Data Capture Automation in the South African Deeds Registry using Optical Character Recognition (OCR)

Ashleigh Favish

A dissertation submitted to the Faculty of Commerce, University of Cape Town, in partial fulfilment of the requirements for the degree of Master of Philosophy.

Supervisor: Co-Pierre Georg

December 25, 2018

MPhil specializing in Financial Technology,
University of Cape Town.

Declaration

I declare that this dissertation is my own, unaided work. It is being submitted for the Degree of Master of Philosophy to the University of Cape Town. It has not before been submitted for any degree or examination.

Signed by candidate

Ashleigh Favish
25 December 2018

Abstract

The impact of apartheid on land registration is still evident within South Africa. The Deeds Registry is facing a current backlog in registering an estimated 900,000 title deeds.

Providing formal ownership, through title, is seen as necessary for unlocking the ‘dead capital’ of unregistered property, fostering access to capital markets and poverty alleviation. Within the current legislative framework, the Deeds Registry only accepts paper documents, which introduces inefficiencies.

To increase the number of deeds processed per day, automation of manual data capture is tested using an OCR pipeline. To adapt to the linguistics used in title deeds, text analysis and parsing is done using Regex. Uploading the scanned title deeds onto IPFS is as an additional security measure included in the pipeline. Previous research has failed to apply these techniques to formal land registration or other South African government institutions.

The preliminary results show that this pipeline has an overall accuracy of 89.6%. This represents the comparison of the expected output to the output extracted using OCR. The results are significantly less accurate when classifying handwritten and stamped information. Thus, further measures are required to increase accuracy for these fields. The OCR accuracy was 98.3% for the fields extracted from typed text characters. This is within the accuracy range of manual data capture. A secondary quality check, which is currently done on manual data capture, would still be necessary to ensure accuracy of inputs. Overall it appears that this application would be appropriate for incorporation into the Deeds Registry to streamline their processes while ensuring title deed validity.

Acknowledgements

I am eternally grateful to G-d for giving me the strength, health, balanced attitude and motivation to complete this research. All my accomplishments are due to you.

I would like to acknowledge my supervisor Co-Pierre Georg for his constant support, availability, insight and encouragement throughout the whole process. Your input was invaluable, and I was humbled by your ability to think on a grand scale and see the possibilities of improving the world.

Thank you to Allan Davids for adding focus and direction to my dissertation. I appreciate all the time taken to guide me and introduce me to interesting people along the way.

I could not have reached this stage in my studies without my beloved parents. Thank you for your unbridled support and love throughout the years from pre-school all the up to a post-graduate level. There will never be words to express my appreciation.

There have been numerous people who have helped me along the way with my research. I am extremely grateful for all the time, support and resources that you have provided me, whether indirectly or directly.

List of Abbreviations and Acronyms

The Act	-	Deeds Registries Act 47 of 1937
AES	-	Advanced Encryption Standard
AI	-	Artificial Intelligence
CSV	-	Comma-Separated Values
CV2	-	Computer Vision 2
DAG	-	Directed Acyclic Graph
DPI	-	Dots Per Inch
ECTA	-	Electronic Communications and Transactions Act
e-DRS	-	Electronic Deeds Registry System
HMM	-	Hidden Markov Models
ICT	-	Information and Communications Technology
ID	-	Identity Document
IV	-	Initialization Vector
IPFS	-	InterPlanetary File System
LSTM	-	Long Short-Term Memory
OCR	-	Optical Character Recognition
OpenCV	-	Open Source Computer Vision Library
PDF	-	Portable Document Format
Regex	-	Regular Expressions
RAM	-	Random Access Memory
RDP	-	Reconstruction and Development Programme

Table of Contents

List of Figures

List of Tables

Chapter 1

1. Introduction

1.1 Background

Property is generally the most important asset held by households. In most Western countries, home ownership is formalized through records at a deeds registry. Formal ownership, often through title deeds, is widely seen as a way to allow properties to become wealth building assets. This contributes to poverty alleviation, economic growth and increased economic participation. Through access to capital markets and increased trust among market participants, the unlocked potential of property can be accessed via the formalization of property rights. This has led many emerging countries to focus on creating formal property registers.

The legacy of apartheid within South Africa is apparent in many service delivery areas, including property registration. With land ownership for most of the population limited during apartheid, insecure property rights were the norm. These restrictions were lifted as democracy set in. The government also implemented large-scale subsidized housing programmes. Therefore, more people entered the property market and more properties needed registration. This resulted in property transactions spiking and a backlog in the registration of title deeds. Currently the Deeds Registry has approximately 7.1 million properties registered. However, they are challenged by a backlog of an estimated 900,000 title deeds requiring formal registration (Centre for Affordable Housing Finance in Africa, 2017). This backlog is likely to increase, as the government anticipates the number of registered land parcels to reach 20 million through land reform measures (Minister of Rural Development and Land Reform 2017). The Deeds Registry is prevented from increasing capacity owing to budgetary constraints. Increasing employee productivity by incorporating appropriate technology is a possible solution to reduce the backlog economically.

As the lodgement process of title deeds remains paper-based under current legislation, there is room for efficiency gains through automation. This paper examines the accuracy and efficiency of using Optical Character Recognition (OCR) technology to automate the data capture functions of the Deeds Registry. This is to increase the number of deeds registered per day and help ease the backlog. As title deed validity is of prime importance to a deeds registry,

incorporating a security layer using InterPlanetary File System (IPFS) is also discussed.

1.2 Objective

This research investigates the accuracy and viability of using an OCR pipeline to perform automatic data capture on a scanned sample of title deeds. The process starts with the encrypted Portable Document Format (PDF) being added to IPFS for title deed validity and security. This pipeline pre-processes the deed before OCR, in order to increase accuracy of text classification. OCR is then performed using open source software Tesseract which converts the images into text. Subsequently, the semi-structured text output is parsed based on common phrases and patterns using Regular Expressions (Regex), a pattern matching notation. The relevant fields are then extracted into a Comma-Separated Values (CSV) file for comparison with the manually captured data. The unique hash of the PDF, from the IPFS upload, is also incorporated into the CSV file. The overall process is aimed at effectively digitising and extracting data from semi-structured scanned documents in a secure manner.

1.3 Summary of Main Findings

The data extraction has an overall accuracy of 89,6% for the fields being extracted. The accuracy level increases to 98.3% for OCR extraction of typed data. The main errors occur for text which is either stamped or handwritten as the OCR recognition is impacted by smudges, skewness and general low quality of characters. The average time taken to run the program, before code optimization, is shorter than that of manual data capture. Based on a computer's ability to work around the clock, the application can process approximately 331 more deeds per day than a manual typist. As errors occur both in manual and automatic data capture, an accuracy check would be required for both options. The uploading of files to IPFS has negligible effects timewise, with the encryption being successful for uploaded files. Overall, the OCR pipeline is effective in extracting typed data from title deeds and therefore can potentially bring efficiencies to the Deeds Registry through automation. For text commonly stamped onto the deed, recommendations such as flagging empty fields or inputs with unusual formats can be incorporated to alert employees of errors.

These findings are to be viewed in a context of a pilot project where a sample of 21 deeds with registration after 2000 were tested. Fifteen of these deeds were simulated based on a conveyancers template. Not all property types or ownership structures were tested, and all deeds were written in English.

1.4 Chapter Layout

This dissertation is categorized into six chapters. Chapter 1, the current chapter, serves to provide a background on the title deed backlog affecting South Africa and how OCR may be a potential aid in reducing the backlog through increasing efficiency. Chapter 2 delves into the literature, describing the current and historical mode of operations of deeds registries as well as examining the role a deeds registry plays in society. This is followed by a description of OCR, its history, processes and various practical applications of OCR. Chapter 3 looks at the institutional environment within South Africa regarding service delivery backlogs and technology implementations within the context of a post-apartheid South Africa. The Deeds Registry in South Africa is then discussed in terms of legislation, history, the key processes followed once a deed is lodged and a general evaluation of these processes. Chapter 4 explains the data that is used for testing the OCR application, as well as the methodology then applied to the data. The results of the pilot project are discussed in Chapter 5 along with the limitations, problem areas, possible recommendations and areas of future research. The conclusion is presented in Chapter 6.

Chapter 2

2. Literature Review

2.1 Deeds Registry

2.1.1 What is a Deeds Registry?

A deeds registry records ownership of immovable property as well as transfers of property. This information provides certainty of title which is a requirement for a title deed to have value (Kochan, 2013). As a repository of information, a deeds registry keeps a secure record of all property related documents for the general public to query and access (Amadi-Echendu, 2017). The title deed forms the legal documentation reflecting ownership of a property and is replaced by a new title deed when property is transferred (Gordon, Nell & Di Lollo, 2011) .

2.1.2 Evolution of the Cadastral System

Recording land ownership has been a part of civilization for thousands of years with Ancient Egypt creating a registry in approximately 3000BC (Larsson, 1991). Other land recording systems were present in China and Rome for the purposes of taxation and various fiscal records respectively (Ting & Williamson, 1999). In later centuries, fiscal cadastres also emerged in recording land ownership in Europe as the mechanism of land taxation (Henssen, 1975). Napoleon instructed the mapping of land ownership, registering of property transfers and title deeds based on partitioning the land into parcels (Ting & Williamson, 1999). These registry records “lay the foundations for modern-day cadastral systems”. The trend for land taxation started diminishing in the 19th and 20th century, prompting governments to link the cadastral systems to title registration and ownership (Williamson, 1997). Williamson further discusses how this past century has witnessed cadastral systems increasing in sophistication to become land information systems through the increased availability of computer technology.

2.1.3 Deeds Registry Importance

Providing property ownership records allows for market facilitation within real estate transactions by establishing the ground truth of ownership (Kochan, 2013). This is beneficial information to society at large and specifically borrowers and

lenders as an efficient and fair market is created (Korngold, 2008). De Soto (2000) argues that when possession rights of assets are poorly documented, these assets are considered “dead capital” by being difficult to buy, sell, value or become an investment. Furthermore, he claims that without formal property rights, nations remain in poverty as the poor cannot access their assets’ value for collateral or business expansion. Particularly in emerging countries, where land is often unregistered, title deeds are seen as a way to reduce poverty and promote investment (Toulmin, 2009). For example, Ghana experienced improved investment facilitation through formalizing land rights (Besley, 1995). Realizing this, the South African national housing policy includes enhancing access to title as a fundamental policy (Department of Housing, 2004).

A deeds registry can therefore be seen as a vital piece of economic infrastructure which allows for functioning of a free market system (Andreas van Wyk, 2018). These economic benefits will only be substantial if there is a functioning financial market and existence of economic incentives for property investment. However social benefits may arise in the form of protecting the rights for socially weaker group and poverty alleviation (Feder & Feeny, 1991).

2.2 Optical Character Recognition

2.2.1 Definition

OCR is a technology that converts images or scans of text into machine-readable code (Pandey et al., 2017) using various forms of pattern recognition, artificial intelligence and machine vision (Bhatia, 2014). Through this conversion it is possible to edit, search and capture the text data extracted. OCR is a technology which functions as the computer’s reading ability, although its performance is below that of humans (Mithe, Indalkar & Divekar, 2013).

2.2.2 Background

OCR is a concept which has been studied extensively with the first patent on the topic filed in 1929 by Tausheck (1935). This idea became a possibility in the 1950s with advancements in computing (Mori, Suen & Yamamoto, 1992) and the first commercial system installation in 1955 (Patel, Patel & Patel, 2012). The main advancements in OCR occurred after 1990 where pattern recognition techniques were effectively combined with artificial intelligence (AI), increasingly powerful computers and high accuracy electronic equipment (Arica & Yarman-Vural, 2001). Therefore in the last 20 years, OCR has become available online, on mobile applications as well as commercial and open source systems (Asif et al., 2014). OCR has extensive applications such as license plate recognition

(Anagnostopoulos et al., 2006), mail sorting (Downton & Leedham, 1990), digitising historical records (Cojocaru et al., 2016) and document reading (Bhatia, 2014). Through OCR, scanned documents become fully searchable with automatic data entry into a database being possible (Asif et al., 2014).

2.2.3 Process

The general process of OCR consists of pre-processing, segmentation, feature extraction and classification of letters and words (Bhatia, 2014; Singh et al., 2010) as shown in Figure 1. The pre-processing generally consists of a series of operations to enhance the image quality such as noise reduction (Bhatia, 2014). Once text blocks are distinguished, various segmentation methods are applied to partition these blocks into lines, words and individual characters (Singh et al., 2010). In order to classify each character, specified features describing the character are extracted such as curvature, directional features and zoning (Bhatia, 2014). Many techniques have been applied to classify characters based on the extracted features such as support vector machines (Ashwin & Sastry, 2002), neural networks and k-nearest neighbours (Park, Govindaraju & Srihari, 2000).

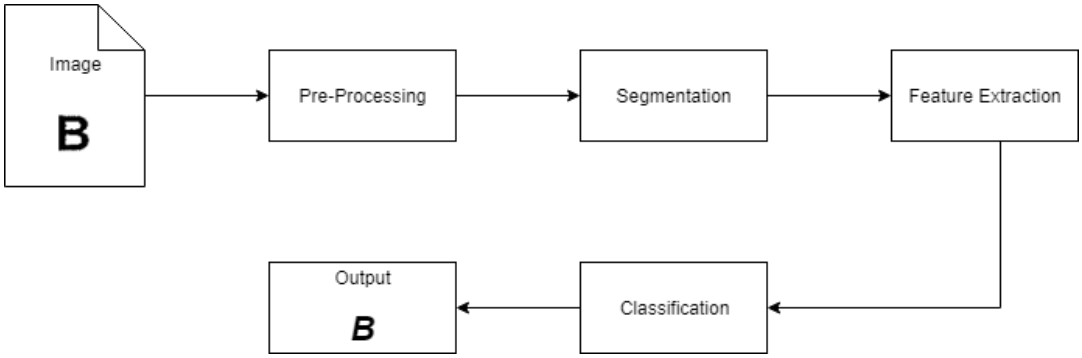


Figure 1: General OCR Process

2.2.4 Accuracy

OCR has recognition capabilities for both handwritten and typed text (Mithe, Indalkar & Divekar, 2013). Accuracy of OCR is dependent on the quality of the input image with accuracy ranging from 71% to 98% (Patel, Patel & Patel, 2012). But with the advancement of OCR applications, strong results are achievable with modern texts (Cojocaru et al., 2016). Due to the variability of handwriting styles, OCR is more complex for handwritten text (Sonkusare & Sahu, 2016) which reduces accuracy. With the rise in Neural Networks for pattern recognition within OCR systems (Bhatia, 2014), the accuracy rates are expected to increase as the models improve with more data. However, more research needs to be done to

improve recognition capabilities for noisy images in both handwritten and machine printed text (Arica & Yarman-Vural, 2001).

