Identifying Real Estate Development Opportunities: Web-Scraping, Regex Patterns & String-Searching Algorithms

by

Oscar Williams
Bachelor of Construction Management & Property
University of New South Wales

Submitted to the Program in Real Estate Development in Conjunction with the Center for Real Estate in Partial Fulfillment of the Requirements for the Degree of Master of Science in Real Estate Development

at the
Massachusetts Institute of Technology
June 2021

Signature of Author____

Center for Real Estate
April 8, 2021

Certified by____

Professor William Wheaton
Professor, Center for Real Estate
Thesis Supervisor

Accepted by____

Professor Siqi Zheng

Samuel Tak Lee Professor of Urban and Real Estate Sustainability
Faculty Director, Center for Real Estate & Sustainable Urbanization Lab

Identifying Real Estate Development Opportunities: Web-Scraping, Regex Patterns & String-Searching Algorithms

by

Oscar Williams

Submitted to the Program in Real Estate Development in Conjunction with the Center for Real Estate on April 8, 2021 in Partial Fulfillment of the Requirements for the Degree of Master of Science in Real Estate Development

ABSTRACT

Web-scraping and data mining algorithms are used extensively by hedge funds, equities traders, digital marketers and in the technology sector more broadly. Contrastingly, the real estate development industry continues to use traditional, manual methods to identify and pursue new development opportunities with the exception of mapping software which has been widely adopted. The lack of adoption of these technologies is primarily due to the difficulty in identifying, retrieving and processing the required data rather than an inherent lack of data. To the contrary, there is a wealth of public and private information available to the real estate development industry that can provide value if collected and analyzed efficiently and at scale using algorithms. To test this hypothesis, the author has built a functioning web-scraping and data collection platform that demonstrates how large amounts of data can be retrieved and processed at scale. This thesis evaluates the effectiveness of using web-scraping algorithms to search for real estate development and land rezoning opportunities from publicly available local Government data. The focus area of the thesis is Sydney, Australia and the subject of the thesis is the *Aiden*[1] platform that is owned by the Principal Investigator and author. The platform uses automated web-scraping algorithms to parse publicly available local Government data for keywords that indicate a prospective development opportunity or an instance of imminent land rezoning. The results of this research demonstrate the effectiveness of adopting web-scraping technologies and the usefulness to real estate development professionals.

The Aiden platform can be accessed at www.aidendata.com using login details that may be provided upon request via oscarw@mit.edu.

[1] *Aiden* Platform URL www.aidendata.com

ACKNOWLEDGEMENTS

Firstly, I would like to thank the MIT community for creating a welcoming, interesting and exciting place for me to learn and study.

Secondly, I would like to thank Jerald Fuller who provided a wealth of technical advice and oversight with respect to web-scraping technology descriptions and processes.

Thirdly, I would like to thank Annie Thompson for her assistance and advice with the statistical analysis.

Lastly, I would like to thank my supervisor, Professor Bill Wheaton, for his direction and oversight.

FIGURES & TABLES

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 THESIS BACKGROUND

The use of technology and automated data analysis in the real estate development industry is in its infancy in comparison to other industries including finance, manufacturing, technology, marketing, telecommunications and many others (Mohanram, 2020). The primary reason for this is that the data used by professionals in the real estate development industry are often fragmented and difficult to identify and access. This thesis focuses on the acquisitions process and the methods used to identify and acquired new real estate development opportunities. Furthermore, this paper contends that current acquisition processes and methods are antiquated, highly manual and ripe for disruption and improvement via the use of search algorithms and mass data collection and analysis.

As a solution, this paper focuses on the *Aiden* platform. *Aiden* is a proprietary web-scraping platform that was built by the author. The platform demonstrates how web-scraping and more specifically, string-searching algorithms, can be used to collect data that can assist real estate development professionals. This paper explores how algorithms can complete work that would otherwise require thousands of hours of manual labor and how *Aiden*, or similar platforms, can prove to be an invaluable investment discovery tool to real estate development professionals.

1.2 PROBLEM STATEMENT

The real estate development industry has yet to widely adopt technologies that enable the retrieval and analysis of large amounts of structured and unstructured data from multiple sources, also known as 'big data' (Winson-Geideman et al., 2017). The primary cause of this is that the data is highly fragmented and difficult to retrieve. The focus area of this research is Sydney, Australia. In the Sydney Metropolitan Area, there are 31 separate Local Government Areas (LGAs) and each of their websites have different structures and different URLs. By law, each Council is required to publish planning information including Council meeting *Agendas* and *Minutes* that include valuable information about rezoning opportunities. Each of these documents are up to 600 pages in length and thousands of hours of labor would be required to review all of these documents each month. The proposed solution to this is the *Aiden* platform. *Aiden* is a web-scraping platform that crawls each of the source URLs in order to find keywords within PDF files. As new documents are discovered each month, the algorithms parse the text for user-defined keywords that may indicate a real estate development or rezoning opportunity.

1.3 RESEARCH AIM & OBJECTIVES

The specific objectives of the research are as follows:

1. To evaluation the technology required to build web-scraping algorithms and a cohesive platform that allows real estate development professions to use them at scale targeting multiple data sources;
2. To evaluate what keywords are likely to signify a real estate development or rezoning opportunity within unstructured data, primarily, local Government meeting reports and minutes;
3. To examine the quality of the output of the *Aiden* platform using 31 data sources comprizing the Local Government Areas in the Metropolitan region of Sydney, Australia.

1.4 SCOPE OF STUDY

The research seeks to evaluate the usefulness of web-scraping technology to professionals in the real estate development industry. Moreover, the study attempts to identify and prescribe technology that can be used to build a platform similar to the *Aiden* platform. The research is limited to the Sydney Metropolitan region and the use of a generic set of keywords collated by the author. The keywords selected are broad and are not specific to a given region or locale. Further research needs to be undertaken to refine and extend the keyword selection. This study does not attempt to evaluate the exact value of the data, nor does it explore the legal considerations that would likely need to be made when collecting this data in a given jurisdiction.

1.5 HYPOTHESIS

As explored in the literature review, web-scraping algorithms can parse vast amounts of sporadic data in a highly efficient and automated way. Furthermore, data retrieval is challenging for professionals in the real estate development industry due to the fact that information is stored in numerous different location that are difficult to access. These two observations lead to the hypothesis that :

> *'Web-scraping algorithms can be a useful tool for real estate development professions to discover new acquisition opportunities'*

1.6 THESIS STRUCTURE

The thesis is composed of the following Chapters:

Chapter 1: provides an introduction to the research, the problem and the hypothesis

Chapter 2: provides a literature review exploring relevant literature relating to web-scraping

Chapter 3: explores the technology used to build the web-scraping technologies and to retrieve and store data

Chapter 4: presents an overview of the *Aiden* platform and the technology used to build it

Chapter 5: presents a discussion of the output and findings

Chapter 6: presents the conclusion

CHAPTER 2: LITERATURE REVIEW

2.1 INTRODUCTION

The literature review explores and evaluates literature that relates to web-scraping, data mining, data retrieval and analysis. Furthermore, the literature review assesses how data is currently being used by the real estate development industry and where there are further opportunities for the use of algorithms and technology more broadly.

2.2 INFORMATION COLLECTION

Information retrieval, processing and storage is growing exponentially and one of the key challenges of private companies in the 21st Century is how to leverage 'big data' to remain competitive (Snell and Menaldo, 2016). At the crux of this phenomenon is how to collect information. Information collection requires data retrieval from myriad sources in a both unstructured and structured form (Claussen and Peukert, 2019). In the field of data collection, there are two phrases that are commonly encountered, web-scraping and data mining (DM). Both terms are used broadly and often interchangeably yet have distinct differences.

