135/9781483399782

Jackson, H. (2007). *Key terms in linguistics*. Continuum,.

James, L. (2016). Transatlantic Passages: Black Identity Construction in West African and West Indian Newspapers, 1935–1950. In D. R. Peterson, E. Hunter, & S. Newell (Eds.), *African Print Cultures* (pp. 49–74). University of Michigan Press; JSTOR.

James-Gilboe, L. (2005). The Challenge of Digitization. *The Serials Librarian*, *49*(1–2), 155–163. https://doi.org/10.1300/J123v49n01_06

Kapsalis, E. (2019). Wikidata: Recruiting the Crowd to Power Access to Digital Archives. *Journal of Radio & Audio Media*, *26*(1), 134–142. https://doi.org/10.1080/19376529.2019.1559520

Klijn, E. (2008). The Current State-of-art in Newspaper Digitization: A Market Perspective. *D-Lib Magazine*, *14*(1/2). https://doi.org/10.1045/january2008-klijn

Koistinen, M., Kettunen, K., & Kervinen, J. (2017). *How to Improve Optical Character Recognition of Historical Finnish Newspapers Using Open Source Tesseract OCR Engine*. 6.

Konya, I., & Eickeler, S. (2014). Logical Structure Recognition for Heterogeneous Periodical Collections. *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 185–192. https://doi.org/10.1145/2595188.2595211

Krtalic, M., & Hasenay, D. (2012, May 8). Newspapers as a source of scientific information in social sciences and humanities: A case study of Faculty of Philosophy, University of Osijek, Croatia. *Session 119 — Users and Portals: Digital Newspapers, Usability, and Genealogy — Newspapers with Genealogy and Local History*. IFLA World Library and Information Congress, Helsinki, Finland. https://www.ifla.org/past-wlic/2012/119-krtalic-en.pdf

Lancaster, F. W. (2003). *Indexing & Abstracting in Theory & Practice* (3 edition). Univ of Illinois Graduate School of.

Landau, R. N., & Wanger, J. (1980). Nonbibliographic on-line data base services. *Journal of the American Society for Information Science*, *31*, 171–180.

Limb, P. (2005). The Digitization of Africa. *Africa Today*, *52*(2), 3–19. JSTOR.

López-Calvo, I., & Mignolo, W. (2016). *Coloniality is not over, it is all over*. 11.

Lorang, E., Soh, L.-K., Datla, M., & Kulwicki, S. (2015). Developing an Image-Based Classifier for Detecting Poetic Content in Historic Newspaper Collections. *Faculty Publications, UNL Libraries*. https://digitalcommons.unl.edu/libraryscience/340

Manovich, L. (2001). *The language of new media*. MIT Press.

Martin, S. E., & Hansen, K. A. (1998). Newspapers of record in a digital age: From hot type to hot link /. Praeger,.

Martínez-González, M. M., & Alvite-Díez, M.-L. (2019). Thesauri and Semantic Web: Discussion of the Evolution of Thesauri Toward Their Integration With the Semantic Web. *IEEE Access*, *7*, 153151–153170. https://doi.org/10.1109/ACCESS.2019.2948028

McCarty, W. (2004). Modeling: A Study in Words and Meanings. In S. Schreibman, R. Siemens, & J. Unsworth, *A Companion to Digital Humanities*. John Wiley & Sons, Incorporated.

McGann, J. J. (2014). *A new republic of letters: Memory and scholarship in the age of digital reproduction /*.

Moens, Marie-Francine. (2000). *Automatic Indexing and Abstracting of Document Texts*. Kluwer Academic Publishers.

Moens, MARIE-FRANCINE, Uyttendaele, C., & Dumortier, J. (1999). Information extraction from legal texts: The potential of discourse analysis. *International Journal of Human-Computer Studies*, *51*(6), 1155–1171. https://doi.org/10.1006/ijhc.1999.0296

Mullan, J. (2007). *Anonymity: A secret history of English literature /*. Princeton University Press,.

Mulvany, N. C. (1994). *Indexing books*. University of Chicago Press,.

Newell, S. (2011, September 22). *Articulating empire: Newspaper readerships in colonial West Africa*. New Formations. https://link.galegroup.com/apps/doc/A291701759/LitRC?sid=lms

Newell, S. (2013a). Anonymity, Pseudonymity, and the Question of Agency in Colonial West African Newspapers. In *The Power to Name* (pp. 1–26). Ohio University Press; JSTOR.

Newell, S. (2013b). Articulating Empire: Newspaper Networks in Colonial West Africa. In *The Power to Name* (pp. 44–62). Ohio University Press; JSTOR.

Newell, S. (2013c). CONCLUSION: "New Visibilities" African Print Subjects and the Birth of the (Postcolonial) Author. In *The Power to Name* (pp. 170–182). Ohio University Press; JSTOR.

Newell, S. (2013d). *The power to name: A history of anonymity in colonial West Africa /*. Ohio University Press.

Nichols, S. (2007). Digital Scholarship: What's All the Fuss? *CLIR*, *58 (July/August)*. https://www.clir.org/2007/07/clir-issues-number-58/

Nuckolls, K. A. (2015). *LC Subject Headings, FAST Headings, and Apps: Diversity Can Be Problematic In the 21st Century*. 9.

O'Callaghan, D., Greene, D., Carthy, J., & Cunningham, P. (2015). An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, *42*(13), 5645–5657. https://doi.org/10.1016/j.eswa.2015.02.055

Owens, T. (2013, June 3). *Freeing Images from Inside Digitized Books and Newspapers*. THATCamp CHNM 2013. http://chnm2013.thatcamp.org/06/03/freeing-images-from-inside-digitized-books-and-newspapers/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Peterson, D. R., Hunter, E., & Newell, S. (Eds.). (2016). *African Print Cultures: Newspapers and Their Publics in the Twentieth Century*. University of Michigan Press; JSTOR. https://doi.org/10.3998/mpub.8833121

Powell, M. N. (2012). *Performing authorship in eighteenth-century English periodicals /*. Bucknell University Press ;

Reddy, V. (2019). *Vasistareddy/python-rlsa* [Python]. https://github.com/Vasistareddy/python-rlsa (Original work published 2018)

Reiter, B. (2020). Fuzzy epistemology: Decolonizing the social sciences. *Journal for the Theory of Social Behaviour*, *50*(1), 103–118. https://doi.org/10.1111/jtsb.12229

Reul, C., Springmann, U., Wick, C., & Puppe, F. (n.d.). *State of the Art Optical Character Recognition of 19th Century Fraktur Scripts using Open Source Engines*. 6.

Reul, C., Springmann, U., Wick, C., & Puppe, F. (2018). State of the Art Optical Character Recognition of 19th Century Fraktur Scripts using Open Source Engines. *ArXiv:1810.03436 [Cs]*. http://arxiv.org/abs/1810.03436

Ripplinger, B. (2001). *Automatic Multilingual Indexing and Natural Language Processing*.