2.2.5 Document Processing Applications

Milligan (2013) asserts that OCR applications are designed principally for digitising conventionally formatted documents often found in corporate and legal settings. However, OCR is often applied to historical documents for preservation and searchability as seen by large-scale digitisation projects of newspaper collections in, for example, Austria (Cojocaru et al., 2016). With the high costs, maintenance and space requirements of physical documents, organizations should consider a paperless office (Ugale, Patil & Musande, 2017). Ugale et al. shows how OCR would be utilized in deploying a paperless document management system, after scanning of paper documents is complete. Once a document has OCR applied to it, data capture is possible. This was shown in a pilot project by Biondich et al. (2002) where OCR was tested in the extraction of structured medical data from paper forms. The findings showed an accuracy in recognition of 92.4% with the system approximately three times as fast as the manual data capture. Other recent research looks at using OCR and other extraction heuristics to extract information from grocery receipts (Ullah et al., 2018).

2.2.6 OCR in a South African Context

Within a South African context, there has been limited research on using OCR. The majority of research focuses on developing models to recognize less common African languages such as Tshivenda (Hocking & Puttkammer, 2016) or Setswana (Kotzé & Wolff, 2017). South African Centre for Digital Language Resources (SADiLaR) is a recent initiative created by the South African Department of Science and Technology (Roux et al., 2016). SADiLaR has a digitisation effort in relation to providing digital language resources for the 11 official languages of South Africa. However, there is a gap in OCR research in relation to automating data capture in government departments.

2.2.7 OCR within Land Registration Systems

There has been research on several areas of digitising documents related to land registration. Chen et al. (1999) presented a system on automatic extraction of information from scanned images of Chinese land register maps using neural networks. A patent application was filed detailing the digitising of non-standard documents such as mortgages, liens and title deeds in order to create easily

searchable property records (Hayes, Beres & Diesch, 2005). A possible transcription method to be used within this invention was OCR. Other patents using OCR for data verification of property records have been filed (Newcomer et al., 2013). In his research into computerising land titling systems within the Australian context, Lang remarks how optical scanners can be incorporated for the improvement of data entry (1981). Even though various papers have mentioned OCR in relation to land titling systems, the research available does not go into detail about how to apply these systems and the performance of such systems.

2.3 Conclusion

It is evident from the research that formal registration of property has been a part of civilization for thousands of years and continues to play a vital role in property markets. As deed registries become more sophisticated, increased levels of computing are used. Since OCR is often used for digitising paper documents, it is not surprising that OCR is suggested as a means of digitising property records and title deeds.

With increasing access to title deeds being a fundamental principle of the South African housing policy, the Deeds Registry within a post-apartheid context will be explored. After analysing the processes of the Deeds Registry, using OCR as a possible way to improve efficiency will be discussed.

Chapter 3

3. Institutional Setting in South Africa

3.1 General Background

After discussing in previous chapters, the history, function and importance of a deeds registry as well as introducing OCR and its multiple applications, this chapter investigates how these ideas relate to the South African context. First the impact of apartheid on society and service delivery policy is explored. The capacity constraints within government departments are assessed with a focus on the high volume of paper documents. The chapter then focuses on how this institutional setting affects property and land registration with a specific analysis of the Deeds Registry, its processes and how OCR could be used to reduce manual data capture times.

3.1.1 Apartheid Legacy

South Africa's history of colonialism and apartheid resulted in an unequal society divided on racial grounds. As democracy became a reality in 1994, the new government was faced with the daunting task of transforming the previous government system into an inclusive and non-racial system of service delivery (Ndou & Sebola, 2016). To address these challenges the government issued the White Paper on the Transformation of the Public Service in 1997 (Nengwekhulu, 2009), which looked specifically at improving the efficiency and effectiveness of service delivery (Department of Public Service and Administration, 1997). However, the post-apartheid expectations concerning service delivery far outstripped the government's material and human resources (Nengwekhulu, 2009). With such a sudden volume increase in people serviced by the government, a service delivery backlog developed across the board. Many of the challenges faced were in relation to overcoming the legacy of apartheid such as a shortage of skilled administrative staff (Koma, 2012). There have been substantial inroads in proving public services, especially to previously underserved demographics, although significant challenges remain. This is particularly true in rural areas and former homelands which have large disparities in service delivery (Powell, 2012).

3.1.2 Capacity Constraints

The government provides services to the public, despite limited financial and material resources. Aims to build public sector capacity are often faced with budgetary constraints limiting employee training (Fourie, 2001). Financial difficulties are pervasive within municipalities, with many having difficulties in providing service delivery in a cost-effective and sustainable manner (Beyers, 2016). With local government being tasked with "doing more with less resources" (Powell, 2012), there appears to be room to increase the productivity of government employees through technology and automation.

3.1.3 High Volumes of Paper Documents and Data Capture

Government departments are often overloaded with paper forms such as drivers' license, marriage and death certificate applications (Netchaeva, 2002). This data is often captured manually into the relevant databases but is time-consuming and error-prone. When a citizen applies for an Identity Document (ID) at Home Affairs, anecdotal evidence shows that it is common for long queues to be faced by applicants, prompting Malusi Gigaba, former Minister of Home Affairs, to declare a "war on queues" [1]. The government has recognized the role information and communications technology (ICT) has in increasing government efficiency, resulting in investment in ICT infrastructure (Mutula & Mostert, 2010). This is apparent through the development of online government services such as e-filing[2]. However, there still is a significant technological divide, as many South Africans do not have access to computers and internet, especially in rural areas where ICT infrastructure is limited. As approximately 45% of South Africans live within rural areas (ibid.), paper forms are expected to be commonplace for the foreseeable future.

3.1.4 Property and Land Registration

Apartheid resulted in weak and insecure property rights for black South Africans, with millions of people living in informal settlements or under communal tenure (Cousins & Hornby, 2005). With the laws restricting property ownership lifted at the conclusion of apartheid, there was a sudden large increase in people with rights to engage in the property market. The government also engaged in

[1] These two articles show the common occurrence of long queues at Home Affairs
https://www.iol.co.za/dailynews/no-end-in-sight-for-home-affairs-long-queues-14416900
https://www.news24.com/SouthAfrica/News/home-affairs-queues-no-quick-fix-20180424

[2] This is South African Revenue Services portal for online tax filing. Manual forms at SARS branches are also accepted.

various land reform policies such as the Reconstruction and Development Programme (RDP), which included the delivery of subsidized housing and the accompanying title deeds. By 2011, approximately 3.25 million housing units and serviced sites had been developed (Gordon, Nell & Di Lollo, 2011). This contributed to transactions spiking, which combined with capacity constraints and other factors, created a large backlog in property registrations

Although housing delivery is at the forefront of the national agenda by the government (Lekota, cited in Lubbe, 2013), the backlog of title deeds persists with an estimate of 900,000 title deeds which still need to be formally registered (Centre for Affordable Housing Finance in Africa, 2017). A significant portion of these title deeds related to RDP houses as discussed in the report by Gordon et al. on the delays of subsidized housing title deeds (Gordon, Nell & Di Lollo, 2011). They explain how title deeds are seen as critical to providing security of tenure and allowing poor households to access the wealth building potential of their houses. Naturally the Deeds Registry would want to employ more people to keep up with the pace of registrations, but this is restricted by budgetary constraints. It is therefore important to understand the Deeds Registry and its processes, to determine areas which could be improved through automation.

3.2 Deeds Registry in South Africa

3.2.1 Development of Land Registration in South Africa

Formal land registration within South Africa started several hundred years ago. Geo Denoon (1943) records the first two freehold grants and formal transfers occurred in 1657 and 1658 respectively. The transfers and associated mortgages “were passed before two Commissioners of the Council of Policy and were countersigned by the Secretary” until 1828, in accordance with Ordinance 39 before the Registrar of Deeds. The Land Survey Act 9 of 1927 and the Deeds Registries Act 47 of 1937 (‘the Act’) were legislated to provide more substance to South African property law through mapping each individual unit (Radloff, 1996). Further legislation to guide deeds registration was passed including the Sectional Titles Act, 1986 (Act 95 of 1986) and the more recent Land Survey Act, 1997 (Act No. 8 of 1997). Common law, customary law and circulars are also referenced by regulations under the Act (Shange, 2010).

3.2.2 Current System

There are eleven deeds registries in South Africa, located in Cape Town, Johannesburg, Kimberley, Pietermaritzburg, Vryburg, King Williams Town, Umtata, Bloemfontein, Pretoria, Nelspruit and Polokwane (Department of Rural

Development and Land Reform, 2018). The office locations are viewable in Figure 2. The mandate of the Deeds Registry is to "register title deeds and documents, manage and maintain the country's land register, provide information related to registration and archive the records". The Deeds Registry is primarily governed by the Act which allocates the responsibilities of deed lodgement to the conveyancer and that of ensuring security of title to the Registrar of Deeds (Department of Rural Development and Land Reform, 2016).

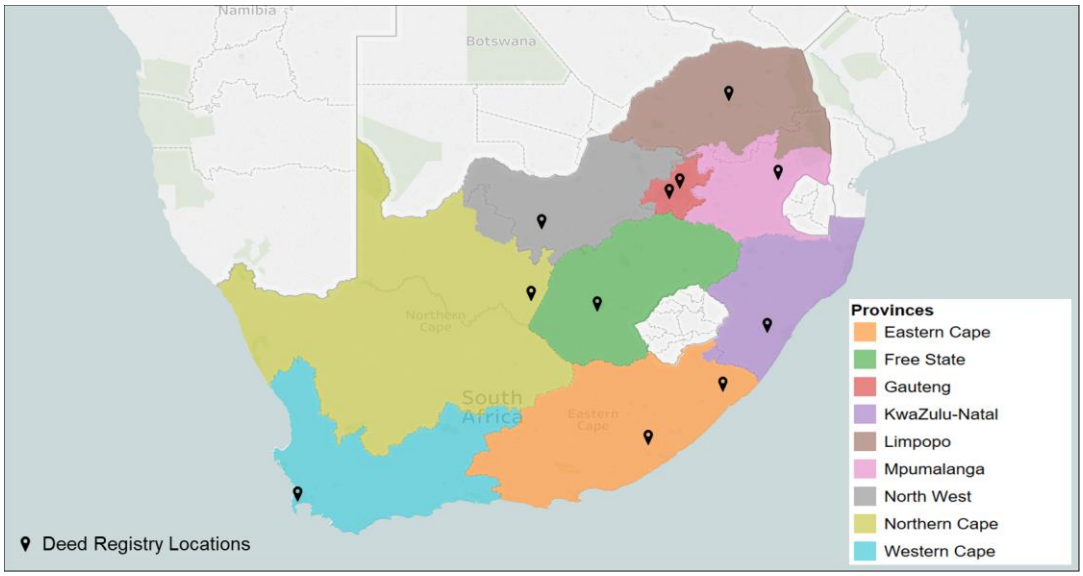


Figure 2: Deeds Registry Locations in South Africa

The Deeds Registry is funded by charging for services based on the Schedule of Fees prescribed by Regulation 84 of the Act (Department of Rural Development and Land Reform, 2018). The Act stipulates in section 7 that any person can inspect the public registers or documents filed at the registry "upon payment of the prescribed fees". This fee applies to every person or entity including the State (South Africa, 1937). The funds received by the registry are dependent on the demand for services provided by the Deeds Registry. With no provision for expansion or option to receive funding from the national government, it can be inferred that the Deeds Registry is unable to employ more people to solve the backlog, without an increase in the prescribed fees. A higher rate of technological integration may therefore be a sustainable solution to increase employee productivity, without needing to increase fees.

Amadi-Echendu (2017) explains, in her research of the South African conveyancing system, how all original signed documents are couriered to conveyancing attorneys for them to manually lodge at the Deeds Registry. This is owing to paper documents being required by Section 10 of the Act as well as a conveyancing system that is only "equipped to accept paper documents" based

on Regulation 20 of the Act. A conveyancer collates all the necessary documents and manually lodges them at the Deeds Registry for them to record the transfer of property into the buyer's name. The documents lodged include the deed of transfer (i.e. the title deed), sales agreement, rates clearance from the municipality, transfer duty receipt/exemption and pest control assessment (Shange, 2010).

3.2.3 e-DRS system

The Deeds Registry acknowledges the major backlog it faces and how technology could help ease the burden. This is shown in their aim to develop an electronic deeds registry system (e-DRS) in order to improve turnaround times, registration accuracy and increase the volume of deed registration (Minister of Rural Development and Land Reform, 2017). The policy document on the e-DRS system was endorsed in 2009, and the Bill explaining certain aspects to be included was circulated in Parliament in 2016 (Amadi-Echendu, 2017) and again in 2018, and has now been approved by the National Assembly. Final approval is still required from the National Council of Provinces and President before implementation (Parliamentary Monitoring Group, 2018). The specifics of how the system will be developed has not been announced, but the Bill mentions that once the system is implemented, it will run concurrently with both manual and online lodgement during the transition (Minister of Rural Development and Land Reform, 2017). There are also aspects from the Electronic Communications and Transactions Act (ECTA) which may be incorporated, such as advanced electronic signatures (Heyink Mark, 2018). But even with the Deeds Registry aiming to develop an electronic system, there may be ways to increase efficiency in the interim.