2.2.1 WEB-SCRAPING & DATA RETRIEVAL

Web-scraping describes the action algorithms that are designed to parse information from data sources. This process is often referenced when discussing search engines such as Google that use 'web-crawlers' to search and extract information from billions of websites around the world each day. Web-scrapers are typically classified as string-searching algorithms that search strings (sentences) for keywords or other target data (Patel, 2020). The format of these algorithms is often written as a regular expression (REGEX) pattern. The most prevalent string-searching algorithm is the '*find and replace'* algorithm commonly used in word processors such as Microsoft Word (Santos, 2018).

Data Mining (DM) is a term used more broadly and often erroneously to describe the collection of 'big data'. 'Big data' is often characterized as a misnomer as it is commonly used in marketing material as a 'buzzword'. There is, however, a technical definition describing the processes of DM (Santos, 2018):

- Classification
- Regression
- Clustering
- Summarization
- Dependency Modelling
- Change and deviation detection

Moreover, there are three common DM processes that are widely used for data retrieval. The three process are Knowledge Discovery in Databases (KDD), SEMMA and CRISP-DM. The sub-processes of each of these methods are outlined below (Santos, 2018):

Knowledge Discovery in Databases (KDD) process

- Pre-processing
- Transformation
- Data Mining
- Interpretation/Evaluation

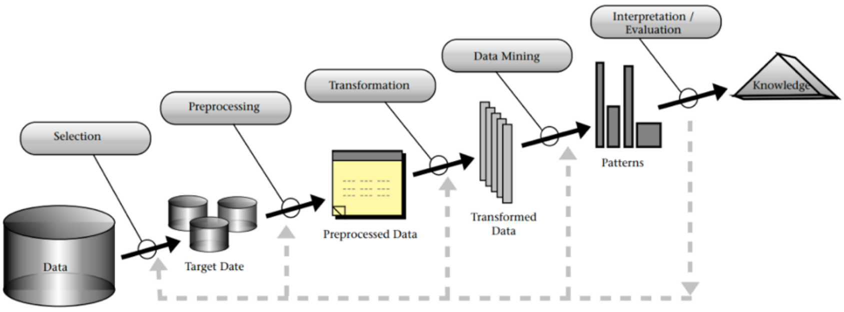


Figure 2.1: The five stages of KKD as described by Santos (Fayyad et al., 1996)

The SEMMA process

- Sample
- Explore
- Modify
- Model
- Assess

The CRISP-DM process (Cross-Industry Standard Process for Data Mining)

- Understanding the business
- Understanding the data
- Data preparation
- Modelling
- Evaluation
- Deployment

For the purposes of the thesis and in order for the *Aiden* platform to be able to undertake its basic operation, we have adopted relatively simple string-searching web-scrapers. There is a wealth of literature that explores the field of DM and more complicated web-scraping algorithms, however, the available information is vast and beyond the scope of this research.

2.3 IDENTIFYING REAL ESTATE REZONING OPPORTUNITIES USING TECHNOLOGY

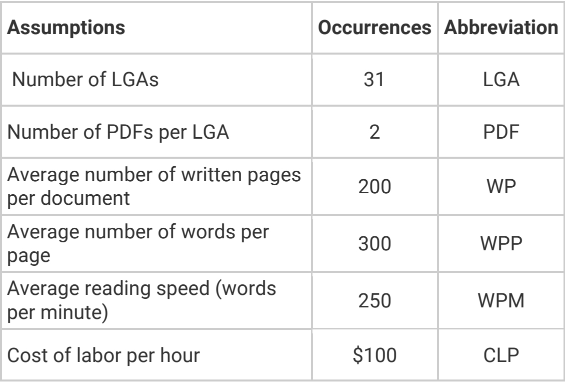
There are numerous ways that real estate professional search, identify and secure investment opportunities. Most of these methods are manual and rely of word-of-mouth and local knowledge. For the purpose of this research, land rezoning opportunities are the focus.

2.3.1 DEFINING ‘OPPORTUNITIES’

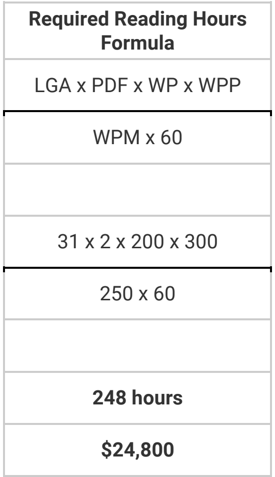
For the purposes of this research real estate development ‘opportunities’ refer to the discovery of information that is not readily available to the wider market and potential investment targets that are identified as a result of the analysis of this information. More specifically, the research assumes that information that indicates a new, Government-led, rezoning is of significant value to professionals and acquisitions managers in the real estate development industry as it provides them with a first-mover advantage to acquire property in a given area. In the focus area of Sydney, Australia, both the local and state Governments frequently rezone or ‘release’ new land for development. Typically, land is rezoned from rural to residential or additional height and density is provided if the area is in an inner-city location. Each of these scenarios present investment opportunities to real estate developers as statutory changes increases the value of the land.

2.4 DESIRED DATA

The focus data for the purpose of this thesis are the meeting agendas and minutes from the websites of the 31 Local Government Areas (LGAs) of the Sydney Metropolitan area. Under the *Government Information (Public Access) Act 2009 No 52*., also known as the GIPA Act, each LGA must upload their monthly meeting agenda and minutes to their website in PDF format. These documents are the focus of this research as they are commonly analyzed by professionals in the property development industry in order to obtain new information about planning decisions and rezoning initiatives. As previously noted, thousands of hours of manual labor would be required to read the documents each month as they are published on 31 separate websites and are up to 600 pages in length each. The time and associated cost of manually reading these documents can be estimated using the following formula:



Assumptions	Occurrences	Abbreviation
Number of LGAs	31	LGA
Number of PDFs per LGA	2	PDF
Average number of written pages per document	200	WP
Average number of words per page	300	WPP
Average reading speed (words per minute)	250	WPM
Cost of labor per hour	$100	CLP

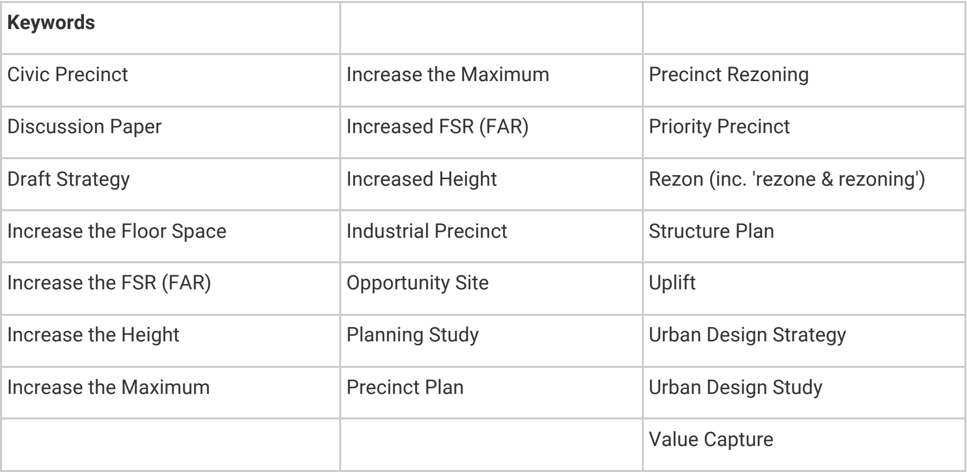


Required Reading Hours Formula
LGA x PDF x WP x WPP
WPM x 60
31 x 2 x 200 x 300
250 x 60
248 hours
$24,800