Röhle, B. R. T. (2012). Digital Methods: Five Challenges. In D. M. Berry (Ed.), *Understanding Digital Humanities* (pp. 67–84). Palgrave Macmillan UK. https://doi.org/10.1057/9780230371934_4

Rutenbeck, J. B. (2009). The New Literacy Challenge: Coming to Terms with the Processes and Practices of Digitization. In J. E. Ruggill, K. S. McAllister, & J. R. Chaney (Eds.), *The computer culture reader* (pp. 90–107). Cambridge Scholars.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620. https://doi.org/10.1145/361219.361220

Salton, Gerard. (1991). Developments in Automatic Text Retrieval. *Science*, *253*(5023), 974–980. https://doi.org/10.1126/science.253.5023.974

Salton, Gerard, Allan, J., Buckley, C., & Singhal, A. (1994). Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts. *Science*, *264*(5164), 1421–1426. https://doi.org/10.1126/science.264.5164.1421

Samuels, R. (2006). *Integrating hypertextual subjects: Computers, composition, and academic labor /*. Hampton Press,.

Schirmer, J. (2009). The Personal as Public: Identity Construction/Fragmentation Online. In J. E. Ruggill, K. S. McAllister, & J. R. Chaney (Eds.), *The computer culture reader* (pp. 61–72). Cambridge Scholars.

Shafait, F., & Smith, R. (2010). Table Detection in Heterogeneous Documents. *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, 65–72. https://doi.org/10.1145/1815330.1815339

Shiri, A., Ruecker, S., Fiorentino, C., Stafford, A., Bouchard, M., & Bieber, M. (2010). *Designing a Semantically Rich Visual Interface for Cultural Digital libraries Using the UNESCO Multilingual Thesaurus*. https://experts.illinois.edu/en/publications/designing-a-semantically-rich-visual-interface-for-cultural-digit

Simon, J. (2015). *Access to Print, Digitized, and Born-Digital Newspapers from Africa: The North American / Global Conundrum*. IFLA WLIC 2015 - Cape Town, South Africa. http://library.ifla.org/1269/

Small, H. G. (1978). Cited Documents as Concept Symbols. *Social Studies of Science*, *8*(3), 327–340. https://doi.org/10.1177/030631277800800305

Smith, R. (2007). An Overview of the Tesseract OCR Engine. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2*, 629–633. https://doi.org/10.1109/ICDAR.2007.4376991

Smith, R. W. (2009). Hybrid Page Layout Analysis via Tab-Stop Detection. *2009 10th International Conference on Document Analysis and Recognition*, 241–245. https://doi.org/10.1109/ICDAR.2009.257

Suominen, O. (2019). Annif: DIY automated subject indexing using multiple algorithms. *LIBER Quarterly*, *29*(1), 1–25. https://doi.org/10.18352/lq.10285

Toepfer, M., & Seifert, C. (2018). Fusion architectures for automatic subject indexing under concept drift. *International Journal on Digital Libraries*. https://doi.org/10.1007/s00799-018-0240-3

Underwood, T. (2014). *Understanding Genre in a Collection of a Million Volumes, Interim Report*. https://doi.org/10.6084/m9.figshare.1281251

van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., & Yu, T. (2014). scikit-image: Image processing in Python. *PeerJ*, *2*. https://doi.org/10.7717/peerj.453

Walsh, J. A. (2012). Comic Book Markup Language: An Introduction and Rationale. *Digital Humanities Quarterly*, *006*(1). http://www.digitalhumanities.org/dhq/vol/6/1/000117/000117.html

Wang, D., & Srihari, S. N. (1989). Classification of newspaper image blocks using texture analysis. *Computer Vision, Graphics, and Image Processing*, *47*(3), 327–352. https://doi.org/10.1016/0734-189X(89)90116-3

Warner, M. (2005). *Publics and counterpublics* (1st pbk ed.). Zone Books,.

Wong, K. Y., Casey, R. G., & Wahl, F. M. (1982). Document Analysis System. *IBM Journal of Research and Development*, *26*, 647–656. https://doi.org/10.1147/rd.266.0647

Woodruff, A. G., & Plaunt, C. (1994). GIPSY: Automated geographic indexing of text documents. *Journal of the American Society for Information Science*, *45*(9), 645–655. https://doi.org/10.1002/(SICI)1097-4571(199410)45:9<645::AID-ASI2>3.0.CO;2-8

Zeni, M., & Weldemariam, K. (2017). Extracting information from newspaper archives in Africa. *IBM Journal of Research and Development*, *61*(6), 12:1-12:12. https://doi.org/10.1147/JRD.2017.2742706

APPENDICES

APPENDIX A: Splitting the Newspaper and extracting the *Mailbox* Column

Command line tools were used to split the PDF document format daily issues into separate pages. Thereafter, the pages were converted into PNG image format.

Steps to split the issues:

1. Split each issue into individual pages. To do this on all the files I used a script which accepted a directory as an input. Directory name was the year the files were collected under. For instance, 1975 contained all the issues for the year 1975. The output comprised individual pages identified by the original filename with the page number appended at the end.

```bash
#!/bin/bash

# Shell script to do pre processing
# Usage - ./pre_process_one.sh 1975

echo "**processing issue $1"

YEAR=$1
PATH_TO_YEAR="/DNIssues/$1"
PATH_TO_OUTPUT="/Data/$1"
mkdir -p $PATH_TO_OUTPUT

COUNTER=0
PDF_FILES=$PATH_TO_YEAR/*.pdf
for pdf in $PDF_FILES
do
    PAGE_COUNT=$(strings < $pdf | sed -n 's|.*/Count -\{0,1\}\([0-9]\{1,\}\).*|\1|p' |
sort -rn | head -n 1)
    echo "Processing $pdf file... #$COUNTER"
    echo "Has $PAGE_COUNT pages"
    filename=$(basename -- "$pdf")
    extension="${filename##*.}"
    filename="${filename%.*}"

    PATH_TO_ISSUE_PAGES=$PATH_TO_OUTPUT/$filename
    mkdir -p $PATH_TO_ISSUE_PAGES

    pdfseparate -f 1 -l $PAGE_COUNT $pdf $PATH_TO_ISSUE_PAGES/$filename-page-%d.$extension
    let COUNTER=COUNTER+1
done
```

Pdfseparate, command-line tool, did the actual split. Below shows how you would split a single file named dds-89071.pdf.

```
pdfseparate -f 1 -l 50 dds-89071.pdf output-page%d.pdf
```