3.2.4 Process at the Deeds Registry

Brian Daniels, Assistant Registrar, recently provided an outline of the Western Cape's deeds registration process to Parliament (2018). After lodgement by a conveyancer, the entire process of deeds registration is completed in a timeframe of 12 days, which ensures the legality and accuracy of the property transfer. The Registrar's signature on the documents confirm that multiple checks and balances have been performed to ensure high accuracy levels. However, with the high volume of between 800 to 1300 deeds lodged per day, this 12 day target is often not reached. The ideal process as described by Daniels and Shange (2010) is displayed in Figure 3 and explained further below.

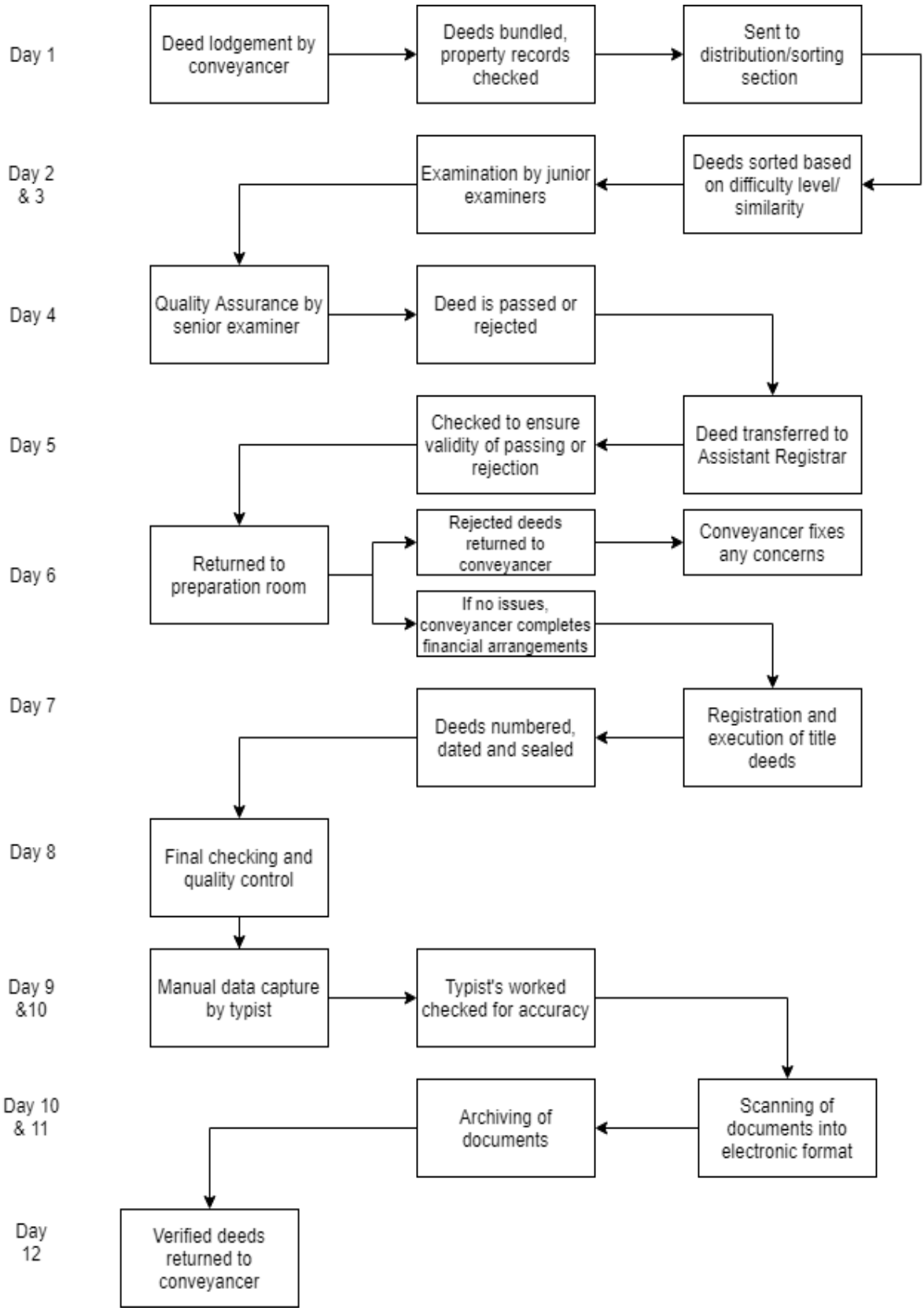


Figure 3: Ideal 12 Day Process of Registering a Deed at the Deeds Registry

3.2.4.1 Process Description

The first seven days of deeds registration are focused on the examination process, which is the core function of the Deeds Registry (Parliament South Africa:, 2018). The process begins with a conveyancer lodging the relevant documents at the deeds office. Shange (2010) explains in depth the examination process performed by each examiner. Before examination occurs, deeds are batched and sorted according to the batch complexity and deed similarity. A junior examiner will check for any errors on the deed or if there are any interdicts on the relevant persons and/or properties. A senior examiner performs additional checks including reviewing if the related legislation has been complied with. An Assistant Registrar or monitor will assess any notes by the previous examiners and "oversee the fairness of rejections". Any rejected deeds are returned to the conveyancer for rectification before re-lodgement. Deeds which have passed all the examinations and have the final financial arrangements in order, are ready for execution and registration. The deed is brought before the Registrar to be signed and sealed. This is the point where the transfer of property ownership occurs (World Bank, 2018b).

After property registration, the deeds go through a numbering process where each deed receives a unique number and is stamped with the registration date (Shange, 2010). A final quality check is performed to ensure no interdicts were received which prohibit the transaction. If there are no issues, a typist captures the data into the Deeds Registry's database with every typed entry being subsequently checked for accuracy. This facilitates the Deeds Registry building up a historical record of properties (Parliament South Africa:, 2018). After data capture, there is an archiving and scanning process where each deed is scanned into an electronic format. The electronic version is kept as a copy by the Deeds Registry and the verified paper copy is returned to the conveyancer.

3.2.4.2 Evaluation of Processes

The Deeds Registry, despite being manual and paper driven is regarded as one of the most secure systems worldwide (Amadi-Echendu, 2017). However, the overall rankings by the World Bank, in relation to the ease of registering property, only place South Africa in position 106 of 190 countries (World Bank, 2018a), as seen in Table 1 below. It is interesting to note that Rwanda, a developing country within Africa, is ranked second, after they implemented a series of land tenure reforms in 2010 (Ali, Deininger & Duponchel, 2016). Being ranked in the bottom half of countries points to the potential existence of

inefficiencies within the Deeds Registry[3], which automation can reduce. This is consistent with Amadi-Echendu (2017) who states that "various inefficiencies exist in the end-to-end process" which is exacerbated by the process being paper based. She further notes that owing to value placed on security, there is a seeming reluctance in South Africa to update the current property registration system to an electronic base, even though the procedures which have been computerized reveal increased efficiencies.

Table 1: Ease of Registering Property Comparison

Economy	Registering Property Rank	Registering Property Score	Procedures (number)	Time (days)	Cost (% of property value)	Land Admin Quality (0-30)
New Zealand	1	94.89	2	1	0.1	26.5
Rwanda	2	93.7	3	7	0.1	28.5
Belarus	5	92.19	2	3	0.0	23.5
Morocco	68	67.86	6	20.5	6.4	19.5
Botswana	80	65.43	4	27	5.1	10
South Africa	106	59.32	7	23	7.8	15
Lesotho	108	58.25	4	43	8.0	9.5
Zimbabwe	109	58.2	5	36	7.6	10

Source: Adapted from Doing Business Report 2019 by World Bank

The first seven days facilitate the core examination process of the Deeds Registry, ensuring legal validity and accuracy of every property transfer (Parliament South Africa, 2018). The review process appears extensive, with examinations done by multiple examiners and an Assistant Registrar as well as additional quality control procedures as described above. There appears to be a potential area of efficiency gain during days 9 to 11, which incorporates data capture, scanning and archiving. The data is manually captured before the documents are scanned and archived. Potential efficiency would be gained by incorporating OCR to convert the scanned title deed and other documents into machine readable text which would be automatically captured into the Deeds Registry database. This would require a reversal of procedures by scanning the documents before data capture, in order to have a digitised version of the documents on which to perform OCR. Since OCR may produce results with inaccuracies, there would still likely be requirements to manually check the accuracy of the OCR outputs, just as is currently done following the typist's manual data entry.

[3] The methodology used to rank countries records the full sequence of procedures necessary to register a property and not just the Deeds Registry process. Further information on the methodology for ranking property registration is available at http://www.doingbusiness.org/en/methodology/registering-property

As mentioned above, there exists a seeming reluctance to incorporate more efficient technology into the South African property registration system owing to concerns of how secure an electronic system will be in comparison to the status quo. Therefore, incorporating additional security measures within a technological framework may reduce these concerns. For example, the OCR data capture pipeline can include the storage of scanned deeds on a secure distributed file system such as IPFS.

After discussing the title deed backlog and potential inefficiencies within the Deeds Registry process, OCR and IPFS were suggested as ways to increase data capture efficiency without compromising on security. The OCR pipeline, with IPFS incorporated, is described in greater depth in the following chapter.

Chapter 4

4. Methodology

The previous chapter proposes that there is a potential efficiency gain by automating the data capture process at the Deeds Registry by applying OCR to the scanned title deeds and thereafter extracting the data. Additional security measures using IPFS to store scanned title deeds is also suggested. This chapter discusses the title deeds used in this pilot project as well as the techniques applied to automate data extraction from the collected and simulated title deeds. The deeds are scanned into the system as PDFs. In order to safely secure the deeds, the PDFs are encrypted and then uploaded onto IPFS with the hash[4] of the PDF incorporated into the CSV. The original PDFs are converted into images where pre-processing is performed using OpenCV. OCR using Tesseract engine is then performed to convert the images into machine readable code. The text output is then analysed using Regex to determine the text of interest. The relevant text is then extracted into a CSV and compared to the correct input to determine accuracy. This application is created in Python using multiple packages, with the notable packages being CV2, pytesseract and re. The full process is summarized in

Figure 4 with a further explanation of the key methods discussed later in the chapter. The average time per manual data input is then compared to the time for data extraction via OCR.

[4] This is the result of cryptographic hash function which takes an input of any size and returns an output of a fixed size where it is nearly impossible to calculate the input from the output. This results in a high level of security.

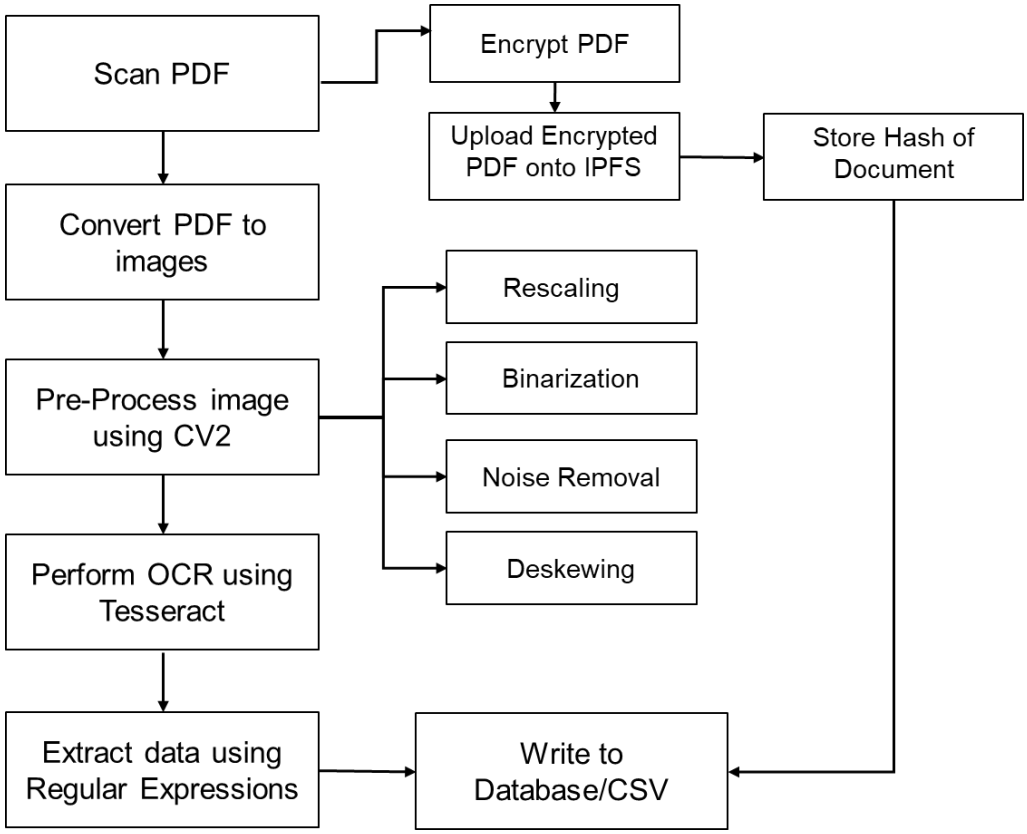


Figure 4: OCR Data Extraction Process

4.1 Data Description

This research undertakes a proof of concept to test the accuracy and viability of using OCR for data extraction from title deeds. The Deeds Registry provides publicly available property ownership information through the acquisition of title deed copies in terms of Section 7 of the Act. A title deed is a legal document outlining the buyer, seller and property details as well as any restrictions, conditions or special entitlements. Title deeds generally follow a similar format and linguistic flow. However, because of different conditions or ownership structures the layout and number of pages varies. For each transfer of property, a new title deed is created.