Table 2.1: Required Reading Hours Formula

2.5 KEYWORDS

An extensive review of relevant literature concluded that there is no readily available research on keywords that may indicate instances of a Government-initiated rezoning. Due to the absence of this this research, the author has composed a list of words that were discovered in actual LGA meeting minute and agenda documents that discuss actual instances of a rezoning. Any keywords may be input into the *Aiden* platform by a user. Once added to the system via the *Management* page, the algorithms will search for those words in every document that it crawls each day. If the keyword is detected, the application will note it as an 'occurrence' and add the document the *Newsfeed*. The paragraph, page number and sentence string will be high-lighted within the document and the user may click and view the information. The layout of the platform and the UI is described further in Chapter 4. For this study, the author has selected the following keywords that are commonly used in rezoning documentation:



Keywords		
Civic Precinct	Increase the Maximum	Precinct Rezoning
Discussion Paper	Increased FSR (FAR)	Priority Precinct
Draft Strategy	Increased Height	Rezon (inc. 'rezone & rezoning')
Increase the Floor Space	Industrial Precinct	Structure Plan
Increase the FSR (FAR)	Opportunity Site	Uplift
Increase the Height	Planning Study	Urban Design Strategy
Increase the Maximum	Precinct Plan	Urban Design Study
		Value Capture

Table 2.2: Keywords used to search for opportunities

CHAPTER 3: SCRAPING TECHNOLOGY

3.1 INTRODUCTION

This chapter describes the scraping technology used to retrieve, process and store the data from the LGA's websites. The descriptions provide insight into how the *Aiden* platform works and are intended to provide a recommended framework for building a similar application.

3.2 DATA SOURCES

The URLs for each of the 31 LGA websites are the data sources and 'scraping targets'. Each website structure was unique, therefore custom algorithms with different regular expression (REGEX) patterns had to be built and maintained for each URL. The individual URLs are provided in the table below:

Name	Scrape URL
North Sydney	https://www.northsydney.nsw.gov.au/Council_Meet...
Bayside Council	https://www.bayside.vic.gov.au/council-minutes
Burwood	http://www.burwood.nsw.gov.au/council_meetings_...
	http://www.burwood.nsw.gov.au/b_and_d_minutes_a...
Camden	https://www.camden.nsw.gov.au/council/council-m...
Campbelltown	https://www.campbelltown.nsw.gov.au/CouncilandC...
Canada Bay	http://www.canadabay.nsw.gov.au/calendar-of-cou...
Canterbury-Bankstown	https://www.cbcity.nsw.gov.au/council/Councilme...
Cumberland	http://cumberland.infocouncil.biz/
Fairfield	http://bpweb.fairfieldcity.nsw.gov.au:8080/fccbps/
Georges River	http://infoweb.georgesriver.nsw.gov.au/grinfoco...
The Hills Shire	https://www.thehills.nsw.gov.au/Council/Meeting...
	https://www.thehills.nsw.gov.au/Council/Meeting...
Hornsby Shire	http://businesspapers.hornsby.nsw.gov.au/
Hunter's Hill	http://www.huntershill.nsw.gov.au/Page/Page.asp...
Inner West	https://innerwest.infocouncil.biz/
Ku-ring-gai	https://eservices.kmc.nsw.gov.au/Infocouncil.Web/
Lane Cove	http://lccweb.lanecove.nsw.gov.au/bps/BusinessP...
Liverpool	http://liverpool.infocouncil.biz/
Mosman	https://mosman.nsw.gov.au/council/meetings/coun...
Northern Beaches	https://www.northernbeaches.nsw.gov.au/council/...
Parramatta	https://businesspapers.parracity.nsw.gov.au/Bus...
Penrith	http://bizsearch.penrithcity.nsw.gov.au/pccbps/
Randwick	http://businesspapers.randwick.nsw.gov.au/

Ryde	http://www.ryde.nsw.gov.au/Council/Council-Meet...
Strathfield	https://www.strathfield.nsw.gov.au/council/coun...
Sutherland	http://www.sutherlandshire.nsw.gov.au/Council/M...
Sydney	http://www.cityofsydney.nsw.gov.au/council/abou...
Waverley	http://waverley.infocouncil.biz/
Willoughby	http://www.willoughby.nsw.gov.au/Council-Meetin...
Woollahra	https://www.woollahra.nsw.gov.au/council/meetin...
Blacktown	https://www.blacktown.nsw.gov.au/About-Council/...
New South Wales	https://live.ipcn.nsw.gov.au/projects?year={year}

Table 3.1: Data sources including URLs

3.3 DATA STORAGE

All of the retrieved data is stored on Amazon's Simple Storage Service (AS3) cloud servers. The private server space is referred to as a 'bucket' and it can be easily integrated into the platform via Amazon's Application Programming Interface (API). Amazon's AS3 is the most widely used cloud storage service in the world.

3.4 REGULAR EXPRESSION (REGEX) PATTERNS

As noted above, each LGA website has a unique URL and structure which requires a different algorithm for each website. A typical REGEX pattern that will locate any user-generated keyword in a PDF document will be in the form:

```
'/\/viewDocument\?docid=([0-9]+)/'
```

Similarly, this pattern will locate any content on the site which is an anchor tag (<a>) and has a link of numeric value as the document ID:

```
'/(([0-9]{1,2}\/[0-9]{2}\/[0-9]{4}) (.+?) - (.+?) \[<a href="(.+?)" target="_blank">View<\/a>]<br>)/'
```

The REGEX patterns and scraping depths vary between the different data sources as illustrated in the table below.

Name	**Scrape URL**	**Regex pattern**	**Scraping Depth**
North Sydney	https://www.northsydney.nsw.gov.au/Council_Meet...	**/(.+)\.pdf/i**	0
Bayside Council	https://www.bayside.vic.gov.au/council-minutes	**/(.+)\.pdf/i**	0
Burwood	http://www.burwood.nsw.gov.au/council_meetings_...	**/(.+)\.pdf/i**	0