2. *Convert* was the other command line tool used to convert each PDF page to PNG with fairly good quality that will be acceptable to tesseract OCR. Given the directory name, the script below recursively converts all PDF files into PNG.

```
#!/bin/bash

# Shell script to do pre processing and convert PDF to PNG
# Usage - ./pre_process_png.sh 1975
# some notes on killing processed i.e. `jobs`, then `kill %1`

# trap ctrl-c
trap "exit" INT

echo "**processing $1 issue pages"

YEAR=$1
PWD=$(pwd) # portable than back-ticks method
PATH_TO_YEAR="$PWD/Data/$1"
PATH_TO_OUTPUT="$PWD/PNG/$1"
mkdir -p $PATH_TO_OUTPUT

ISSUE_COUNTER=1
for d in $PATH_TO_YEAR/*/ ; do #  loop through all the directories in the year
    COUNTER=1 # start from number 2 since page 1 is copyright data from CRL
    PDF_FILES="$d*.pdf"
    for pdf in $PDF_FILES
    do
        echo "Processing $pdf file... #$COUNTER of issue #$ISSUE_COUNTER"
        filename=$(basename -- "$pdf")
        extension="${filename##*.}"
        filename="${filename%.*}"

        OUTPUT_DIR=$(basename $d)
        PATH_TO_PNGS=$PATH_TO_OUTPUT/$OUTPUT_DIR
        mkdir -p $PATH_TO_PNGS

        convert -density 300 $pdf -depth 2 -strip -background white -alpha off
$PATH_TO_PNGS/$filename.png || echo "$pdf" >> ./to_png_errors_$1.log # bash version of
try/catch error
        let COUNTER=COUNTER+1
    done
    let ISSUE_COUNTER=ISSUE_COUNTER+1
done
```

The key parameters here were density and background. Density parameter specifies the number of pixels per inch on the output image, more pixels mean better quality. Background sets a background standardized color on all the images, in this case, white.

```
convert -density 300 dds-89486-page-8.pdf -depth 8 -strip -background white -alpha off dds-89486-
page-3.png
```

3. Since the page 8 was the relevant page of every issue – *Mailbox* column – only 1,513 page images were collected for the letter column.

APPENDIX B: The Remediation Steps

The methods section (see Chapter 4) and application section (see Chapter 5) were all done either on Python scripts or Jupyter Notebook in line with the goals and constraints of this project, albeit as prototypes. The following sequence of steps outline these processes showing links to the script or notebook involved. These computational scripts and notebooks are also located on a public Github repository located at https://github.com/ooduor/lettersiterate. Appendix C describes in detail the generated data by these scripts and notebooks.

The Jupyter Notebooks including results are too big to be appended in full. The deposited Notebooks on GitHub as a public repository have limited results for copyright reasons. However, the Notebook links provided here are meant to allow the reader further explore the data and thesaurus sources.

1. Key Python scripts
    a. `main_steps.py` – Image analysis, segmentation and OCR-ing.
    Below are different command-line options to run the script depending on the mode, that is, whether processing a page image or a folder consisting of pages images. Debug option ensures each step generates the processing done so far on the image and the state of the image.

```
./main_steps.py --image data/PNG/dds-89412-page-8.png –debug

./main_steps.py --image data/test/dds-89491-page-8.png --debug --empty

./main_steps.py --dir data/PNG/1978 --no-empty

./main_steps.py data/PNG/1974/dds-8939 --no-empty
```

    b. `txt_pre_proc.py` – Spell checker, conjoining hyphenated words, newline and tab characters, etc.
    Processes the plain text files from OCR step to perform spell-checking and remove noise characters in preparation for topic modeling and training the Annif indexer.

```
./txt_pre_proc.py --txt data/TXT/1975/dds-89458-page-8-article-3.txt --debug

./txt_pre_proc.py --dir data/TXT/1975 –no-empty
```

    c. `column_dates.py` – Extract the date the column was published as shown on the page image.

Processes the page images to extract the dates indicated at the top right corner of the page. This is the date it was published on the daily.

```
./column_dates.py --dir /PNG/1974/dds-8939 --no-empty
```

```
./column_dates.py --dir /PNG/1975 --no-empty
```

2. Extract bylines from OCR letters (extract_letter_bylines.ipynb)
   a. Uses TXT_XML directory where this Notebook serializes both the plain text and XML files after segmentation and OCR-ing.
   b. Run for each year 1974–1978. Update the `proc_year` variable to process that year.
   c. Output is stored in `bylines_and_files` directory
3. Generate Byline Similarities of names using jellyfish (cluster_similar_reader_names_jellyfish.ipynb)
   a. Use `bylines_and_files` TSVs and output in the same directory
   b. Run for each year 1974–1978.
4. Extract titles from raw letters (extract_letter_titles.ipynb)
   a. Output is found in `titles_and_files` directory located in the same directory as the Notebook.
   b. Run for each year 1974–1978
5. Trace Dates guided by Category keywords (trace_dates_guided_by_keywords.ipynb)
   a. Output is found in `files_and_dates` directory.
   b. Run for each year 1974–1978.
   c. After running for each of the 5 years, run Notebook section titled "Merge all into one".
   d. This combines all the TSVs into one, titled `combined_with_count.tsv`.
6. Average letters extracted per column (mailbox_avg_column_letters.ipynb)
   a. Output is found in `letters_count_per_date` directory.
   b. Run for each year 1974–1978. Each TSV is prefixed with COUNT_ e.g., COUNT_1974.tsv
   c. 'Average' is used here to mean not exact match or ground truth but based on the segmentation and subsequent OCR.
7. Topic Model using NMF (topic_modeling_nmf.ipynb)
   a. Output found in `annif/letters-unesco` folder.
   b. Top 20 files under each topic are stored with they vocabulary terms
   c. Run for each year 1974–1978
8. Split into Training and Evaluation Datasets (split_training_eval_datasets.ipynb)
   a. Output found in `annif/letters-unesco/training` and `annif/letters-unesco/evaluation` folders.
9. Annifier (annif_unseen.py)
   a. Run for each year 1974–1978 to get suggestions of subjects and append subjects into custom XML file.

Directory structure in Figure B.1 below situates the above scripts, notebooks and data where the research was conducted. Data at every stage was stored under the directory on line 8. All the Python scripts (lines 2-5) and Notebooks (lines 52 - 58) have access to the data directory where the page images, XML, plain text, and TSV files are stored.

```
1  ./lettersiterate/
2  ├── annif_unseen.py
```

```
├── column_dates.py
├── main_steps.py
├── txt_pre_proc.py
├── column_years
├── data
│   ├── annif
│   │   ├── letters-custom
│   │   └── letters-unesco
│   │       ├── evaluation
│   │       └── training
│   ├── bylines_and_files
│   ├── column_dates
│   ├── dates_and_quotes
│   ├── files_and_dates
│   ├── letters_count_per_date
│   ├── PNG
│   │   ├── 1974
│   │   ├── 1975
│   │   ├── 1976
│   │   ├── 1977
│   │   └── 1978
│   ├── projects
│   │   ├── letters-omikuji-bonsai-en
│   │   │   └── omikuji-model
│   │   ├── letters-omikuji-parabel-en
│   │   │   └── omikuji-model
│   │   ├── letters-tfidf
│   │   └── omikuji-parabel-en
│   │       └── omikuji-model
│   ├── thesauri
│   │   └── unesco
│   ├── titles_and_files
│   ├── TXT_PROC
│   │   ├── 1974
│   │   ├── 1975
│   │   ├── ...
│   ├── TXT_XML
│   │   ├── 1974
│   │   ├── 1975
│   │   ├── ...
│   ├── vocabs
│   │   ├── correspondence-en
│   │   ├── letters-en
│   │   └── letters-unesco
│   └── XML
│       ├── 1974
│       ├── 1975
│       ├── ...
├── notebooks
│   ├── cluster_similar_reader_names_jellyfish.ipynb
│   ├── create_training_corpus.ipynb
│   ├── extract_date_references.ipynb
│   ├── extract_letter_bylines.ipynb
│   ├── generate_nodes_and_edges.ipynb
│   ├── mailbox_avg_column_letters.ipynb
│   └── extract_letter_titles.ipynb
├── scripts
│   └── create_annif_vocab_from_unesco.py
└── output
```