The ownership structure and type of property effect the way that the data is stored within the deeds database. For example, a couple married in community of property will have two identical entries containing the transfer details, with the only differences being the names and identity numbers of each partner. This will also be the case for each party in a fractionally owned/bought property. A sectional title property which has a parking or separate storeroom, will also have duplicate entries which differ in property sizes. The structure of the fields to be

extracted from each title deed, with example entries, are shown in Table 2 below. The first entry represents a standard sale transaction. The second and third rows entry show a couple married in community of property buying a freehold title and the last two rows represent the sale of a sectional title property.

Table 2: Example Extract of Title Deed Data Extraction

Name Buyer	Surname Buyer	Buyer ID	Name Seller	Surname Seller	Seller ID	Date Transact	Date Register	ERF Number	Title Deed Number	Sect Plan No.	Sect No.	Size	Transaction Price	Bonded Amount	Share
Harry	Potter	8007310048036	Hogwarts		2018/345678/01	01-01-18	04-01-18	123	T1234/2018			300	R150,000.00	R0.00	100%
Prop LTD		2015/123456/01	Mark	Anthon	8301140230808	04-07-06	07-07-06	321	T5678/2006			1500	R600,000.00	R80,000.00	100%
Prop LTD		2015/123456/02	Cleo	Patra	6901010028012	04-07-06	07-07-06	321	T5678/2006			1500	R600,000.00	R80,000.00	100%
John	Smith	7504120548012	Peter	Parker	8507080028082	06-08-11	09-08-11		ST12345/2011	SS 1234/2005	301	400	R400,000.00	R50,000.00	100%
John	Smith	7504120548012	Peter	Parker	8507080028082	06-08-11	09-08-11		ST12345/2011	SS 1234/2005	87	10	R400,000.00	R50,000.00	100%

Row 1: Standard sale
Row 2 & 3: Buyers married in community of property
Row 4 & 5: Sectional title with sale of apartment and parking

The title deeds used in this proof of concept include representation of fractional ownership, marriages in community of property, sectional titles, government subsidized housing and several combinations of the above. A sample of 6 title deeds were collected from the City of Cape Town and other sources. These title deeds are from 2000 onward in order to better approximate the scanning quality and format of current deeds. A template of a title deed was acquired and used to simulate an additional 15 title deeds. The composition of the 21 title deeds analysed is in Table 3.

Table 3: Composition of Deeds Analysed

Types of Deeds	**Number of Deeds Analysed**
Standard Freehold Title	7
Sectional Title	5
Buyers Married in Community of Property	3
Subsidized RDP housing	3
Fractional Ownership	4
Total	21*

**Some deeds contained multiple properties such as being an RDP house being provided for a couple married in community of property.*

4.2 Image Pre-Processing

Open Source Computer Vision Library (OpenCV) is an open source computer vision library with machine learning capabilities (OpenCV Team, 2018). This package has all the necessary tools for image pre-processing. For improved OCR results, the Tesseract wiki recommends that images be pre-processed through rescaling, binarization, noise removal and deskewing (Tesseract, 2018).

4.2.1 Rescaling

Tesseract has better performance on images with a Dots Per Inch (DPI) of at least 300. For images with a lower DPI, enlarging the image is necessary. There are various interpolation methods for rescaling images such as pixel resize, bilinear and bicubic interpolation. Bicubic interpolation is robust and preserves fine details (Nuno-Maganda & Arias-Estrada, 2005) and was therefore chosen for this application where details are important for character recognition.

4.2.2 Binarization

Binarization converts an image to black and white based on certain threshold parameters. A simple binarization threshold was used which converts pixels with a threshold value greater than 127 to black, otherwise to white. Figure 5 shows an example of this simple binary thresholding.

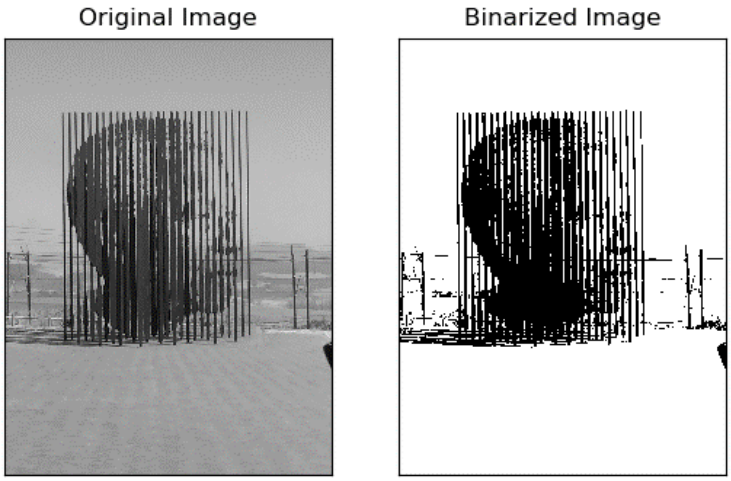


Figure 5: Simple Binarization

4.2.3 Noise Removal

Scanned images often have ‘noise’, which appears as randomly distributed dots and negatively affects OCR accuracy (Tesseract, 2018). Noise reduction is often performed using mathematical morphology which is a combination of erosion and dilation (Chinnasarn, Rangsanseri & Thitimajshima, 1998). Erosion shrinks the greyscale value of an image and dilation has the opposite effect (Yu-qian et al., 2006). The combination of erosion and dilation, called opening and closing, improves the performance of mathematical morphology in salt and pepper noise reduction (Chinnasarn, Rangsanseri & Thitimajshima, 1998). A sample of 100 images with random noise were generated. After noise removal via mathematical morphology, OCR was performed on the images. The accuracy of the recognition improved from a mean of 50.4% to 93.3% after noise reduction had been performed. Figure 6 shows an example of the image transformation.

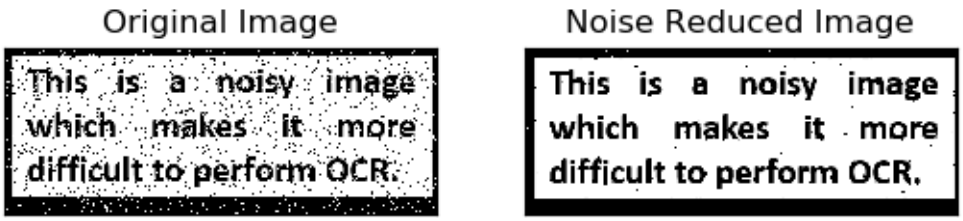


Figure 6: Noise Reduction

4.2.4 Deskewing

With a skew image, Tesseract's line segmentation is impacted significantly, reducing the OCR accuracy (Tesseract, 2018). As the extent of skewness is unknown, the algorithm detects the co-ordinates containing the block of text and finds the minimum angle needed for deskewing. The deskewing of a sample deed extract is shown in Figure 7 below.

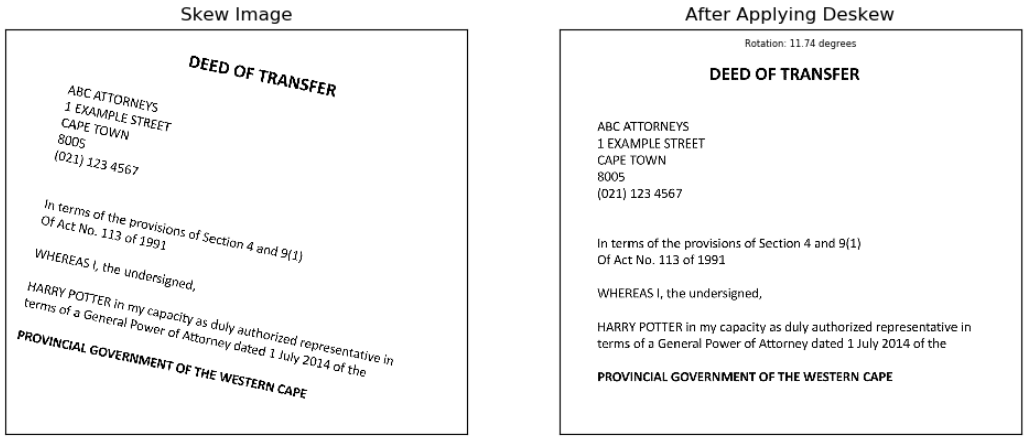


Figure 7: Deskewing Images

4.3 Tesseract

Tesseract is a widely used open-source OCR programme which started its development at HP between 1984 and 1994 (Smith, 2007) and is now developed and maintained by Google. The process, as explained by Smith, follows a traditional pipeline. Component analysis is performed to determine outlines which are nested into "blobs". These blobs are arranged into text lines which are divided into words based on spacing (Asif et al., 2014). The words are then divided into characters by chopping the blob. The characters are classified by a static classifier, which can recognise complete and broken characters. The characters are then regrouped to form words (Ohlsson, 2016) which are classified using an adaptive classifier allowing for learning and improved accuracy (Asif et al., 2014).

Tesseract 4.00 uses a Long Short-Term Memory (LSTM) Neural Network for recognition which has shown significant accuracy improvements on document images. The existing model has been extensively trained for Latin-based languages, including English, in several thousand fonts (Benda Krisztián, 2018). LSTMs have become more popular than Hidden Markov Models (HMM) as the

performance improves through learning context over time as opposed to only characters (Sabir, Rawls & Natarajan, 2017).

The output from Tesseract provides the text as well as the position details and accuracy confidence for each word. Regex was used to process the string output from Tesseract. The position details were not used in this context as the text position for title deeds is not standardized, with deeds ranging in length from three to thirty pages.

4.4 Regex

Regex is a pattern notation allowing for the parsing text in a powerful, flexible and efficient manner and can be used as a full text processing language (Friedl, 2006). Text extraction tasks can be addressed using Regex in a plethora of practical cases by recognising the underlying syntactical pattern (Bartoli et al., 2016). A case study using Regex to extract clinical data from medical documents showed high accuracy and speed using a Regex application (Turchin et al., 2006). Since Regex works well with semi-structured or unstructured text (Bartoli et al., 2016), it is appropriate for title deed data extraction which can be considered semi-structured. For example, a sectional title number has a standard format starting with “ST” and ending with the year of transfer which can be represented by the regular expression below.

$$r_a = ST\backslash s\backslash d + /\backslash d\{4\}$$

OCR output can sometimes contain out of place symbols and characters. This would occur if there is any noise remaining after pre-processing, or there are lines, stamp borders or images which are misclassified by Tesseract as letters or symbols. Regex was therefore used not only for text extraction, but also for text cleansing through removal of characters not expected in a title deed. The following characters were removed using Regex.

$$r_b = \text{\textbf{\textregistered}}*|;\%!\$\#@\sim<\geq=`$$

The remaining noise left after image pre-processing is often classified by Tesseract as repetitive characters such as “eee” or “oo”. Regex was therefore used for additional cleansing to remove any standalone character groupings which contained unexpected repetitive characters which could hinder data extraction. The combination of “SS” and “CC” is excluded as they form part of the format for a sectional plan number and closed corporation respectively.

$$r_c = \backslash b([a-b, d-r, t-z])\backslash 1\{1,\}\backslash b$$

4.5 Timing of Manual Capture

To assess the benefit of using the OCR to reduce time of data capture, it is necessary to compare the time of the OCR process against that of manual data capture. The time of manual data capture was calculated by averaging personal times for each of the sample deeds. As a professional typist is likely faster in data capture, the personal times were reduced by 25% to be closer aligned to reality. The average of all the title deed times was then computed. The time taken for OCR is based on my personal computer's specifications[5] and therefore, if the program was run on other computers, processing times would differ. There would also be room for time reduction if the programme was run on a server, parallel computing was incorporated or a computer with higher random-access memory (RAM) was used.

4.6 IPFS

IPFS is a peer-to-peer distributed file storage and sharing system where each file and its associated blocks are linked with a unique cryptographic hash. This unique hash ensures that any file uploaded to IPFS is immutable as any change to the file is tracked and detectable. The Merkle Directed Acyclic Graph (DAG) structure provides useful properties which ensure content uploaded to IPFS is uniquely identified through a checksum, is tamper resistant and will not be duplicated. Because of the unique hash, content addressing is performed by using the hash to access content. The distributed nature of IPFS has no single point of failure (Benet, 2014).

Since content uploaded is retrievable by anybody who knows the hash of a file, it is necessary to encrypt the files to ensure privacy. This is also necessary as the Deeds Registry finances itself through sale of deeds data. Cloud computing has similar issues of data security and privacy. A recent application uses the Advanced Encryption Standard (AES) algorithm for file encryption to enhance data security (Rewagad & Pawar, 2013). AES was therefore chosen as the encryption algorithm within this application. The encryption process is displayed in Figure 8. The file is divided into smaller chunks which are then encrypted with a password [6] and an initialization vector[7] (IV) which prevents unauthorized use. The file is then uploaded to IPFS and will only be decryptable with the password used to encrypt the file.