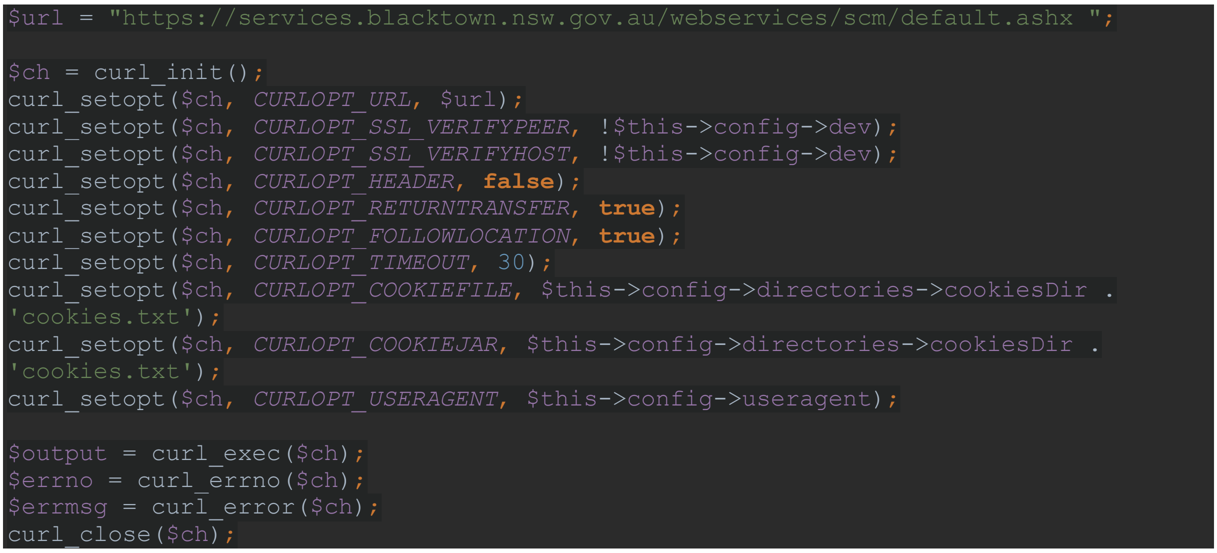
	http://www.burwood.nsw.gov.au/b_and_d_minutes_a...	/(.+)\.pdf/i	0
Camden	https://www.camden.nsw.gov.au/council/council-m...	/(.+)\.pdf/i	0
Campbelltown	https://www.campbelltown.nsw.gov.au/CouncilandC...	/(.+)\.pdf/i	0
Canada Bay	http://www.canadabay.nsw.gov.au/calendar-of-cou...	/(.+)\.pdf/i	1
Canterbury-Bankstown	https://www.cbcity.nsw.gov.au/council/Councilme...	/(.+)\.pdf/i	0
Cumberland	http://cumberland.infocouncil.biz/	/(.+)\.pdf/i	0
Fairfield	http://bpweb.fairfieldcity.nsw.gov.au:8080/fccbps/	/(.+)\.pdf/i	0
Georges River	http://infoweb.georgesriver.nsw.gov.au/grinfoco...	/(.+)\.pdf/i	0
The Hills Shire	https://www.thehills.nsw.gov.au/Council/Meeting...	/(.+)\.pdf/i	0
	https://www.thehills.nsw.gov.au/Council/Meeting...	/(.+)\.pdf/i	0
Hornsby Shire	http://businesspapers.hornsby.nsw.gov.au/	/(.+)\.pdf/i	0
Hunter's Hill	http://www.huntershill.nsw.gov.au/Page/Page.asp...	/(.+)\.pdf/i	1
Inner West	https://innerwest.infocouncil.biz/	/(.+)\.pdf/i	0
Ku-ring-gai	https://eservices.kmc.nsw.gov.au/Infocouncil.Web/	/(.+)\.pdf/i	0
Lane Cove	http://lccweb.lanecove.nsw.gov.au/bps/BusinessP...	/(.+)\.pdf/i	3
Liverpool	http://liverpool.infocouncil.biz/	/(.+)\.pdf/i	0
Mosman	https://mosman.nsw.gov.au/council/meetings/coun...	/\?ext=pdf\&id=(.+)/	0
Northern Beaches	https://www.northernbeaches.nsw.gov.au/council/...	/(.+)\.pdf/i	0
Parramatta	https://businesspapers.parracity.nsw.gov.au/Bus...	/(.+)\.pdf/i	0
Penrith	http://bizsearch.penrithcity.nsw.gov.au/pccbps/	/(.+)\.pdf/i	0
Randwick	http://businesspapers.randwick.nsw.gov.au/	/(.+)\.pdf/i	0
Ryde	http://www.ryde.nsw.gov.au/Council/Council-Meet...	/(.+)(agenda|minutes)\.pdf/i	1
Strathfield	https://www.strathfield.nsw.gov.au/council/coun...	/(.+)\.pdf/i	1
Sutherland	http://www.sutherlandshire.nsw.gov.au/Council/M...	/(.+)\.pdf/i	0
Sydney	http://www.cityofsydney.nsw.gov.au/council/abou...	/(.+)\.pdf/i	1
Waverley	http://waverley.infocouncil.biz/	/(.+)\.pdf/i	0
Willoughby	http://www.willoughby.nsw.gov.au/Council-Meetin...	/(.+)\/DocumentViewer\.ashx\?dsi=([0-9]+)/	0
Woollahra	https://www.woollahra.nsw.gov.au/council/meetin...	/(.+)\.pdf/i	0
Blacktown	https://www.blacktown.nsw.gov.au/About-Council/...	/(.+)\.pdf/i	1
New South Wales	https://live.ipcn.nsw.gov.au/projects?year={year}	/(.+)\.pdf/i	1

Table 3.2: Data source URLs and REGEX patterns

3.5 THE WEB-SCRAPING PROCESS

The web-scraping process that the *Aiden* platform follows is summarized below. As each LGA has a different Hypertext Markup Language (HTML) structure, alternate approaches are required to initiate the data command-line tool, cURL.

1. Standard cURL Initialisation

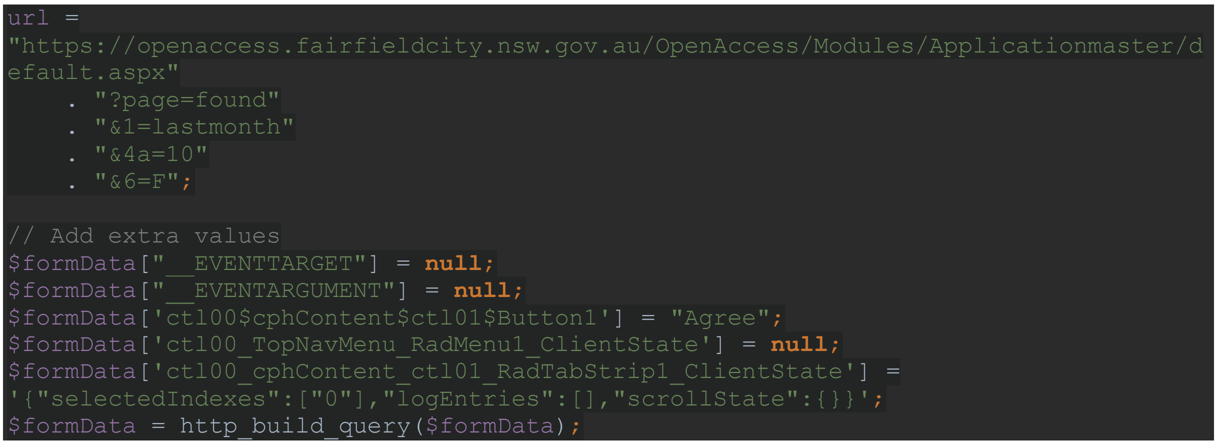


```
$url = "https://services.blacktown.nsw.gov.au/webservices/scm/default.ashx ";

$ch = curl_init();
curl_setopt($ch, CURLOPT_URL, $url);
curl_setopt($ch, CURLOPT_SSL_VERIFYPEER, !$this->config->dev);
curl_setopt($ch, CURLOPT_SSL_VERIFYHOST, !$this->config->dev);
curl_setopt($ch, CURLOPT_HEADER, false);
curl_setopt($ch, CURLOPT_RETURNTRANSFER, true);
curl_setopt($ch, CURLOPT_FOLLOWLOCATION, true);
curl_setopt($ch, CURLOPT_TIMEOUT, 30);
curl_setopt($ch, CURLOPT_COOKIEFILE, $this->config->directories->cookiesDir .
'cookies.txt');
curl_setopt($ch, CURLOPT_COOKIEJAR, $this->config->directories->cookiesDir .
'cookies.txt');
curl_setopt($ch, CURLOPT_USERAGENT, $this->config->useragent);

$output = curl_exec($ch);
$errno = curl_errno($ch);
$errmsg = curl_error($ch);
curl_close($ch);
```

The code above will visit the URL *'https://services.blacktown.nsw.gov.au/webservices/scm/default.ashx'* and will retrieve the entire HTML structure from the page and stored into the output variables.