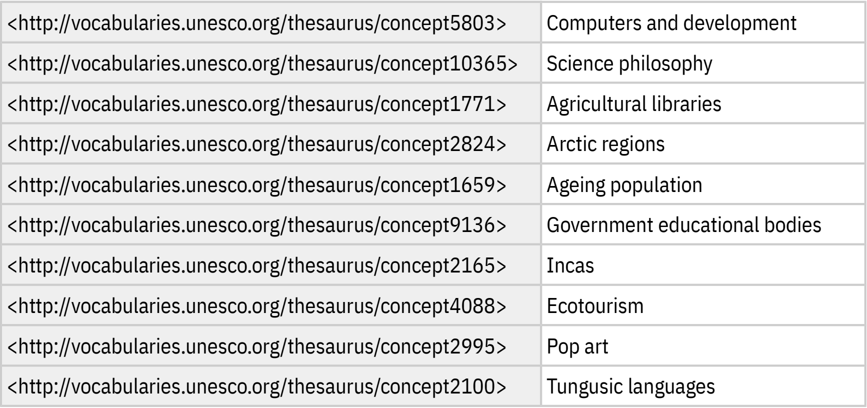
Figure B.1: Structure of the scripts, computational notebooks and data generated in the project.

APPENDIX C: Organization of Data Files

Approximately 20,337 *Mailbox* “letters” were extracted in the end; that is 4,424 in 1978, 4,838 in 1977, 3,886 in 1976, 3,645 in 1975, and 3,544 in 1974. “Letters” because a significant portion of these documents were either too inaccurate to be used intelligibly or form a complete letter. Following the directory “tree” in Appendix B above, I detail the various data types stored in the directories and how they were generated and modified throughout the exploration. In each directory, I describe the file type and structure of the files (of lack thereof). The data are titled by the folder structure beginning from the main directory:

`/data/vocabs`

Contains vocabulary files generated separately from the UNESCO Thesaurus. These files contain the TSV (Tab Separated Values) that are generated in the methods section (see Chapter 4). The table below is a section of unesco-en.tsv file showing how the Thesaurus terms are stored. The URI in the first column is also a key to identify the associated term on the second column. This one of the compatible data extended subject format that Annif uses (see https://github.com/NatLibFi/Annif/wiki/Document-corpus-formats).



<http://vocabularies.unesco.org/thesaurus/concept5803>	Computers and development
<http://vocabularies.unesco.org/thesaurus/concept10365>	Science philosophy
<http://vocabularies.unesco.org/thesaurus/concept1771>	Agricultural libraries
<http://vocabularies.unesco.org/thesaurus/concept2824>	Arctic regions
<http://vocabularies.unesco.org/thesaurus/concept1659>	Ageing population
<http://vocabularies.unesco.org/thesaurus/concept9136>	Government educational bodies
<http://vocabularies.unesco.org/thesaurus/concept2165>	Incas
<http://vocabularies.unesco.org/thesaurus/concept4088>	Ecotourism
<http://vocabularies.unesco.org/thesaurus/concept2995>	Pop art
<http://vocabularies.unesco.org/thesaurus/concept2100>	Tungusic languages

Table C-1: Section of the extracted UNESCO Thesaurus concepts.

The Annif `loadvoc` command loads the vocabulary to be used for training data. The `letters-omikuji-bonsai-en` is one of the several robust algorithms supported by Annif (see https://github.com/NatLibFi/Annif/wiki/Backend%3A-Omikuji) and within the time and resource constraints of the research project.

```
annif loadvoc letters-omikuji-bonsai-en /data/vocabs/unesco-en.tsv
```

`/data/bylines_and_files`

These TSV files were generated at step 3 of Appendix B and were useful for initial network analysis to visualize relationships between *Mailbox* letters using bylines. Two files were generated for each year that was processed to store the “nodes” and “edges” of individual *Mailbox* letters e.g., `1975_jellyfish_nodes.tsv` and `1975_jellyfish_egdes.tsv`. Nodes represent all “names” extracted from byline and publication date, while edges show connections between individual letters. The edges, however, are *Mailbox* letters whose byline names have a high similarity. These files are compatible with Network Analysis software like Gephi to visualize each connection. As shown in Figure C.1, using Gephi made possible to visualize top contributors and a timeline of their contributions.