---

[5] Processor: Intel® Core™ i7-4510U CPU @ 2.00GHz, 2601 Mhz, 2 Cores, 4 Logical Processors, RAM 8GB

[6] Password is hashed using the SHA256 algorithm

[7] The IV is a random number which is used to initiate the encryption process and make it more difficult for dictionary hackers
https://www.techopedia.com/definition/26858/initialization-vector

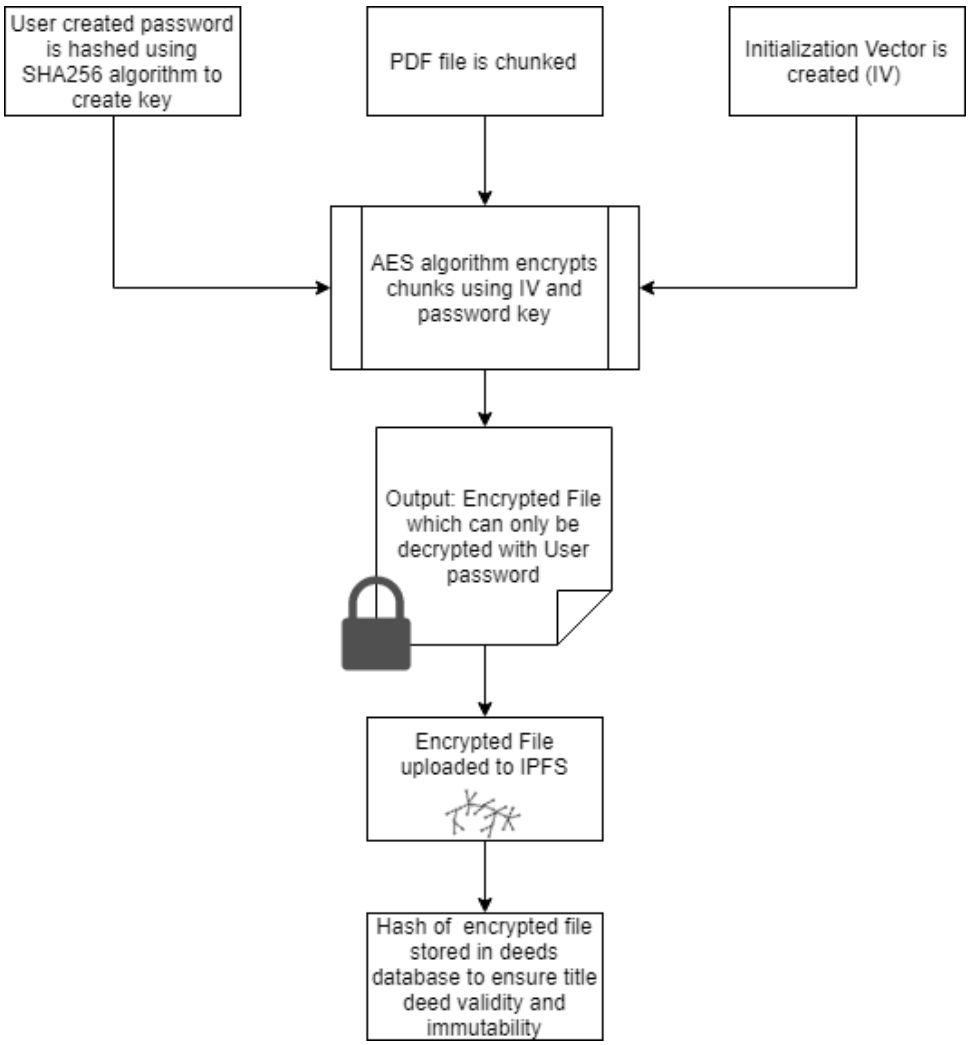


Figure 8: Process of AES encryption and upload to IPFS

By linking each scanned title deed with a unique hash, the security of electronic title deeds is enhanced as any slight alteration to the title deed will result in a completely different hash, thereby being detectable. The unique hash generated by the IPFS system is then integrated with the deeds database by storing the hash with the corresponding title deed. This is to add additional security to electronic scanned deeds by making it is possible to check the validity of the data in the database by retrieving the deed with the hash stored in the IPFS filesystem, decrypting the deed and comparing the document to the database entry. Any deeds tampered with will be detectable and the original deed will be recoverable on the IPFS network.

Chapter 5

5. Results

5.1 General

The previous chapters discuss the importance of a deeds registry and having formal title over property, as well as the current backlog issues plaguing the South African Deeds Registry. After a discussion of the workflow of the Deeds Registry, it was suggested that implementing an OCR system for automatic data capture may introduce efficiency gains over manual typing of data entries. This chapter delves into the results and insights gained through the application of the methodology explained in the previous chapter. This chapter has further analysis of the limitations of the pilot project, describing areas which have lower levels of OCR accuracy and listing various recommendations based on the results. Several areas which require further research are then discussed briefly.

The percentages measuring accuracy are calculated based on the number of row comparisons. There are additional row entries for fractional ownership, couples married in community of property and sectional title properties. Therefore, there are more rows than title deeds to record the multiple owners or sections of a property. The number of deeds is included in Table 4 for additional reference. For entries which are only relevant to certain deeds, there are fewer number of comparisons. This would be the case for companies or government agencies without a surname, entries specific to sectional title properties, fractional ownership as well as unmortgaged property.

There are no data capture accuracy measures for IPFS as the hash stored in the database is not extracted from the deed using OCR. The encrypted files were however successfully added to IPFS and can be retrieved using their hash. The title deed file retrieved using IPFS was encrypted and can only be opened through decryption with the user password. The additional time to incorporate the encryption and uploading of the deed to IPFS is around 1 second which is 0.5% of average processing time.

5.1.1 Findings

The CSV row output from the OCR pipeline when compared to the expected output based on manual data capture showed an overall accuracy of 89.6%. Of the 16 columns which were extracted, 13 had accuracy levels exceeding 95%, with 10 fields having no errors in capture from the deeds analysed. The accuracy levels ranged from 0% to 55.3% for the remaining fields representing the title deed number, date of registration and the mortgage amount of a bonded property. These columns will be the discussed further in the following section. The summary of the results is displayed in Table 4.

Table 4: Results of Accuracy levels for Data Capture

Extracted Columns	Accuracy*				No. of Comparisons**	Number of Deeds
	Captured Correctly	Captured Misspellings	Incorrect Capture	No Capture		
Buyer First Name	100%	0%	0%	0%	47	21
Buyer Surname	100%	0%	0%	0%	45	19
Buyer ID	100%	0%	0%	0%	47	21
Seller First Name	100%	0%	0%	0%	47	21
Seller Surname	100%	0%	0%	0%	18	8
Seller ID	100%	0%	0%	0%	43	18
Erfno	100%	0%	0%	0%	25	13
Title number	55.3%	4%	40.4%	0%	47	21
Size of Property	100%	0%	0%	0%	47	21
Date Transaction	95.7%	4.3%	0%	0%	47	21
Date Registration	8.7%	21.7%	0%	69.5%	46	20
Section Plan Number	100%	0%	0%	0%	22	5
Section Number	95.5%	4.5%	0%	0%	22	5
Price	97.9%	2.1%	0%	0%	47	21
Share	100%	0.00%	0%	0%	22	4
Bond Amount	0%	0%	0%	100%	22	9
Average	89.6%	1.0%	2.7%	6.7%	36.5	15.2

**Accuracy percentages are based on number of comparisons*

*** The number of comparisons is larger than the number of deeds as there are multiple row entries for fractional ownership, couples married in community of property as well as sectional title properties*

The time taken to personally perform manual data capture for the 21 deeds averaged 4:33 minutes. When reduced by 25% to incorporate the typing efficiency of a professional typist, this reduces to a time to 3:24 minutes. This is 21 seconds longer on average than it takes for the OCR pipeline. Based on an eight-hour work day, an extra 16 deeds can be processed per day. When adjusted for computers operating for 24 hours, this number increases to 331 additional deeds that can be captured daily. These results are summarised in Table 5: Comparison between Manually Capture and OCR Application. As mentioned above, the Western Cape Deeds Registry receives between 800 and

1300 deeds a day. Therefore, to capture the data for all those deeds within 1 day, between 6 and 9 employees are required.

Table 5: Comparison between Manually Capture and OCR Application

Measures	Manual Adjusted*	OCR Application	Difference
Average (mm:ss)	3:24	3:03	0:21
Deeds Captured per Day (8 hours/8 hours)	141	157	16
Deeds Captured per Day (8 hours/24 hours)	141	472	331

** The average time taken for data capture was 4:33. This was reduced by 25% to account for the additional efficiencies for a professional typist*

5.1.2 Areas of Difficulty for OCR

5.1.2.1 Title Deed Numbers

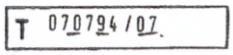
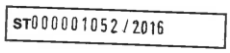
The title deed numbers found on the deeds were either typed or stamped, with significantly different results respectively. The breakdown of results based on these properties is shown in Table 6.

Table 6: Breakdown of Results for Title Deed Number

Extracted Columns	Accuracy				No. of Comparisons	Number of Deeds
	Captured Correctly	Captured Misspellings	Incorrect Capture	No Capture		
Typed	86.7%	6.7%	6.6%	0%	30	16
Stamped	0.00%	0%	100%	0%	17	5
Total	55.3%	4%	40.4%	0%	47	21

The OCR pipeline was unable to extract any of the correct stamped title deed numbers, indicating the difficulty in extracting stamped data. As the title deeds analysed refer to conditions and/or restrictions from previous deeds of the property, typed title deed numbers of the previous deeds were captured instead of the correct current number. Table 7 shows several of the stamped title deed numbers. Although most of the title deed number is deciphered using the OCR pipeline, the underlining, drawn in letters and general lower quality of the stamps, results in unexpected classifications of the characters. As the output is not in the general format of a title deed number, the pattern using Regex is not found, resulting in poor levels of data extraction.

Table 7: Output Comparison for Title Deed Stamps

Manual	OCR	Stamp Screenshot
T 070794/07	TS D7079e 707	
ST000001052/2016	STO0G001052/ 2016	
ST00012002/2013	34 2690007762/ 2013	

The accuracy for typed numbers was much higher, with a rate of 86.7%, although there were some misspellings and capturing of the wrong title deed number that occurred. This indicates that errors with OCR can crop up, even with typed characters.

5.1.2.2 Date of Registration

The registration date has a significantly low rate of capture of 29.4% with the majority of captured line items containing errors in spelling. The registration date is stamped at the end of the title deed and is often skew, smudged or unclear. The sample of deeds also includes a handwritten date of registration. Screenshots of several of the stamps are in Figure 9: Date of Registration Snapshots. Of the examples below, only “29 JUN 2017” and “25 FEB 2012” were captured but still contained errors. The others are less clear resulting in non-capture.

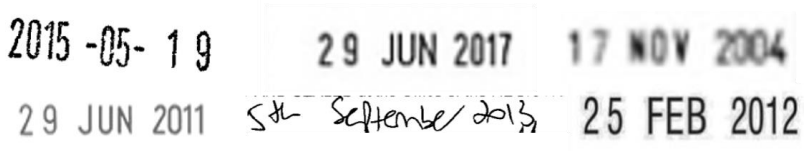


Figure 9: Date of Registration Snapshots

5.1.2.3 Mortgage Stamps

Retrieving the mortgage bond amounts has multiple challenges resulting in the lowest accuracy of data fields extracted. Many of the challenges occur owing to the bond amount being handwritten within stamps which are regularly smudged, faded or unclear. Various other obstructions such as signatures also

impact OCR recognition. Another factor was that the mortgage stamps were often skew relative to the printed text which affects the deskewing process. Screenshots of various mortgage stamps from the data set are shown in Figure 10.

Figure 10: Various Mortgage Stamps found in Title Deeds

5.1.2.4 Overall

The lower performance of stamps has been shown in the results of other research. Dey et al. (2016) references how the presence of stamps causes OCR systems to deteriorate in accuracy, especially for any stamps which overlap on document text. Mihov et al. (2005) also finds that handwriting and stamps bear recognition errors using OCR. Therefore, in the current format of title deeds, the automatic capture for stamped fields is likely to have low accuracy. However, with the recommendations below implemented, accuracy levels are likely to improve.

5.1.3 Recommendations

5.1.3.1 Fields with low accuracy

The fields with the lowest accuracy relate to stamped or handwritten information. For all information that is stamped on, a standardized format should be used as this allows for better data capture using Regex. Additionally, the stamp must be stamped carefully, firmly and parallel to the rest of the text. This will reduce the occurrence of fading, smudging and skewness which reduces the accuracy. It may be the case that the stamps require a different variation of image pre-processing, in comparison to the rest of the document. If so, it would be possible to include another component into the application which identifies that there is a stamp and applies a more suitable image pre-processing. OCR can then be applied to that portion of the deed to try get better extraction.

Handwriting should be replaced with either typing or stamps which tend to be more consistent.

For extracting the bond amount, the signature overlapping the other words further reduces accuracy. Therefore, the signature should either be in a separate box or below the stamp in order to prevent partial cover. Another option would be to have a digital stamp pre-printed on the deed in order to ensure clarity of text and horizontal placement.

As the title deed number is a key component on each data entry, ensuring accuracy in this field is non-negotiable. As PDF documents of the title deed are likely to be named after the title deed number, the file name could be parsed to extract the correct number. This will avoid errors in relation to OCR without increasing data capture time.

5.1.3.2 Other Recommendations

Regex can be used to ensure that entries are in the required format. As many of the fields captured from the title deeds are in a standardized format, Regex can be incorporated to flag any unexpected output such as dates falling outside of the calendar or an ID number with an incorrect number of digits. This can alert the employee checking for accuracy. This recommendation applies to both manual and automatic data capture. Another addition to improve accuracy would be the inclusion of Fuzzy Matching. This could be incorporated, for example, by matching suburb names to a list of the suburbs within South Africa to get the closest match. This would ensure small errors would be picked up. Other text analysis programmes such as Natural Language Toolkit (NLTK) could be combined for deeper text analysis.

Although the writing style of conveyancers is generally very similar, there are some phrases which are worded differently. This makes using Regex more difficult. If this application was to be incorporated within the Deeds Registry, it would be recommended that the Registrar issue guidelines to conveyancers about writing style and formats for the different types of deeds, to encourage a higher level of consistency.

As SADiLaR has already invested in OCR research and technology within South Africa, the Deeds Registry should investigate a potential partnership with them. SADiLaR may have access to off-the-shelf OCR products which are likely to have higher OCR accuracy for data extraction. The accuracy levels for these products should therefore be tested and assessed for appropriateness in relation to title deeds.