2. Accept website Terms and Conditions before redirecting to the source data page

The majority of the sources have a terms and conditions page that needs to be reviewed and accepted. The algorithm has been designed to accept the terms and conditions in order to navigate to the data source.



```
url =
"https://openaccess.fairfieldcity.nsw.gov.au/OpenAccess/Modules/Applicationmaster/d
efault.aspx"
    . "?page=found"
    . "&1=lastmonth"
    . "&4a=10"
    . "&6=F";

// Add extra values
$formData["__EVENTTARGET"] = null;
$formData["__EVENTARGUMENT"] = null;
$formData['ctl00$cphContent$ctl01$Button1'] = "Agree";
$formData['ctl00_TopNavMenu_RadMenu1_ClientState'] = null;
$formData['ctl00_cphContent_ctl01_RadTabStrip1_ClientState'] =
'{"selectedIndexes":["0"],"logEntries":[],"scrollState":{}}';
$formData = http_build_query($formData);
```

```
$requestHeaders = [
    "Accept:
text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/apng,*/*;q=0
.8",
    "Accept-Encoding: none",
    "Content-Type: application/x-www-form-urlencoded",
    "Content-Length: " . strlen($formData),
    "Host: openaccess.fairfieldcity.nsw.gov.au",
    "Referer:
https://openaccess.fairfieldcity.nsw.gov.au/OpenAccess/Modules/Applicationmaster/de
fault.aspx?page=found&1=lastmonth&4a=10&6=F"
];

$ch = curl_init();
curl_setopt($ch, CURLOPT_URL, $url);
curl_setopt($ch, CURLOPT_POST, true);
curl_setopt($ch, CURLOPT_POSTFIELDS, $formData);
curl_setopt($ch, CURLOPT_HTTPHEADER, $requestHeaders);
curl_setopt($ch, CURLOPT_SSL_VERIFYPEER, 2);
curl_setopt($ch, CURLOPT_SSL_VERIFYHOST, 2);
curl_setopt($ch, CURLOPT_HEADER, false);
curl_setopt($ch, CURLOPT_RETURNTRANSFER, true);
curl_setopt($ch, CURLOPT_FOLLOWLOCATION, true);
curl_setopt($ch, CURLOPT_TIMEOUT, 30);
curl_setopt($ch, CURLOPT_COOKIEFILE, __DIR__ . '/../cookies/cookies.txt');
curl_setopt($ch, CURLOPT_COOKIEJAR, __DIR__ . '/../cookies/cookies.txt');
curl_setopt($ch, CURLOPT_USERAGENT, 'Mozilla/5.0 (X11; Ubuntu; Linux x86_64;
rv:60.0) Gecko/20100101 Firefox/60.0');

$output = curl_exec($ch);
$errno = curl_errno($ch);
$errmsg = curl_error($ch);

curl_close($ch);
```

The code above will notify the LGA website's server that the confirmation page has been accepted and the LGA server will return a unique text file known as a 'cookie' to use in the standard cURL procedure.

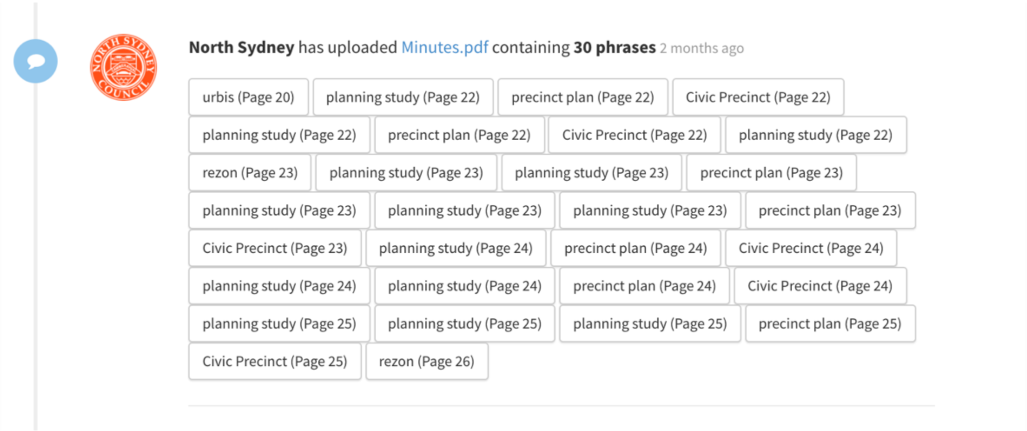
3. Step Query

Amongst the data sources, 3 of the 31 LGA websites have an intermediate page between the *Minutes* and *Agendas* page and the page containing the target PDF files. In order to navigate this, a custom step query was introduced into the algorithms that service those websites. The code snippet below will execute a process that clicks through the intermediate page in order to reach the final page and the source data.

```
$formData = $this->getAspFormDataByUrl($url);
$formData['ctl00$MainBodyContent$mContinueButton'] = "Next";
$formData['ctl00$mHeight'] = 653;
$formData['ctl00$mWidth'] = 786;

// Page gives different output not allowing us to scrape the addresses, change
option to 2 when called by scrapeMeta
if ($calledByScrapeMethod === true) {
    $formData['mDataGrid:Column0:Property'] =
'ctl00$MainBodyContent$mDataList$ctl03$mDataGrid$ctl04$ctl00';
}
else {
    $formData['mDataGrid:Column0:Property'] =
'ctl00$MainBodyContent$mDataList$ctl03$mDataGrid$ctl02$ctl00';
}

$formData['__LASTFOCUS'] = null;
$formData = http_build_query($formData);

$requestHeaders = [
    "Host: ebiz.campbelltown.nsw.gov.au",
    "Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8",
    "Accept-Language: en-GB,en;q=0.5",
    "Accept-Encoding: none",
    "Referer: https://ebiz.campbelltown.nsw.gov.au/ePathway/Production/Web/GeneralEnquiry/EnquiryLists.aspx?ModuleCode=LAP",
    "Content-Type: application/x-www-form-urlencoded",
    "Connection: keep-alive",
    "DNT: 1",
];

$ch = curl_init();
curl_setopt($ch, CURLOPT_URL, $url);
curl_setopt($ch, CURLOPT_SSL_VERIFYPEER, false);
curl_setopt($ch, CURLOPT_SSL_VERIFYHOST, false);
curl_setopt($ch, CURLOPT_HTTPHEADER, $requestHeaders);
curl_setopt($ch, CURLOPT_POST, true);
curl_setopt($ch, CURLOPT_POSTFIELDS, $formData);
curl_setopt($ch, CURLOPT_HEADER, false);
curl_setopt($ch, CURLOPT_RETURNTRANSFER, true);
curl_setopt($ch, CURLOPT_FOLLOWLOCATION, true);
curl_setopt($ch, CURLOPT_TIMEOUT, 30);
curl_setopt($ch, CURLOPT_COOKIEFILE, __DIR__ . '/../cookies/cookies.txt');
curl_setopt($ch, CURLOPT_COOKIEJAR, __DIR__ . '/../cookies/cookies.txt');
curl_setopt($ch, CURLOPT_USERAGENT, 'Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:60.0) Gecko/20100101 Firefox/60.0');

$output = curl_exec($ch);
$errno = curl_errno($ch);
$errmsg = curl_error($ch);
curl_close($ch);
```