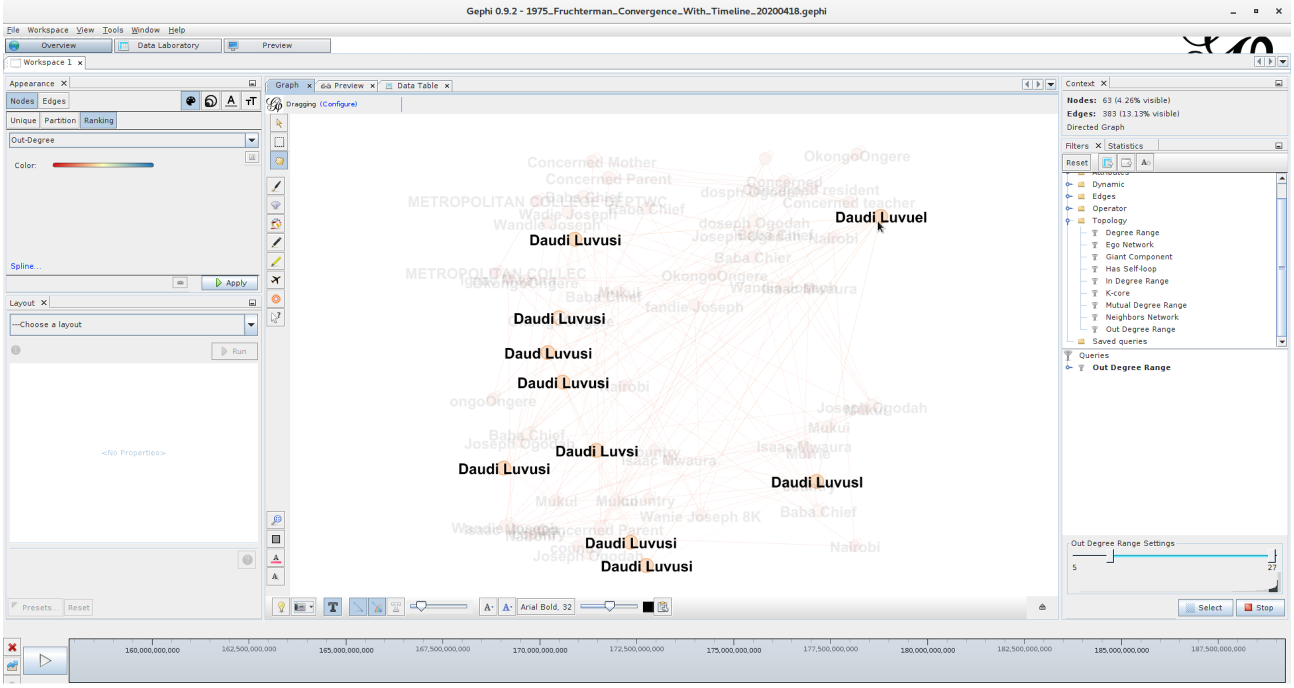


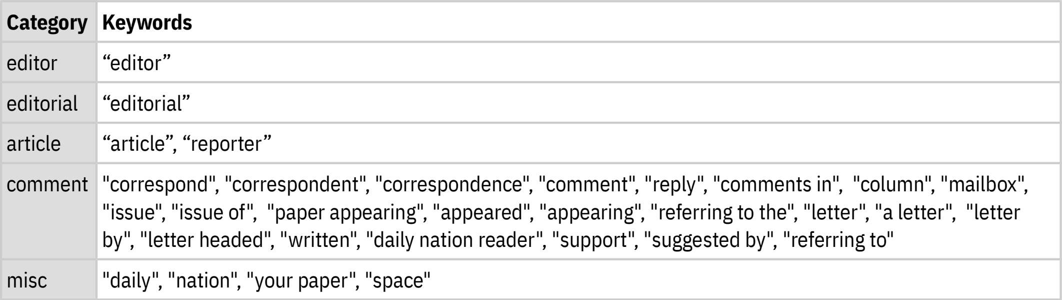
Figure C.1: Screenshot of Gephi visualization of nodes with a high number of connections in the year 1975.

`/data/column_dates`

These TSV files were generated at step 1c of Appendix B. These TSV record the dates each *Mailbox* letter was published for reference. A few corrections were made by hand where the script was unable to extract the date.

/data/files_and_dates

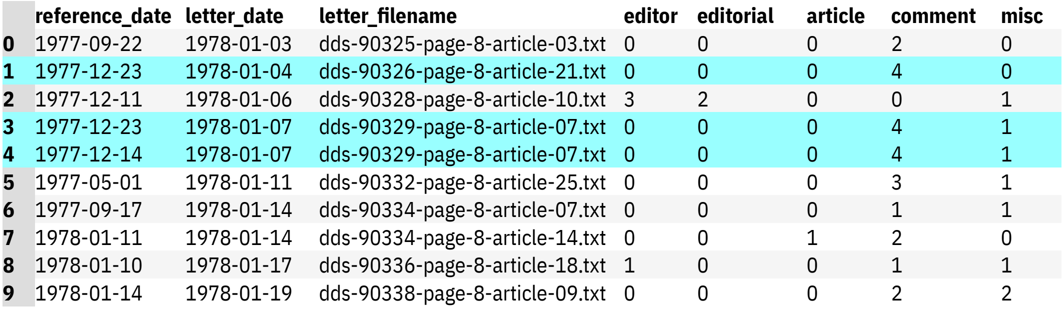
These TSV files were generated at step 5 of Appendix B. Each TSV file records characteristic of plain text files. This includes date references extracted from the body of the *Mailbox* letter, the date the letter was published, and weighting values that might be useful to associate whether it is in references to an editorial, article or *Mailbox* column. The weighting value is determined based on a number of keyword matches in the body text. Table C.1 shows the list of keywords and their categories.



Category	Keywords
editor	“editor”
editorial	“editorial”
article	“article”, “reporter”
comment	"correspond", "correspondent", "correspondence", "comment", "reply", "comments in", "column", "mailbox", "issue", "issue of", "paper appearing", "appeared", "appearing", "referring to the", "letter", "a letter", "letter by", "letter headed", "written", "daily nation reader", "support", "suggested by", "referring to"
misc	"daily", "nation", "your paper", "space"

Table C.1: Keywords used to determine the nature of a *Mailbox* letter. This has a bias towards finding references within the scope of the *Mailbox* column.

Mailbox letters with higher occurrences of comment keywords meant they most likely were referencing a previous letter than an article or editorial. This is crude but useful to narrow down the data to letters of correspondence as highlighted in Table C.2.



	reference_date	letter_date	letter_filename	editor	editorial	article	comment	misc
0	1977-09-22	1978-01-03	dds-90325-page-8-article-03.txt	0	0	0	2	0
1	1977-12-23	1978-01-04	dds-90326-page-8-article-21.txt	0	0	0	4	0
2	1977-12-11	1978-01-06	dds-90328-page-8-article-10.txt	3	2	0	0	1
3	1977-12-23	1978-01-07	dds-90329-page-8-article-07.txt	0	0	0	4	1
4	1977-12-14	1978-01-07	dds-90329-page-8-article-07.txt	0	0	0	4	1
5	1977-05-01	1978-01-11	dds-90332-page-8-article-25.txt	0	0	0	3	1
6	1977-09-17	1978-01-14	dds-90334-page-8-article-07.txt	0	0	0	1	1
7	1978-01-11	1978-01-14	dds-90334-page-8-article-14.txt	0	0	1	2	0
8	1978-01-10	1978-01-17	dds-90336-page-8-article-18.txt	1	0	0	1	1
9	1978-01-14	1978-01-19	dds-90338-page-8-article-09.txt	0	0	0	2	2

Table C.2: Section of 1978 TSV showing the first 10 records.