5.1.4 Limitations

This pilot project was performed on a sample size of 21 deeds with 71% of deeds being simulated. A larger sample would be necessary to ensure widespread application across the multiple deed offices. All the deeds tested were written in English. However, title deeds may be written in other languages such as Afrikaans. Even though a variety of title deed formats were used in the pilot project, farm ownership, usufructs and liens were not tested. There were no transactions involving foreign buyers and sellers where the passport number would serve as the ID. Furthermore, only deeds registered after 2000 were used. This was specified due to newer deeds better representing the printing and scanning quality which the Deeds Registry deals with during data capture.

5.1.5 Discussion

The general accuracy of the automated data capture is at a relatively high standard of 89.6% where the correct data is extracted with no misspellings. When the results are adjusted to exclude any non-typed data, the accuracy improves to 98.3%. From the manual capture process completed personally, 9 misspellings were discovered resulting in an accuracy of 98.5%. For most fields that are to be extracted, the application performs in a similar accuracy to manual data capture. Therefore, just as the Deeds Registry process outlined in Section 3.2.4 requires the typist's work to be checked for accuracy so to the OCR extracted fields would require this check.

The other three fields which have low accuracy and capture rates need to be specifically addressed before the entire pipeline is integrated into the system. This can be done by implementing the recommendations discussed above and assessing the improvement in accuracy. If the improvement is not substantial, the application can be used to extract the high accuracy fields and manual data capture will be required for the remaining fields.

The average time taken for the OCR pipeline is shorter than that of the adjusted manual capture time. Considering that computers can operate 24 hours a day, the number of deeds that can be automatically captured is above 300% more than those captured by a single employee. With further optimization of the code, incorporating parallel computing and increasing computing power, the time to run the OCR pipeline would be substantially reduced. This would allow for additional deeds to be processed daily at a reduced cost.

Including the IPFS and encryption has little impact on the processing time. The additional security features of immutability and tamper detection seem to be worth the additional processing time.

Therefore, this pilot project indicates that using OCR for automatic data extraction is possible for the South African Deeds Registry. If there will still be a significant number of years until the e-DRS system is fully operational, the pilot project should be expanded to test the application on a larger sample of title deeds as the manual lodgement process continues. The uploading of encrypted files to IPFS can be implemented for both past and current deeds, without requiring OCR. Once the e-DRS system is developed and deeds are lodged electronically, IPFS can still be incorporated as a security measure.

5.2 Areas of Further Research

5.2.1 Increase Complexity and Volume of Title Deed Samples

The sample of title deeds on which OCR was applied was limited in size and scope. In order to ensure the versatility and applicability of the tested application, the volume and complexity of title deeds will need to be expanded. The additional deeds tested should be from the Deeds Registry archives as this provides a more realistic sample of deeds handled. The increase in complexity will include various other forms of property within South Africa such as farms and usufruct rights. The combination of various ownerships structures such as a couple married in community of property taking part in a fractional ownership structure. As the title deeds sampled only represent deeds from 2000 onward, older deeds can go through the OCR pipeline to assess how this application performs in archiving.

5.2.2 Other Applicable industries and government applications

As discussed in Chapter 3, government departments within South Africa are often overloaded with large volumes of forms and documents needing to manually be captured into a database. Owing to unequal access to ICT infrastructure, particularly in rural areas, paper forms are likely to remain a reality that faces the government. Further research needs to be done applying this technique to other government forms which are received on a constant basis. This could include forms such as ID applications and tax filing forms. Although the OCR does not have the same level of accuracy for handwriting, it may be the case that structured forms which have boxes for people to write characters, may create more uniformity of written text where a specific OCR classifier could be

trained. There is also room to research if different designs for forms could improve the OCR results.

5.2.3 Security within Technology

Owing to the concern of increased security risks associated with technology, there is room to research additional ways to incorporate security within a technological framework. IPFS was discussed above as a potential addition to incorporate additional validity and immutability into title deed records. Another option is a distributed and immutable database such as blockchain. Blockchain is increasingly being used as the basis for addressing land and property matters including deed registries. This technology is being proposed as a solution to reduce time-consuming and expensive intermediary functions, proving ownership of property while including the security and resilience owing to the decentralization, fault tolerance and immutability ingrained within blockchain (Graglia & Mellon, 2018). Many countries have investigated how blockchain can be used as a basis for their land registration systems[8]. Within South Africa, more research is required to assess if blockchain can be implemented effectively, which is the most appropriate underlying structure and how to incorporate this technology within the current legislative landscape. IPFS can also be integrated with blockchain as a transport protocol (Benet, 2014). This integration may further enhance security by drawing on the stronger features of both technologies.

Within the current IPFS inclusion, various other encryption methods can be tested. If the Deeds Registry will want to send title deeds using the IPFS system, asymmetric encryption may be more suitable. This means that the document will be encrypted with the public key of the recipient. Only the recipient will be able to decrypt the file as their private key is needed[9]. More security layers within encryption could be included such as digital signatures. Therefore, if the Deeds Registry incorporates IPFS, depending on their usage, research should be performed to determine the most appropriate and secure encryption method.

[8] UK: https://www.gov.uk/government/news/hm-land-registry-to-explore-the-benefits-of-blockchain
Bermuda: http://www.royalgazette.com/business/article/20180628/bermudas-land-registry-to-go-on-blockchain
Kenya: https://www.bbc.com/news/world-africa-43640885

[9] An article about sharing files via IPFS using asymmetric encryption is found at https://medium.com/@mycoralhealth/learn-to-securely-share-files-on-the-blockchain-with-ipfs-219ee47df54c

Chapter 6

6. Conclusion

The large backlog facing the Deeds Registry hinders the ability of unregistered property owners to unlock the “dead capital” in their possession. Registered land parcels are expected to more than double through land reforms in the near future. Increasing the Deeds Registry capacity is therefore necessary. Since the deed lodgement process legislatively requires manual lodgement, various inefficiencies result. The focus of this paper was to test the application of an OCR pipeline to automatically extract data from scanned title deeds while adding an IPFS security layer.

In comparison to manual data capture, accuracy levels for typed OCR text extraction are at similar levels. However, the accuracy level for stamped or handwritten text is substantially lower. Therefore, additional validation checks and other recommendations are necessary to improve accuracy.

The ability to automate the process allows for an extra 331 deeds that can be processed daily. IPFS integrates smoothly into the system and links effectively to the CSV output. The encrypted files are retrievable through IPFS but cannot be decrypted without the relevant password.

Overall, the preliminary results indicate that the OCR pipeline is effective in automatic data capture of South Africa title deeds and can increase the number of deeds captured daily. The pilot project has room for expansion to broaden the scope and address the limitation of the small and less varied sample. It is therefore recommended that the Deeds Registry investigate the potential for applying OCR to their data capturing process.

Bibliography

Ali, D., Deininger, K. & Duponchel, M. 2016. *Sustaining the success of the systematic Land Tenure Registration in Rwanda.* The World Bank.

Amadi-Echendu, A.P. 2017. *Towards a framework for the integration of data and data sources in the automation and dematerialisation of land administration systems.* University of Pretoria.

Anagnostopoulos, C.N.E., Anagnostopoulos, I.E., Loumos, V. & Kayafas, E. 2006. *A license plate-recognition algorithm for intelligent transportation system applications.*

Andreas van Wyk. 2018. *South African deeds journal.*

Arica, N. & Yarman-Vural, F.T. 2001. *An overview of character recognition focused on off-line handwriting.*

Ashwin, T.V. & Sastry, P.S. 2002. *A font and size-independent OCR system for printed Kannada documents using support vector machines.*

Asif, A., Hannan, S.A., Perwej, Y. & Vithalrao, M.A. 2014. *An overview and applications of optical character recognition.*

Bartoli, A., De Lorenzo, A., Medvet, E. & Tarlao, F. 2016. *Inference of regular expressions for text extraction from examples.*

Benda Krisztián. 2018. *TrainingTesseract 4.00.* Available: https://github.com/tesseract-ocr/tesseract/wiki/TrainingTesseract-4.00.

Benet, J. 2014. *IPFS - Content Addressed, Versioned, P2P File System.*

Besley, T. 1995. *Property rights and investment incentives: Theory and evidence from Ghana.*

Beyers, L.J.E. 2016. *Service delivery challenges facing municipalities: a case study of Fetakgomo local municipality in Sekhukhune District Municipality, Limpopo Province.*

Bhatia, E.N. 2014. *Optical character recognition techniques: a review.*

Biondich, P.G., Overhage, J.M., Dexter, P.R., Downs, S.M., Lemmon, L. & McDonald, C.J. 2002. *A modern optical character recognition system in a real world clinical setting: some accuracy and feasibility observations.*

Centre for Affordable Housing Finance in Africa. 2017. *HOUSING FINANCE IN AFRICA A review of some of Africa's housing finance markets.*

Chen, L., Liao, H., Wang, J. & Fan, K. 1999. *Automatic data capture for geographic information systems.*

Chinnasarn, K., Rangsanseri, Y. & Thitimajshima, P. 1998. *Removing salt-and-pepper noise in text/graphics images.*

Cojocaru, S., Colesnicov, A., Malahov, L. & Bumbu, T. 2016. *Optical Character Recognition Applied to Romanian Printed Texts of the 18th–20th Century.*

Cousins, B. & Hornby, D. 2005. *Will Formalising Property Rights Reduce Poverty in South Africa's' Second Economy'? Questioning the Mythologies of Hernando de Soto.*

De Soto, H. 2000. *The mystery of capital: why capitalism triumphs in the West and fails everywhere else.*

Department of Housing. 2004. *Breaking new ground: A comprehensive plan for the development of sustainable human settlements.*

Department of Public Service and Administration. 1997. *White paper on transforming public service delivery:(Batho Pele White Paper).*

Department of Rural Development and Land Reform. 2016. *Deeds Registries Amendment Bill, 2016.* Available: http://pmg-assets.s3-website-eu-west-1.amazonaws.com/160304Deedsregistriesamendmentbill.pdf .

Department of Rural Development and Land Reform. 2018. *Deeds Registration Service Commitment Charter Standards: 2018.* Available: http://www.ruraldevelopment.gov.za/publications/deeds-registration/file/6916 .

Dey, S., Mukhopadhyay, J. & Sural, S. 2016. *Removal of Gray Rubber Stamps.*

Downton, A.C. & Leedham, C.G. 1990. *Preprocessing and presorting of envelope images for automatic sorting using OCR.*

Feder, G. & Feeny, D. 1991. *Land tenure and property rights: Theory and implications for development policy.*

Fourie, D. 2001. *The generation of additional financial resources to facilitate the HRD of the public service.*

Friedl, J.E. 2006. *Mastering Regular Expressions: Understand Your Data and Be More Productive.*

GEO DENOON. 1943. *Development of Methods of Land Registration in South Africa.*

Gordon, R., Nell, M. & Di Lollo, A. 2011. *Investigation into the delays in issuing Title Deeds to Beneficiaries of Housing Projects funded by the capital subsidy.*

Graglia, J.M. & Mellon, C. 2018. *Blockchain and Property in 2018: At the End of the Beginning.*

Hayes, B., Beres, T. & Diesch, M. 2005. *Posting data to a database from non-standard documents using document mapping to standard document types.* U.S. Patent Application 11/083,546:.

Henssen, J. 1975. *Cadastres, including some aspects of assessment of real property.*

Heyink Mark. 2018. *A giant leap - 2. A Small Step in Electronic Signature - A Giant Leap for Electronic Deeds Registration.* Lexis Nexis.

Hocking, J. & Puttkammer, M. 2016. *Optical character recognition for South African languages.*

Kochan, D.J. 2013. *Certainty of Title: Perspectives After the Mortgage Foreclosure Crisis on the Essential Role of Effective Recording Systems.*

Koma, S.B. 2012. *The evolution of developmental local government in South Africa: Issues, trends and options.*

Korngold, G. 2008. *Legal and Policy Choices in the Aftermath of the Subprime and Mortgage Financing Crisis.*

Kotzé, G. & Wolff, F. 2017. *Developing and evaluating a pipeline for Setswana OCR.*

Lang, A.G. 1981. *Computerised Land Title and Land Information.*

Larsson, G. 1991. *Land registration and cadastral systems: tools for land information and management.*

Lubbe, L. 2013. *Sectional title property in South Africa: an accounting and auditing perspective.* University of the Free State.

Mihov, S., Schulz, K.U., Ringlstetter, C., Dojchinova, V., Nakova, V., Kalpakchieva, K., Gerasimov, O., Gotscharek, A. et al. 2005. *A corpus for comparative evaluation of OCR software and postcorrection techniques.*

Milligan, I. 2013. *Illusionary order: Online databases, optical character recognition, and Canadian history, 1997-2010.*

Minister of Rural Development and Land Reform. 2017. *Electronic Deeds Registration Systems (B35-2017).* Available: https://pmg.org.za/bill/749/.

Mithe, R., Indalkar, S. & Divekar, N. 2013. *Optical character recognition.*

Mori, S., Suen, C.Y. & Yamamoto, K. 1992. *Historical review of OCR research and development.*

Mutula, S.M. & Mostert, J. 2010. *Challenges and opportunities of e-government in South Africa.*

Ndou, S. & Sebola, M. 2016. *Capacity building in local government: an analysis for application of competency-based training in South Africa.*

Nengwekhulu, R.H. 2009. *Public service delivery challenges facing the South African public service.*

Netchaeva, I. 2002. *E-government and e-democracy: a comparison of opportunities in the north and south.*

Newcomer, D.A., Seely, J.S., Branham, D.L. & Kosan, P. 2013. *Property record document data verification systems and methods.* U.S. Patent 8,401,301:.

Nuno-Maganda, M.A. & Arias-Estrada, M.O. 2005. *Real-time FPGA-based architecture for bicubic interpolation: an application for digital image scaling.*

Ohlsson, V. 2016. *Optical Character and Symbol Recognition using Tesseract.* Luleå University of Technology.

OpenCV Team. 2018. *About - OpenCV library.* Available: https://opencv.org/about.html.