4. Scan the HTML document

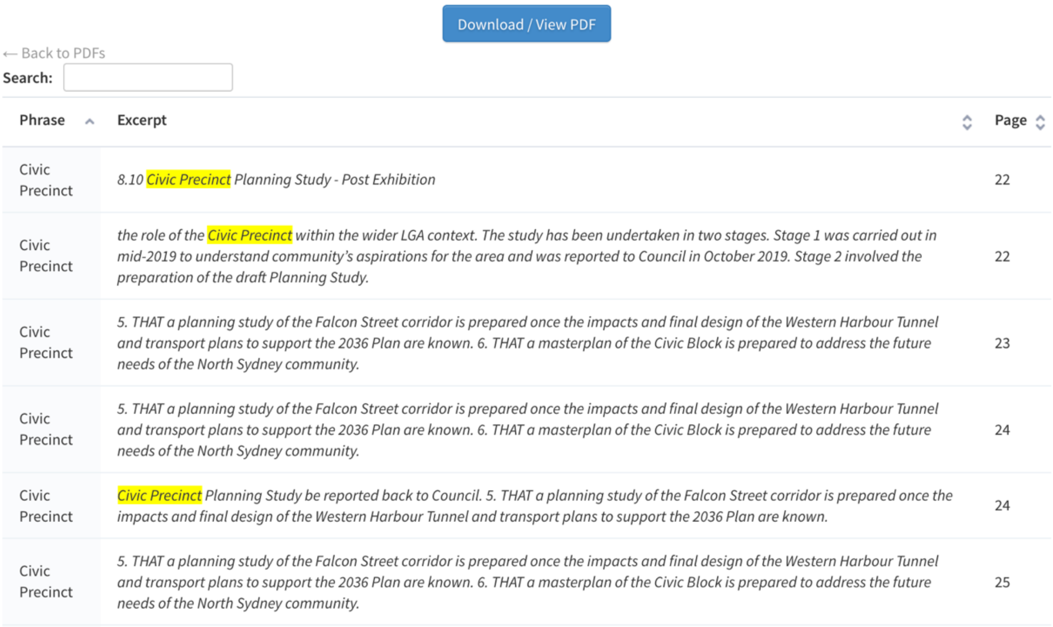
This step will scan the page and gather the data where the keywords have been identified. Each council has a different HTML structure and different REGEX patterns are being used for each website. Once the target data is extracted, it is then stored and the user is notified daily via email of any occurrences. The user can also access the data at any time in the future via the platform. The keywords are highlighted in the text as illustrated below.



North Sydney has uploaded Minutes.pdf containing **30 phrases** 2 months ago

urbis (Page 20) | planning study (Page 22) | precinct plan (Page 22) | Civic Precinct (Page 22) | planning study (Page 22) | precinct plan (Page 22) | Civic Precinct (Page 22) | planning study (Page 22) | rezon (Page 23) | planning study (Page 23) | planning study (Page 23) | precinct plan (Page 23) | planning study (Page 23) | planning study (Page 23) | planning study (Page 23) | precinct plan (Page 23) | Civic Precinct (Page 23) | planning study (Page 24) | precinct plan (Page 24) | Civic Precinct (Page 24) | planning study (Page 24) | planning study (Page 24) | precinct plan (Page 24) | Civic Precinct (Page 24) | planning study (Page 25) | planning study (Page 25) | planning study (Page 25) | precinct plan (Page 25) | Civic Precinct (Page 25) | rezon (Page 26)

Figure 3.1: Newsfeed sample with keywords



Phrase	Excerpt	Page
Civic Precinct	*8.10 Civic Precinct Planning Study - Post Exhibition*	22
Civic Precinct	*the role of the Civic Precinct within the wider LGA context. The study has been undertaken in two stages. Stage 1 was carried out in mid-2019 to understand community's aspirations for the area and was reported to Council in October 2019. Stage 2 involved the preparation of the draft Planning Study.*	22
Civic Precinct	*5. THAT a planning study of the Falcon Street corridor is prepared once the impacts and final design of the Western Harbour Tunnel and transport plans to support the 2036 Plan are known. 6. THAT a masterplan of the Civic Block is prepared to address the future needs of the North Sydney community.*	23
Civic Precinct	*5. THAT a planning study of the Falcon Street corridor is prepared once the impacts and final design of the Western Harbour Tunnel and transport plans to support the 2036 Plan are known. 6. THAT a masterplan of the Civic Block is prepared to address the future needs of the North Sydney community.*	24
Civic Precinct	*Civic Precinct Planning Study be reported back to Council. 5. THAT a planning study of the Falcon Street corridor is prepared once the impacts and final design of the Western Harbour Tunnel and transport plans to support the 2036 Plan are known.*	24
Civic Precinct	*5. THAT a planning study of the Falcon Street corridor is prepared once the impacts and final design of the Western Harbour Tunnel and transport plans to support the 2036 Plan are known. 6. THAT a masterplan of the Civic Block is prepared to address the future needs of the North Sydney community.*	25

Figure 3.2: Retrieved PDF sample with keywords highlighted

CHAPTER 4: THE AIDEN PLATFORM

4.1 INTRODUCTION

This chapter provides an overview of the key technologies and modules that were used to build the *Aiden* platform. Furthermore, the layout, structure and User Interface (UI) of the platform is discussed.

4.2 TECHNOLOGY STACK

The technology stack used to build the *Aiden* platform unifies numerous elements, plugins and code modules. These technologies are summarized below, however, a detailed analysis of how they interact is beyond the scope of this research. The technologies used are grouped into front-end (user facing), back-end and framework. In addition, a flowchart describing the platform is provided along with sample images of the UI.

4.2.1 FRONT-END

Languages:

- HTML5
- CSS3
- Javascript
- jQuery

Plugins

- Bootstrap v4.0.0
- jquery.data Tables Version: 1.9.4
- moment.js Version : 2.5.1
- jquery-timeago.js
  bootstrap-datepicker.js
- bootstrap date-picker (provides a flexible date-picker widget in the Bootstrap style)
- raphael,js
- jquery.flot.js

4.2.2 BACK-END

Languages:

- PHP
- MySQL

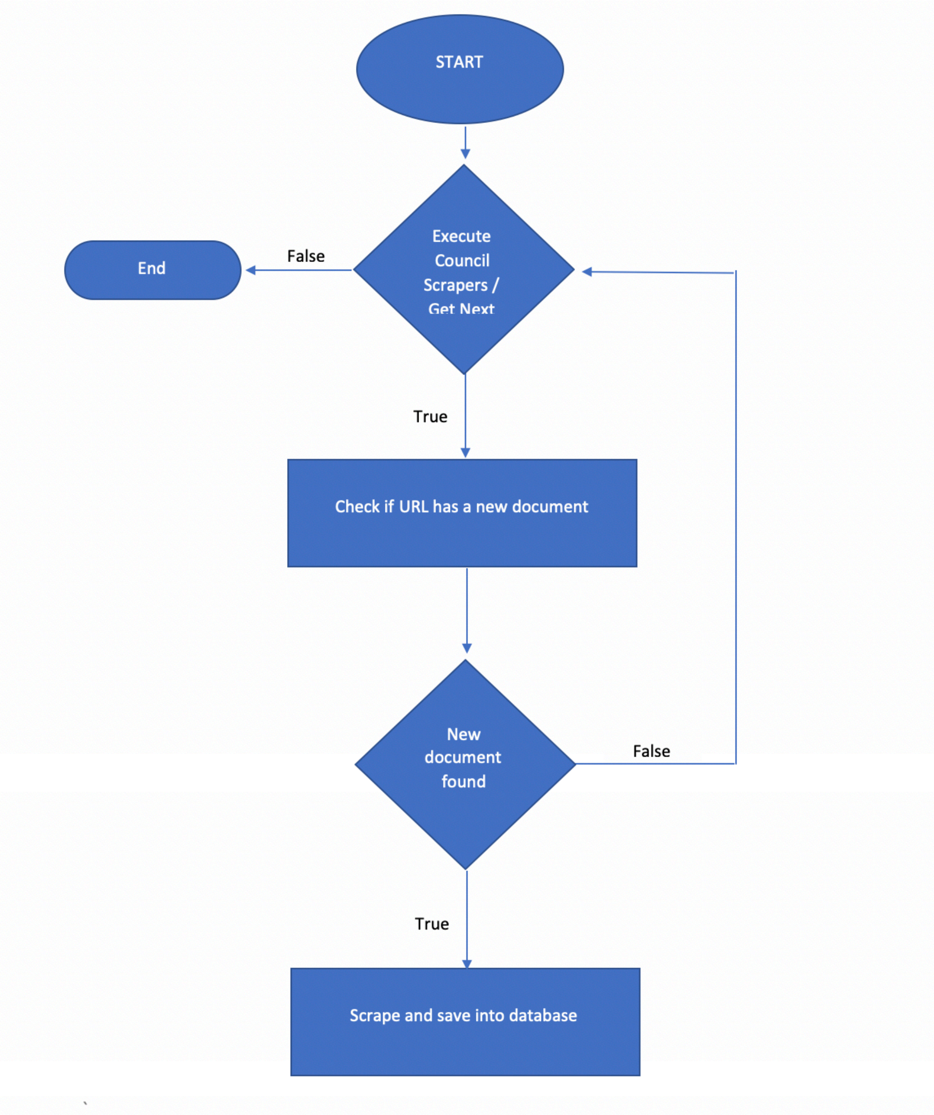
Plugins:

- cURL

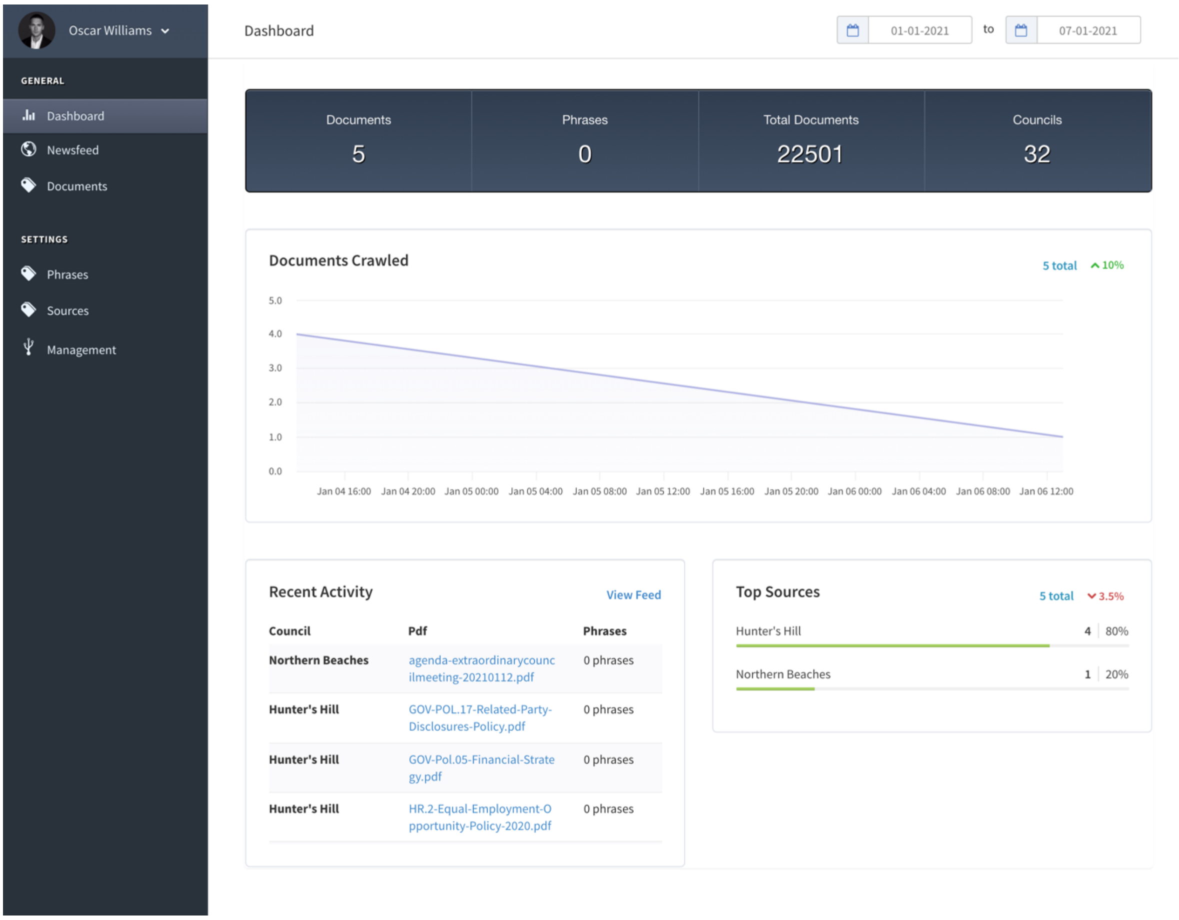
4.2.3 FRAMEWORK

- Phalcon

4.3 FLOWCHART


START
Execute Council Scrapers / Get Next
False
End
True
Check if URL has a new document
New document found
False
True
Scrape and save into database

4.4 USER INTERFACE (UI)



Documents	Phrases	Total Documents	Councils
5	0	22501	32

Documents Crawled

5 total ^10%

Recent Activity

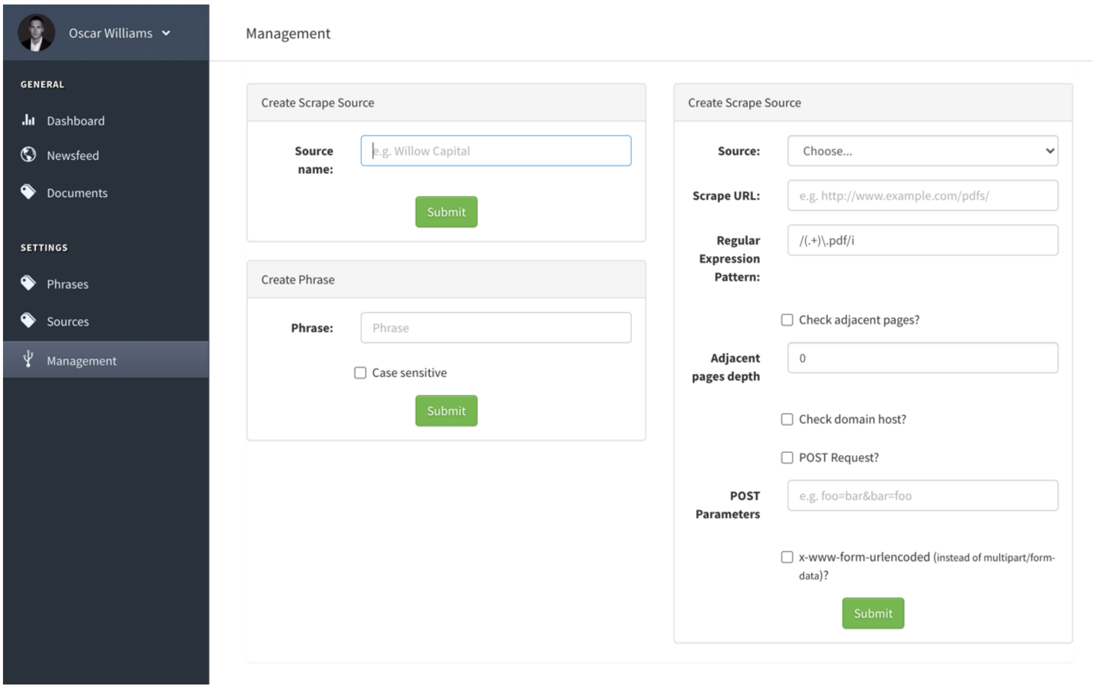
Council	Pdf	Phrases
Northern Beaches	agenda-extraordinarycouncilmeeting-20210112.pdf	0 phrases
Hunter's Hill	GOV-POL.17-Related-Party-Disclosures-Policy.pdf	0 phrases
Hunter's Hill	GOV-Pol.05-Financial-Strategy.pdf	0 phrases
Hunter's Hill	HR.2-Equal-Employment-Opportunity-Policy-2020.pdf	0 phrases

Top Sources

5 total ⌄3.5%

Hunter's Hill	4	80%
Northern Beaches	1	20%

Figure 4.1: Dashboard UI sample



Create Scrape Source

Source name: e.g. Willow Capital

Create Phrase

Phrase: Phrase

□ Case sensitive

Create Scrape Source

Source: Choose...