/data/titles_and_files

These TSV files were generated for each year at step 4 of Appendix B. Each TSV file contain titles extracted from *Mailbox* letters. The titles are extracted from plain text files as the first paragraph whenever it met certain conditions specifically 1) length (fewer than 100 characters) and 2) if it is not the only paragraph in the text file.

title	txt_name	title_length
Poodle that never came	dds-89398-page-8-article-01.txt	23
Let's kick them_all _out	dds-89398-page-8-article-04.txt	25
Coast TV	dds-89398-page-8-article-05.txt	9
DRUG ABUSE 'A CANCER THAT KILLS SLOWLY	dds-89398-page-8-article-07.txt	39
The future is not rosy	dds-89398-page-8-article-09.txt	23
The Arabs	dds-89398-page-8-article-11.txt	10
Hoanting still goes on	dds-89398-page-8-article-14.txt	23
We are afraid WORKS ON in the dark	dds-89398-page-8-article-16.txt	35
We must 'avoid being ruled by the heart':	dds-89398-page-8-article-19.txt	42
THE NANDI	dds-89398-page-8-article-23.txt	10

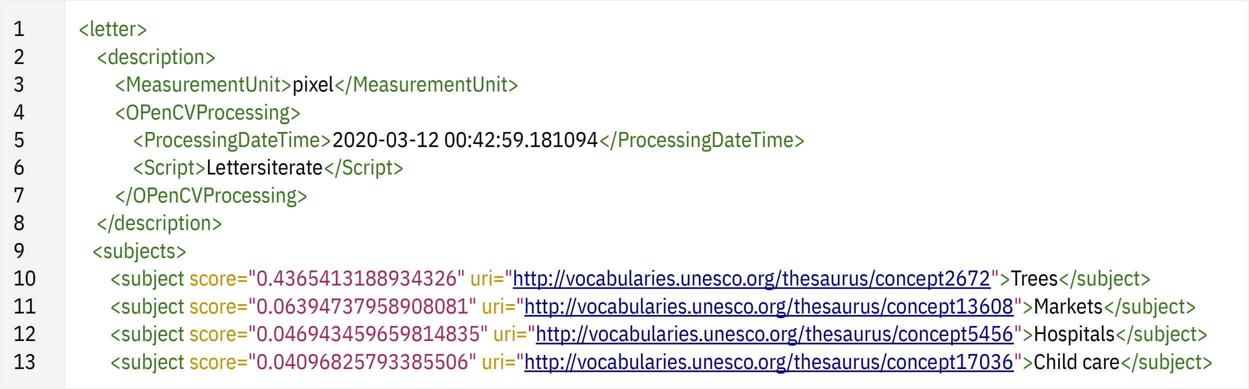
Table C.3: Section of 1975 TSV showing the first 10 records showing titles for *Mailbox* titles.

/data/TXT_XML

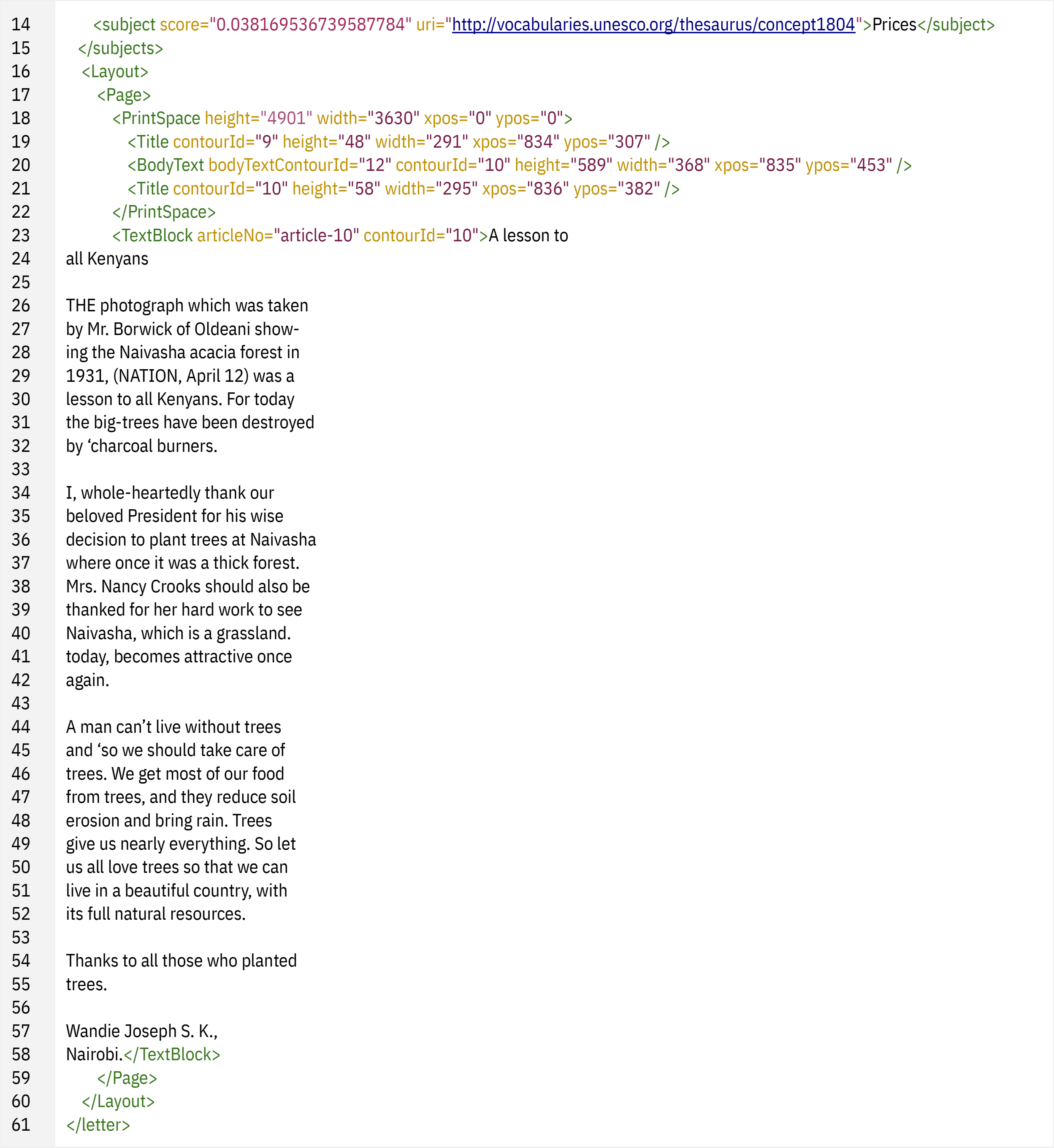
These plain text and XML files are generated in the initial step of Appendix B. Contains 20,327 plain text and an equal number of XML files extracted by the OCR processing.