Pandey, A., Sharma, V., Paanchbhai, S., Hedaoo, N. & Zade, S.D. 2017. *Optical Character Recognition (OCR).*

Park, J., Govindaraju, V. & Srihari, S.N. 2000. *OCR in a hierarchical feature space.*

Parliament South Africa:. 2018. *Title Deeds backlog: Western Cape Department of Human Settlements and Land Reform briefing; Deeds Office: Process of registering deeds | PMG.* Available: https://pmg.org.za/committee-meeting/26921/ [Nov 11, 2018].

Parliamentary Monitoring Group. 2018. *Electronic Deeds Registration Systems | PMG.* Available: https://pmg.org.za/bill/749/ [Nov 28, 2018].

Patel, C., Patel, A. & Patel, D. 2012. *Optical character recognition by open source OCR tool tesseract: A case study.*

Powell, D. 2012. *Imperfect transition–local government reform in South Africa 1994-2012.*

Radloff, F.G.T. 1996. *Land Registration and Land Reform in South Africa * Available: https://heinonline.org/HOL/P?h=hein.journals/jmlr29&i=831 .

Rewagad, P. & Pawar, Y. 2013. *Use of digital signature with diffie hellman key exchange and AES encryption algorithm to enhance data security in cloud computing.*

Roux, J., Calzolari, N., Choukri, K., Declerck, T., Goggi, S. & Grobelnik, M. 2016. *South African National Centre for Digital Language Resources.*

Sabir, E., Rawls, S. & Natarajan, P. 2017. *Implicit Language Model in LSTM for OCR.*

Shange, M.B. 2010. *A system-based approach to land registration analysis and improvements: a case study of the KwaZulu-Natal deeds registration system.* University of KwaZulu-Natal.

Singh, R., Yadav, C.S., Verma, P. & Yadav, V. 2010. *Optical character recognition (OCR) for printed Devanagari script using artificial neural network.*

Smith, R. 2007. *An overview of the Tesseract OCR engine.*

Sonkusare, M. & Sahu, N. 2016. *A survey on handwritten character recognition (HCR) techniques for English alphabets.*

South Africa. 1937. *Deeds Registries Act, 47 of 1936.*

Tauschek, G. 1935. *Reading machine.* US Patent 2026330A:.

Tesseract. 2018. *ImproveQuality.* Available: https://github.com/tesseract-ocr/tesseract/wiki/ImproveQuality.

Ting, L. & Williamson, I.P. 1999. *Cadastral trends: A synthesis.*

Toulmin, C. 2009. *Securing land and property rights in sub-Saharan Africa: the role of local institutions.*

Turchin, A., Kolatkar, N.S., Grant, R.W., Makhni, E.C., Pendergrass, M.L. & Einbinder, J.S. 2006. *Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes.*

Ugale, M.K., Patil, S.J. & Musande, V.B. 2017. *Document management system: A notion towards paperless office.*

Ullah, R., Sohani, A., Rai, A., Ali, F. & Messier, R. 2018. *OCR Engine to Extract Food-Items, Prices, Quantity, Units from Receipt Images, Heuristics Rules Based Approach.* Available: https://www.researchgate.net/publication/323640080_OCR_Engine_to_Extract_Food-Items_Prices_Quantity_Units_from_Receipt_Images_Heuristics_Rules_Based_Approach.

Williamson, I.P. 1997. *The justification of cadastral systems in developing countries.*

World Bank. 2018a. *Doing Business 2019 Training for Reform.* The World Bank.

World Bank. 2018b. *Doing Business 2019 Training for Reform Economy Profile South Africa.* Available: http://www.doingbusiness.org/content/dam/doingBusiness/country/s/south-africa/ZAF.pdf [Nov 13, 2018].

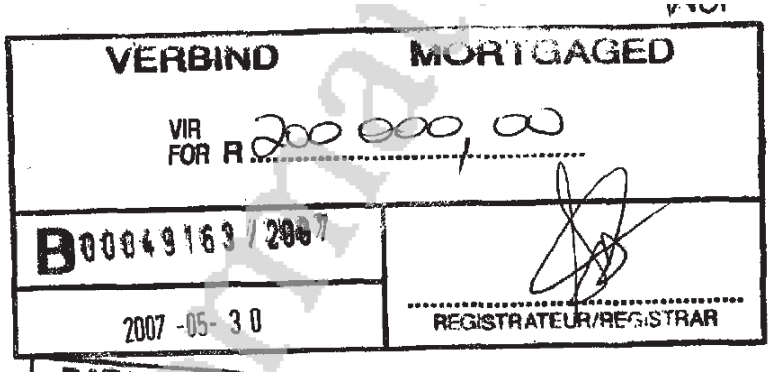
Yu-qian, Z., Wei-hua, G., Zhen-cheng, C., Jing-tian, T. & Ling-Yun, L. 2006. *Medical images edge detection based on mathematical morphology.*

Appendix

Appendix A: Sample Title Deed[10]

Prepared by me

CONVEYANCER
HILTON SMITH



DEED OF TRANSFER

ABC ATTORNEYS
1 BEACH STREET
CAPE TOWN
1234
(021) 911 01123

BE IT HEREBY MADE KNOWN THAT

JOHN BROWN

appeared before me, REGISTRAR OF DEEDS at Pretoria, he the said Appearer, being duly authorised thereto by Powers of Attorney signed at Pretoria on 9 October 2011 granted to him by

FORTUNE AVIATION LIMITED
Registration Number 2004/0938472/71

[10] All personal information on title deed has been changed

And the appearer declared that his said principical had truly and legally sold the undermentioned property on 11 November 2011 for an amount of R618 000,00 (SIX HUNDRED AND EIGHTEEN THOUSAND RAND);

NOW, THEREFORE, I, hereby cede and transfer, the State, however,r eserving its rights, to and on behalf of

MICHELLE CARTER
Identity Number: 9501174800084
Unmarried

in full and free property to

ERF 2098 MILNERTON
IN THE CITY OF CAPE TOWN
DIVISION CAPE
PROVINCE WESTERN CAPE

IN EXTENT: 77 (SEVENTY SEVEN) Square Metres

AS WILL APPEAR from General Plan No S.G.NO. 10982/2010

AND HELD by CERTIFICATE OF CONSOLIDATED TITLE No. T54321/2012

SUBJECT to the conditions contained in Deed of Transfer No. T12345/1987.

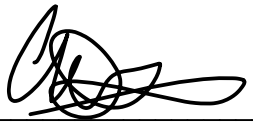
SIGNED at CAPE TOWN on **10** day of **February 2012.**

DULY AUTHORIZED AGENT

Before me

CONVEYANCER
HILTON SMITH

Registered at Cape Town on 25 FEB 2012



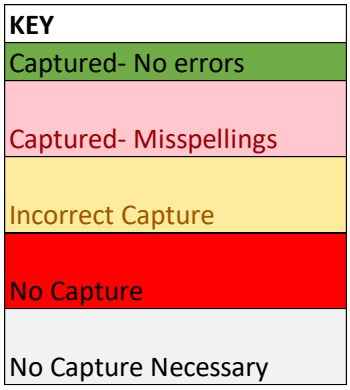
REGISTRAR OF DEEDS

Appendix B: Glossary

Blockchain	-	A blockchain is a distributed digital ledger of transactions which are validated by nodes on the network. As each block in the chain cryptographically depends on the previous blocks, the transactions are considered immutable. This is because any changes to a transaction on a block require all subsequent blocks to be altered which is intentionally difficult[11].
Cadastral System	-	This system encompasses the maps of land parcels and the title and mortgage registers of a country[12].
Conveyancer	-	A conveyancer is a lawyer who assists in registering properties and mortgages by compiling the relevant documents to lodge[13].
(Cryptographic) Hash	-	When an input of any size goes through cryptographic hashing, a hash of a fixed length is produced. This hashing function is a one-way function as the input cannot be deciphered from the hash. The same input will always create the same hash[14].
Freehold Title	-	This title is in relation to a property in which the owner has full rights over the property and land underneath[15] such as a house.
Fractional Ownership	-	This is an ownership structure of property where two or more people co-own a property in specified shares.
Sectional Title	-	This is in reference to a property where ownership is of units/sections within a complex or development such as apartments and townhouses. This type of ownership requires compliance with the Sectional Titles Act.[16]

[11] https://techterms.com/definition/blockchain
[12] https://mmm.fi/en/land-surveying-and-spatial-information/cadastral-system-and-surveying
[13] https://www.justlanded.com/english/South-Africa/South-Africa-Guide/Property/Conveyance
[14] https://hackernoon.com/cryptographic-hashing-c25da23609c3
[15] https://economictimes.indiatimes.com/definition/freehold-property
[16] https://www.property24.com/articles/sect-title-vs-freehold-which-is-best/12318

Appendix C: CSV Output from OCR Pipeline (with comparison)



Notes:

*The share portion is automatically 100% for all properties. This default only changes when there is a fractionally owned property based on the share owned.

** There are no fields for comparison for the hash value compared to the manual data capture. This is because the hash is only generated when the scan is uploaded to IPFS and not found on the title deed.

The password used to encrypt files is "abc123" and can be used to decrypt files at these hash locations.

Table 8: OCR Automatic Capture Output

buyerfirstname	buyersurname	buyerID	sellerfirstname	sellersurname	sellerID	erfno	title_num	size	datetransactio	dateregistration	sectionplan_num	section_number	price	bondamount	share*	filehash**
DAVID	WILLIAMS	730114 1234 77 8	PIZZA COMPANY LTD		1996/010341/07	1234 SEA POINT	T125689/2012	360	09-Nov-12				R2 120 000,00	0	0.05	GtUgEMK1LrfPdUoRB2YhK j75PJL
JOYCE	SMITH	750816 1234 79 8	PIZZA COMPANY LTD		1996/010341/07	1234 SEA POINT	T125689/2012	360	09-Nov-12				R2 120 000,00	0	0.6	GtUgEMK1LrfPdUoRB2YhK j75PJL
JACOB	BROWN	750816 1234 79 8	PIZZA COMPANY LTD		1996/010341/07	1234 SEA POINT	T125689/2012	360	09-Nov-12				R2 120 000,00	0	0.35	GtUgEMK1LrfPdUoRB2YhK j75PJL
ABRAHAM	CARLSON	7703015800081	VICTORIA	GARZA	8805014800084	4321 PINELANDS	T7390283/201 2	402	09-Nov-12				R1 125 000,00	0	0.1	D4WGnswkRbxXHxgEQEP sSTan4hsKu
DEANNA	GRAHAM	4901094800088	VICTORIA	GARZA	8805014800084	4321 PINELANDS	T7390283/201 2	402	09-Nov-12				R1 125 000,00	0	0.45	D4WGnswkRbxXHxgEQEP sSTan4hsKu
GRANT	MEYER	6203095800086	VICTORIA	GARZA	8805014800084	4321 PINELANDS	T7390283/201 2	402	09-Nov-12				R1 125 000,00	0	0.45	D4WGnswkRbxXHxgEQEP sSTan4hsKu
CANDICE	DAY	5108014800086	MIGUEL	HORTON	8805014800084	302 RONDEBOSCH	T8034801/201 0	160	09-Nov-12	04-Dec-10			R1 125 000,00	0	0.2	u9rGWUxRougZAoV1gWrP 1oLXWco
MELODY	CASEY	6708034800088	MIGUEL	HORTON	8805014800084	302 RONDEBOSCH	T8034801/201 0	160	09-Nov-12	04-Dec-10			R1 125 000,00	0	0.5	u9rGWUxRougZAoV1gWrP 1oLXWco
CHRISTIE	FLEMING	8008014800085	MIGUEL	HORTON	8805014800084	302 RONDEBOSCH	T8034801/201 0	160	09-Nov-12	04-Dec-10			R1 125 000,00	0	0.15	u9rGWUxRougZAoV1gWrP 1oLXWco
TINA	BRIGS	6408034800085	MIGUEL	HORTON	8805014800084	302 RONDEBOSCH	T8034801/201 0	160	09-Nov-12	04-Dec-10			R1 125 000,00	0	0.15	u9rGWUxRougZAoV1gWrP 1oLXWco
DIANNA	WATSON	8611174800082	WONDERPRISES LIMTED		1971/09829182/ 98	329 GOODWOOD	T63/710	107	18-Nov-07	28-Dec-07			R1 418 000,00	0	100%	xCxLAsdqufBHtVk4BaHuVi DbKk
SEAN	ALLEN	8309075800083	WONDERPRISES LIMTED		1971/09829182/ 98	329 GOODWOOD	T63/710	107	18-Nov-07	28-Dec-07			R1 418 000,00	0	100%	xCxLAsdqufBHtVk4BaHuVi DbKk
CARL	TORRES	6507055800088	RAPTOR NAVIGATIONS LIMTED		1971/09829182/ 98	3231 WYNBERG	T63710/2006	147	10-Jul-04				R1 738 000,00	0	100%	vNxSrok6ZuXXsg5tSJKvAD bgi
DOROTHY	THOMAS	6202144800089	RAPTOR NAVIGATIONS LIMTED		1971/09829182/ 98	3231 WYNBERG	T63710/2006	147	10-Jul-04				R1 738 000,00	0	100%	vNxSrok6ZuXXsg5tSJKvAD bgi
HARRIET	TYLER	5601034800086	TITANIUM SPORTS LIMTED		1998/018392/0 9	2098 MILNERTON	T12345/1987	77	09-Dec-12				R1 018 200,00	0	100%	oSY8ZJY6voM2edx4pfTMn KNCvkF
MICHELLE	CARTER	9501174800084	FORTUNE AVIATION LIMTED		2004/093847 2/71	2098 MILNERTON	T12345/1987	77	11-Nov-11	09-Feb-12			R618 000,00	0	100%	ssmRmZAZ65TrLGhfdGsY eJLCXu
JUDY	SANCHEZ	6204104800083	PRODIGY SOLUTION LIMTED		2010/982017 2/33	2098 CONSTANTIA	T528302/2011	260	13-Apr-11				R3 612 300,00	0	100%	Z5m2wiS13J13zpY7aK2qtZ snr
PETER	WILLIAMS	730624 1234 78 9	GOVERNMENT OF THE WESTERN CAPE			1234 VUKUZENZELE	T54321/19201 2	68	09-Jul-17				R120 000,00	0	100%	mtPeNztZQ9SrcMngAK6nT iqTMu
JOYCE	WILLIAMS	750323 1234 78 8	GOVERNMENT OF THE WESTERN CAPE			1234 VUKUZENZELE	T54321/19201 2	68	09-Jul-17				R120 000,00	0	100%	mtPeNztZQ9SrcMngAK6nT iqTMu
HARRY	POTTER	720202 1234 78 9	GOVERNMENT OF THE WESTERN CAPE			1998 PHILIPPI	T12345/1979	73	09-Jul-17				R18 800,00	0	100%	T5X9qmNSDgwosfd75DcU BfB6R4
HERMIONE	GRANGER	800324 1234 78 9	GOVERNMENT OF THE WESTERN CAPE			1999 PHILIPPI	T12345/1979	82	09-Jul-17				R20 00,00	0	100%	GPTCAdbZhhY8Cww4tUS mUiU6MQ
OLIVER	WOODS	6104104800085	GUILLERMO	SINGLETON	7004105800085		ST 829301/2010	90	09-May-12		SS 111/2008	202	R1 130 000,00	0	100%	okNNMSTNakBN9LCJtAXb tcPnzR
OLIVER	WOODS	6104104800085	GUILLERMO	SINGLETON	7004105800085		ST 829301/2010	20	09-May-12		SS 111/2008	217	R1 130 000,00	0	100%	okNNMSTNakBN9LCJtAXb tcPnzR
DEBBIE	BALLARD	5402064800082	ERICK	HENDERSON	5402105800083		ST 834017/2012	55	01-May-17	09-Jun-17	SS 333/2002	303	R2 750 000,00	0	100%	UypVh1eH7TT8Ff3MdJqx4 7Rtg
DEBBIE	BALLARD	5402064800082	ERICK	HENDERSON	5402105800083		ST 834017/2012	10	01-May-17	09-Jun-17	SS 333/2002	304	R2 750 000,00	0	100%	UypVh1eH7TT8Ff3MdJqx4 7Rtg
BERNADETTE	MORTON	6505064800083	JOSEFINA	MATTHEWS	7706065800083		ST 3450983/2010	105	31 April 2017		SS 444/2002	444	R3 905 000,00	0	100%	ZCuMDk4M5n5y3TGNGW8 v92LHppPH
BERNADETTE	MORTON	6505064800083	JOSEFINA	MATTHEWS	7706065800083		ST 3450983/2010	19	31 April 2017		SS 444/2002	134	R3 905 000,00	0	100%	ZCuMDk4M5n5y3TGNGW8 v92LHppPH
THOMAS	MILLER	780331 1034 77 8	DAVID SAMUEL	WILSON	851214 1234 77 8		ST 98765/2010	16	09-Apr-17	08-Jun-17	SS 222/2008	100	R1 125 000,00	0	100%	oY2tVyLgGSZPiF5Ssd7xT RNhh9
THOMAS	MILLER	780331 1034 77 8	DAVID SAMUEL	WILSON	851214 1234 77 8		ST 98765/2010	47	09-Apr-17	08-Jun-17	SS 222/2008	101	R1 125 000,00	0	100%	QmTRBxgtYaRpVYD1KNq oY2tVyLgGSZPiF5Ssd7xT RNhh9