/data/XML

These XML files are generated at step 9 of Appendix B. Contains 14,847 XML files that also contains embedded subject indexes. The subject indexes are described using `annif_unseen.py` script which uses trained model to suggest at least 4 subjects. Figure C.3 shows an XML file with subject indexes (lines 10 - 14).



```
1   <letter>
2     <description>
3       <MeasurementUnit>pixel</MeasurementUnit>
4       <OPenCVProcessing>
5         <ProcessingDateTime>2020-03-12 00:42:59.181094</ProcessingDateTime>
6         <Script>Lettersiterate</Script>
7       </OPenCVProcessing>
8     </description>
9     <subjects>
10      <subject score="0.4365413188934326" uri="http://vocabularies.unesco.org/thesaurus/concept2672">Trees</subject>
11      <subject score="0.06394737958908081" uri="http://vocabularies.unesco.org/thesaurus/concept13608">Markets</subject>
12      <subject score="0.046943459659814835" uri="http://vocabularies.unesco.org/thesaurus/concept5456">Hospitals</subject>
13      <subject score="0.04096825793385506" uri="http://vocabularies.unesco.org/thesaurus/concept17036">Child care</subject>
```



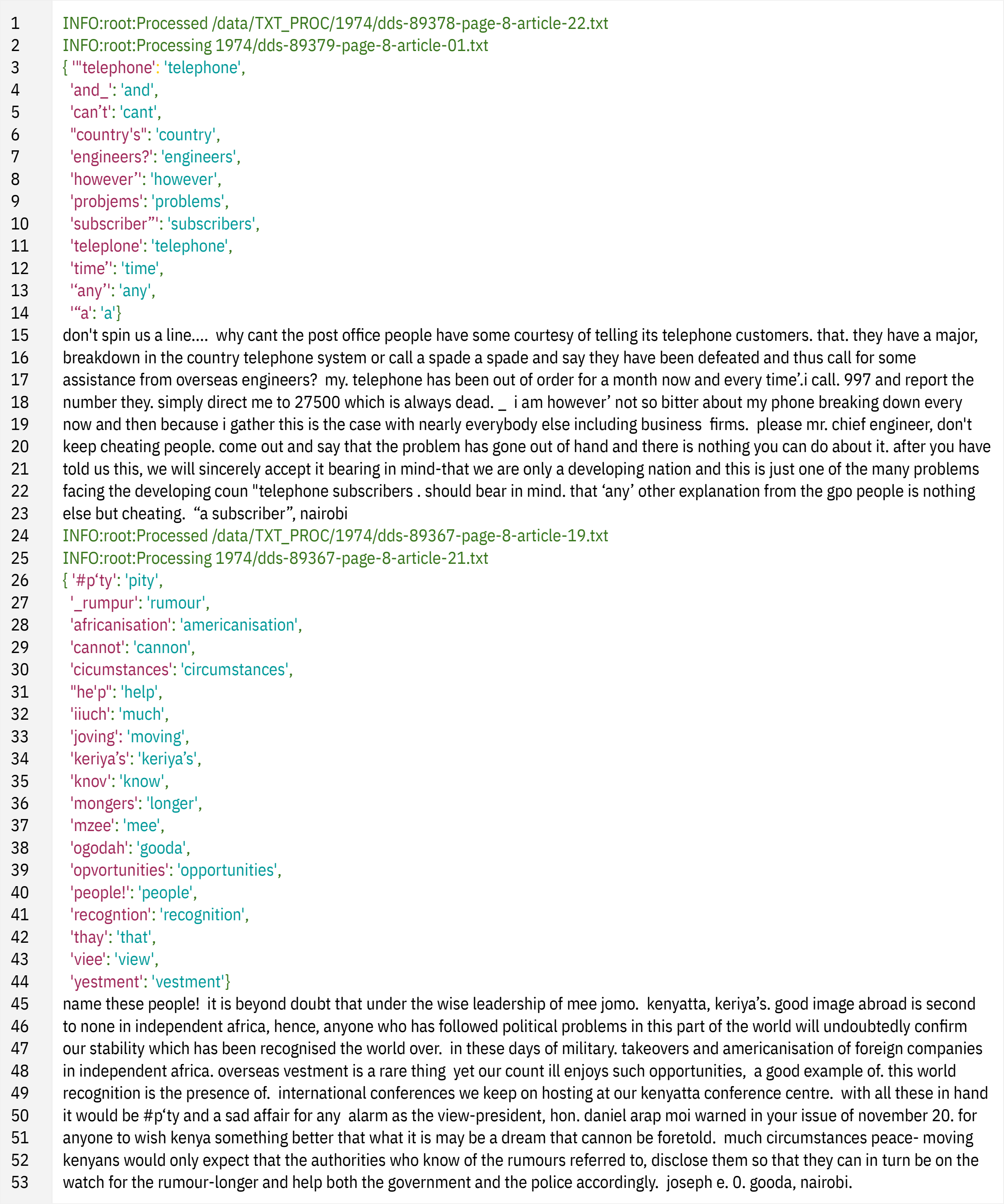
```
    <subject score="0.038169536739587784" uri="http://vocabularies.unesco.org/thesaurus/concept1804">Prices</subject>
  </subjects>
  <Layout>
    <Page>
      <PrintSpace height="4901" width="3630" xpos="0" ypos="0">
        <Title contourId="9" height="48" width="291" xpos="834" ypos="307" />
        <BodyText bodyTextContourId="12" contourId="10" height="589" width="368" xpos="835" ypos="453" />
        <Title contourId="10" height="58" width="295" xpos="836" ypos="382" />
      </PrintSpace>
      <TextBlock articleNo="article-10" contourId="10">A lesson to
all Kenyans

THE photograph which was taken
by Mr. Borwick of Oldeani show-
ing the Naivasha acacia forest in
1931, (NATION, April 12) was a
lesson to all Kenyans. For today
the big-trees have been destroyed
by ‘charcoal burners.

I, whole-heartedly thank our
beloved President for his wise
decision to plant trees at Naivasha
where once it was a thick forest.
Mrs. Nancy Crooks should also be
thanked for her hard work to see
Naivasha, which is a grassland.
today, becomes attractive once
again.

A man can’t live without trees
and ‘so we should take care of
trees. We get most of our food
from trees, and they reduce soil
erosion and bring rain. Trees
give us nearly everything. So let
us all love trees so that we can
live in a beautiful country, with
its full natural resources.

Thanks to all those who planted
trees.

Wandie Joseph S. K.,
Nairobi.</TextBlock>
    </Page>
  </Layout>
</letter>
```

Figure C.3: XML file *dds-89807-page-8-article-10.xml* from 1976 with suggested subject indexes.

`/data/TXT_PROC`

These XML files are generated at step 1b of Appendix B. Contains 20,327 plain text files that have been processed with automated spell checker and remove noise characters in the text. As shown in Figure C.4, in lines 3-14 and 26-44, words on the left were replaced by words on the right as the spell-checker script processed each plain text file.

Below these line blocks is the resulting text with corrected and in some cases, miscorrected words. However, this was an important step as it correct trivial typographical errors and improves the subject indexing to some extent.

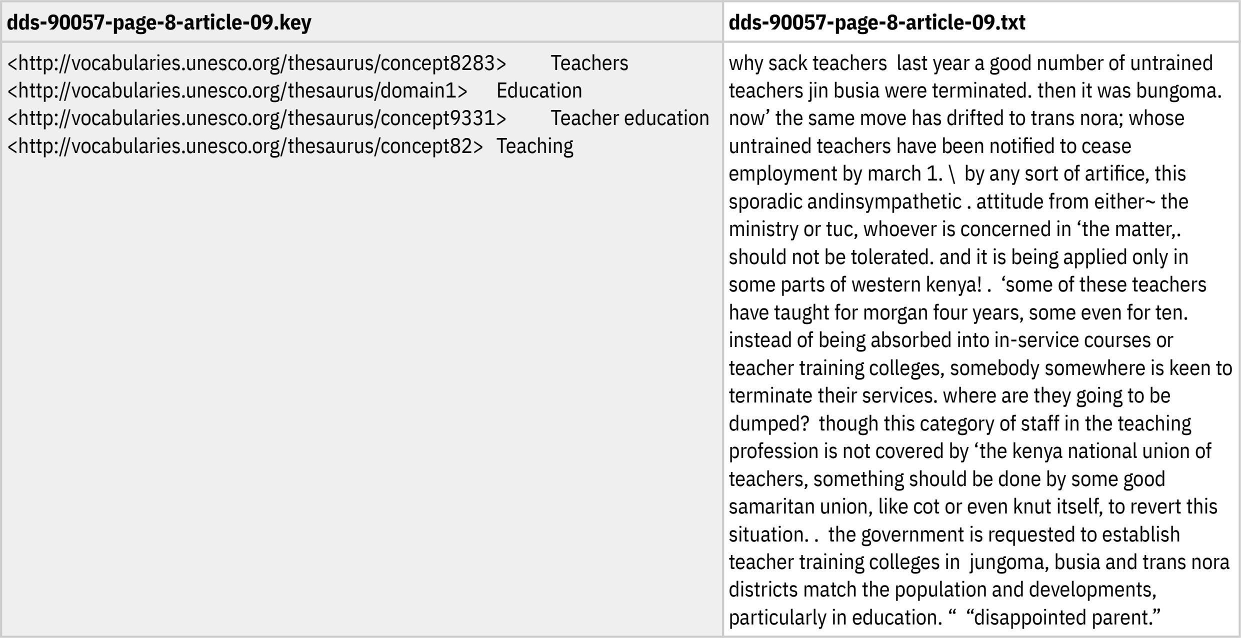


```
INFO:root:Processed /data/TXT_PROC/1974/dds-89378-page-8-article-22.txt
INFO:root:Processing 1974/dds-89379-page-8-article-01.txt
{ '"telephone': 'telephone',
  'and_': 'and',
  'can’t': 'cant',
  "country's": 'country',
  'engineers?': 'engineers',
  'however’': 'however',
  'probjems': 'problems',
  'subscriber”': 'subscribers',
  'teleplone': 'telephone',
  'time’': 'time',
  '“any’': 'any',
  '“a': 'a'}
don't spin us a line....  why cant the post office people have some courtesy of telling its telephone customers. that. they have a major,
breakdown in the country telephone system or call a spade a spade and say they have been defeated and thus call for some
assistance from overseas engineers?  my. telephone has been out of order for a month now and every time’.i call. 997 and report the
number they. simply direct me to 27500 which is always dead. _  i am however’ not so bitter about my phone breaking down every
now and then because i gather this is the case with nearly everybody else including business  firms.  please mr. chief engineer, don't
keep cheating people. come out and say that the problem has gone out of hand and there is nothing you can do about it. after you have
told us this, we will sincerely accept it bearing in mind-that we are only a developing nation and this is just one of the many problems
facing the developing coun "telephone subscribers . should bear in mind. that ‘any’ other explanation from the gpo people is nothing
else but cheating.  “a subscriber”, nairobi
INFO:root:Processed /data/TXT_PROC/1974/dds-89367-page-8-article-19.txt
INFO:root:Processing 1974/dds-89367-page-8-article-21.txt
{ '#p‘ty': 'pity',
  '_rumpur': 'rumour',
  'africanisation': 'americanisation',
  'cannot': 'cannon',
  'cicumstances': 'circumstances',
  "he'p": 'help',
  'iiuch': 'much',
  'joving': 'moving',
  'keriya’s': 'keriya’s',
  'knov': 'know',
  'mongers': 'longer',
  'mzee': 'mee',
  'ogodah': 'gooda',
  'opvortunities': 'opportunities',
  'people!': 'people',
  'recogntion': 'recognition',
  'thay': 'that',
  'viee': 'view',
  'yestment': 'vestment'}
name these people!  it is beyond doubt that under the wise leadership of mee jomo.  kenyatta, keriya’s. good image abroad is second
to none in independent africa, hence, anyone who has followed political problems in this part of the world will undoubtedly confirm
our stability which has been recognised the world over.  in these days of military. takeovers and americanisation of foreign companies
in independent africa. overseas vestment is a rare thing  yet our count ill enjoys such opportunities,  a good example of. this world
recognition is the presence of.  international conferences we keep on hosting at our kenyatta conference centre.  with all these in hand
it would be #p‘ty and a sad affair for any  alarm as the view-president, hon. daniel arap moi warned in your issue of november 20. for
anyone to wish kenya something better that what it is may be a dream that cannon be foretold.  much circumstances peace- moving
kenyans would only expect that the authorities who know of the rumours referred to, disclose them so that they can in turn be on the
watch for the rumour-longer and help both the government and the police accordingly.  joseph e. 0. gooda, nairobi.
```

Figure C.4: A section of the spell-checker output of events during processing on plain text files.

/data/annif

The “.key” files are generated at step 7 of Appendix B and stored together with plain text files copied from spell-checker step. Data organized for Annif to use in training and evaluating the assignment of subject terms. As the contents in Table C.4 show, each plain text file (right column) is accompanied by a *.key* (left column) file with thesaurus terms that describe topics in the document.



dds-90057-page-8-article-09.key	dds-90057-page-8-article-09.txt
<http://vocabularies.unesco.org/thesaurus/concept8283> Teachers <http://vocabularies.unesco.org/thesaurus/domain1> Education <http://vocabularies.unesco.org/thesaurus/concept9331> Teacher education <http://vocabularies.unesco.org/thesaurus/concept82> Teaching	why sack teachers last year a good number of untrained teachers jin busia were terminated. then it was bungoma. now’ the same move has drifted to trans nora; whose untrained teachers have been notified to cease employment by march 1. \ by any sort of artifice, this sporadic andinsympathetic . attitude from either~ the ministry or tuc, whoever is concerned in ‘the matter,. should not be tolerated. and it is being applied only in some parts of western kenya! . ‘some of these teachers have taught for morgan four years, some even for ten. instead of being absorbed into in-service courses or teacher training colleges, somebody somewhere is keen to terminate their services. where are they going to be dumped? though this category of staff in the teaching profession is not covered by ‘the kenya national union of teachers, something should be done by some good samaritan union, like cot or even knut itself, to revert this situation. . the government is requested to establish teacher training colleges in jungoma, busia and trans nora districts match the population and developments, particularly in education. “ “disappointed parent.”

Table C.4: A section of the spell-checker output of events during processing on plain text files.

/data/thesauri

The UNESCO Thesaurus downloads retrieved from http://vocabularies.unesco.org/exports/thesaurus/latest/. The `unesco-thesaurus.rdf` in particular is parsed to create the vocabulary used in training the Annif model for subject indexing and assigned to *Mailbox* letters.