Appendix D: CSV Output from Manual Capture

Notes:

The errors found in manual capture have been corrected but the blocks where errors were found are highlighted.

Table 9: Manual Data Capture

buyerfirstname	buyersurname	buyerID	sellerfirstname	sellersurnam	sellerID	erfno	title_num	size	datetransaction	dateregistration	sectionplan_num	section_number	price	bondamount	share
DAVID	WILLIAMS	730114 1234 77 8	Pizza Company LTD		1996/010341/077	1234 SEA POINT	T125689/2012	360	09-Nov-12	24-Jan-13			R2 120 000.00	0	0.05
JOYCE	SMITH	7508161234 798	Pizza Company LTD		1996/010341/077	1235 SEA POINT	T125689/2012	360	09-Nov-12	24-Jan-13			R2 120 000.00	0	0.6
JACOB	BROWN	750816 1234 798	Pizza Company LTD		1996/010341/077	1236 SEA POINT	T125689/2012	360	09-Nov-12	24-Jan-13			R2 120 000.00	0	0.35
ABRAHAM	CARLSON	7703015800081	VICTORIA	GARZA	8805014800084	4321 PINELANDS	T7390283/2012	402	09-Nov-12	24-Jan-13			R1 125 000.00	0	0.1
DEANNA	GRAHAM	4901094800088	VICTORIA	GARZA	8805014800084	4321 PINELANDS	T7390283/2012	402	09-Nov-12	24-Jan-13			R1 125 000.00	0	0.45
GRANT	MEYER	6203095800086	VICTORIA	GARZA	8805014800084	4321 PINELANDS	T7390283/2012	402	09-Nov-12	24-Jan-13			R1 125 000.00	0	0.45
CANDICE	DAY	5108014800086	MIGUEL	HORTON	8805014800084	302 RONDEBOSCH	T8034801/2010	160	09-Nov-12	14-Dec-10			R1 1250 000.00	0	0.2
MELODY	CASEY	6708034800088	MIGUEL	HORTON	8805014800084	302 RONDEBOSCH	T8034801/2010	160	09-Nov-12	14-Dec-10			R1 1250 000.00	0	0.5
CHRISTIE	FLEMING	8008014800085	MIGUEL	HORTON	8805014800084	302 RONDEBOSCH	T8034801/2010	160	09-Nov-12	14-Dec-10			R1 1250 000.00	0	0.15
TINA	BRIGS	6408034800085	MIGUEL	HORTON	8805014800084	302 RONDEBOSCH	T8034801/2010	160	09-Nov-12	14-Dec-10			R1 1250 000.00	0	0.15
DIANNA	WATSON	8611174800082	WONDERPRISES LIMITED		1971/098291829/8	329 GOODWOOD	T63710/2006	107	18-Nov-07	28-Dec-07			R1 418 000.00	0	100%
SEAN	ALLEN	8309075800083	WONDERPRISES LIMITED		1971/098291829/8	329 GOODWOOD	T63710/2006	107	18-Nov-07	28-Dec-07			R1 418 000.00	0	100%
CARL	TORRES	6507055800088	RAPTOR NAVIGATIONS LIMITED		1971/098291829/8	3231 WYNBERG	T752930/2008	147	10-Jul-04	17-Nov-04			R1 738 000.00	0	100%
DOROTHY	THOMAS	6202144800089	RAPTOR NAVIGATIONS LIMITED		1971/098291829/8	3231 WYNBERG	T752930/2008	147	10-Jul-04	17-Nov-04			R1 738 000.00	0	100%
HARRIET	TYLER	5601034800086	TITANIUM SPORTS LIMITED		1998/0183921/09	2098 MILNERTON	T12345/1987	77	09-Dec-12	24-Jan-13			R1 018 200.00	0	100%
MICHELLE	CARTER	9501174800084	FORTUNE AVIATION LIMITED		2004/0938472/71	2098 MILNERTON	T12345/1987	77	11-Nov-11	25-Feb-12			R618 000.00	R2000 000.00	100%
JUDY	SANCHEZ	6204104800083	PRODIGY SOLUTION LIMITED		2010/9820172/33	2098 CONSTANTIA	T528302/2011	260	13-Apr-11	29-Jun-11			R3 612 300.00	0	100%
PETER	WILLIAMS	730624 1234 78 9	GOVERNMENT OF THE WESTERN CAPE			1234 VUKUZENZELE	T54321/192012	68	09-Jul-17	24-Jan-13			R120 000.00	0	100%
JOYCE	WILLIAMS	7503231234788	GOVERNMENT OF THE WESTERN CAPE			1234 VUKUZENZELE	T54321/192012	68	09-Jul-17	24-Jan-13			R120 000.00	0	100%
HARRY	POTTER	720202 1234 78 9	GOVERNMENT OF THE WESTERN CAPE			1998 PHILIPPI	T12345/1979	73	09-Jul-17	24-Jan-13			R18 800.01	0	100%
HERMIONE	GRANGER	8003241234789	PROVINCIAL GOVERNMENT OF THE WESTERN CAPE			1999 PHILIPPI	T12345/1979	82	09-Jul-17	24-Jan-13			R20 700.00	0	100%
OLIVER	WOODS	6104104800085	GUILLERMO	SINGLETON	7.00411E+12		ST 829301/2010	90	09-May-12	29-Jun-17	SS 111/2008	202	R1 130 000.00	0	100%
OLIVER	WOODS	6104104800085	GUILLERMO	SINGLETON	7.00411E+12		ST 829301/2010	20	09-May-12	29-Jun-17	SS 111/2008	211	R1 130 000.00	0	100%
DEBBIE	BALLARD	5402064800082	ERICK	HENDERSON	5.40211E+12		ST 834017/2012	55	01-May-17	29-Jun-17	SS 333/2002	303	R2 750 000.00	R510 000.00	100%
DEBBIE	BALLARD	5402064800082	ERICK	HENDERSON	5.40211E+12		ST 834017/2012	10	01-May-17	29-Jun-17	SS 333/2002	304	R2 750 000.00	R510 000.00	100%
BERNADETTE	MORTON	6505064800081	JOSEFINA	MATTHEWS	7.70607E+12		ST 3450983/2010	105	31 April 2017	10-Feb-10	SS 444/2002	444	R3 905 000.00	R3 000 000.00	100%
BERNADETTE	MORTON	6505064800081	JOSEFINA	MATTHEWS	7.70607E+12		ST 3450983/2010	15	31 April 2017	10-Feb-10	SS 444/2002	134	R3 905 000.00	R3 000 000.00	100%
THOMAS	MILLER	780331 1034 77 8	DAVID SAMUEL	WILSON	851214 1234 77 8		ST 98765/2010	16	09-Apr-17	29-Jun-17	SS 222/2008	100	R1 125 000.00	0	100%
THOMAS	MILLER	780331 1034 77 8	DAVID SAMUEL	WILSON	851214 1234 77 8		ST 98765/2010	47	09-Apr-17	29-Jun-17	SS 222/2008	101	R1 125 000.00	0	100%

Appendix E: List of Packages and Software

Required programmes external to python required for code to work:

1. Image Pre-Processing: ImageMagick
2. OCR: Tesseract
3. IPFS

These programmes are available for download at the following links:

1. https://imagemagick.org/script/download.php
2. https://github.com/tesseract-ocr/tesseract/wiki/Downloads
3. https://docs.ipfs.io/introduction/install/

The following are the python packages used:

PDF Conversion:

1. io
2. wand
3. PIL
4. import cv2
5. numpy
6. os

IPFS and Encryption:

1. ipfsapi
2. Cryptodomex

Image Pre-Processing

1. cv2
2. numpy

OCR

1. pytesseract

Regex (Text Parsing)

1. re

Write to CSV

1. csv

Appendix F: Description of Code-Base

The following files are included in the OCR package which can be found on the github page:

- allpreprocessing: This file has the OCR results for the simulated deeds and links to the create deed function, so the CSV can be generated.

- convertpdf: This contains function to convert pdf to images and stores the image paths in a list for future use. This function also encrypts the pdf file (using functions from encrypt file) and adds the encrypted file to IPFS. If IPFS daemon is not running, the conversion to images will continue and a message will print indicating that IPFS part of function could not be completed and file hash will be an empty string. The function takes the file path to a pdf. If the file is a pdf, the function will return the list of image paths and the file hash for the encrypted pdf. If the file is not a pdf, the function will return the file path and the file hash.

- deskew: This function takes an image, detects any skewness and rotates image accordingly.

- encrypt: This file contains AES encryption and decryption functions using a user created password. This uses pycryptodome for encryption and decryption.

- fullocr: This contains the function which takes functions from the other files to create a full pipeline from pdf to data capture in a CSV. The input parameters are the pdf file path and possible parameters to adjust text cleaning process. This function then calls the function from convertpdf to convert pdf to images, encrypt pdf and upload encrypted pdf to IPFS. The images are then pre-processed before OCR by calling the clean_image function from imageclean. OCR using pytesseract is performed with the text output appended to a list for each output page. The text output then goes through the create_deed function from regex which extracts the relevant data using regular expressions and writes it to a CSV file.

- imageclean: This has the function for pre-processing of images to improve performance of OCR. The pre-processing function contains binarization, rescaling, removing noise, deskewing and converting image to greyscale. This function takes an image path as a parameter.

- ipfs: run this file to ensure connection to IPFS is working. This should be done before using the fullocr package or else files will not be uploaded to IPFS. This is a tester file and doesn't link to any other file.

- methodology: This contains the code for the demonstrating how different parts of the pre-processing appears visually. This file was used to create the images within methodology chapter of dissertation.

- regex: This file has the compiled regular expressions and function used for extracting relevant data from OCR results and cleaning OCR results for noisy characters.

- titledeed: This file contains the classes DeedsRegsitry and TitleDeed. For the DeedsRegsitry class, there is a method to create a new registry (i.e. database/CSV) if none exists. The TitleDeed class stores the relevant data extracted as attributes. The methods within this class are create_dictionary and write_to_csv. create_dictionary takes the attributes from the object and creates the relevant number of dictionaries based on the type of deed and how many rows need to be recorded. write_to_csv then takes these dictionaries and writes them to a CSV file where each dictionary represents one row of the CSV.

Appendix G: Github Link

This is the link to my github repository:

https://github.com/ashfavish/OCR_Deeds_Registry

The following resources can be found uploaded to the repository

1) The PDFS of the simulated title deeds
2) The code-base for the OCR process used