Scrape URL: e.g. http://www.example.com/pdfs/

Regular Expression Pattern: /(.+)\.pdf/i

□ Check adjacent pages?

Adjacent pages depth: 0

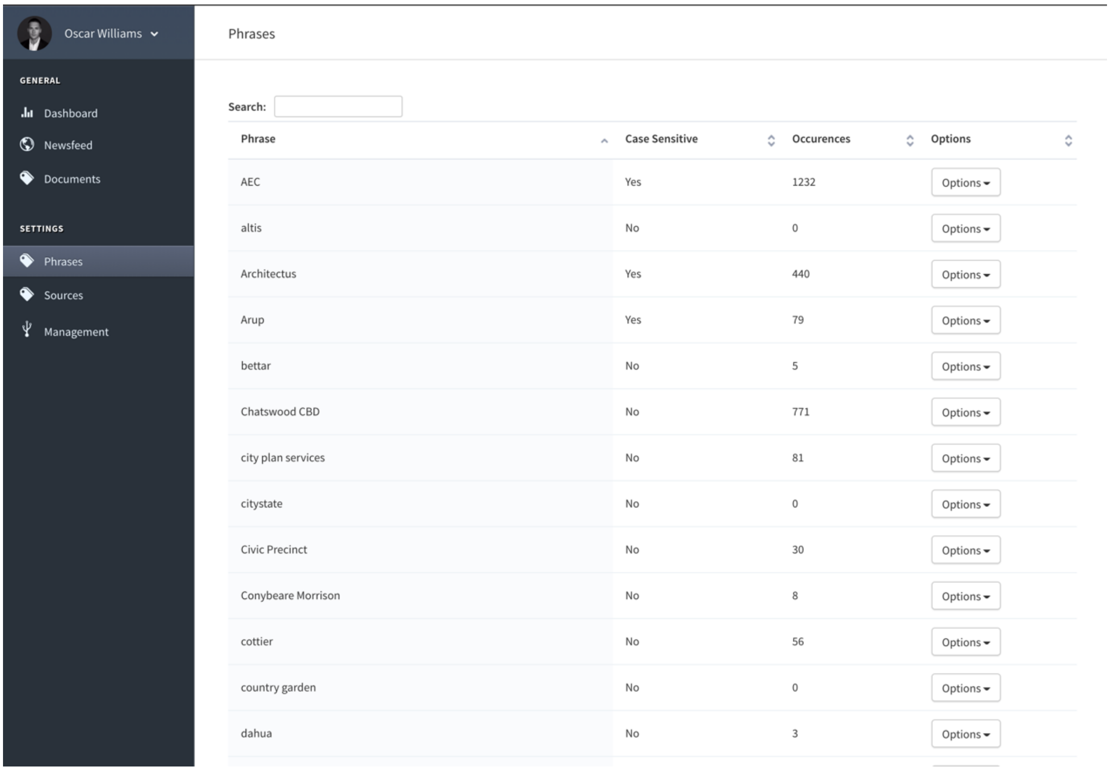
□ Check domain host?

□ POST Request?

POST Parameters: e.g. foo=bar&bar=foo

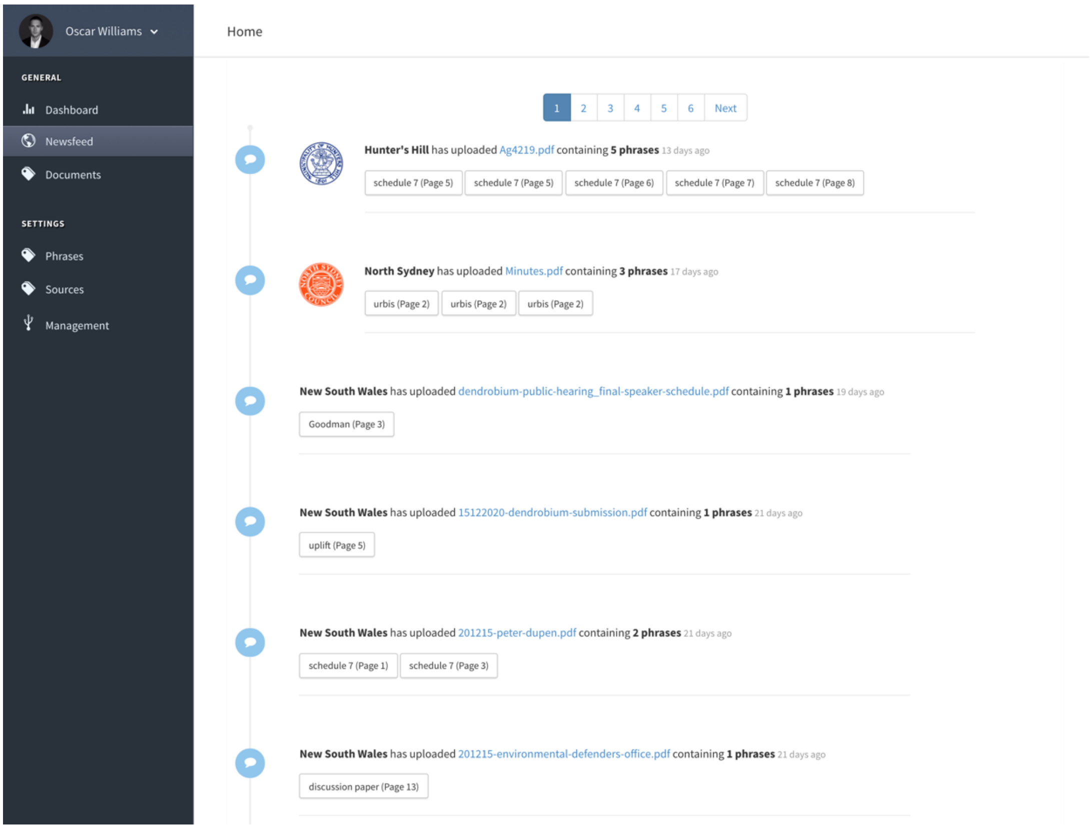
□ x-www-form-urlencoded (instead of multipart/form-data)?

Figure 4.2: Management page sample where keywords can be added



Phrase	Case Sensitive	Occurences	Options
AEC	Yes	1232	Options
altis	No	0	Options
Architectus	Yes	440	Options
Arup	Yes	79	Options
bettar	No	5	Options
Chatswood CBD	No	771	Options
city plan services	No	81	Options
citystate	No	0	Options
Civic Precinct	No	30	Options
Conybeare Morrison	No	8	Options
cottier	No	56	Options
country garden	No	0	Options
dahua	No	3	Options

Figure 4.3: Phrases sample where keywords are listed



Hunter's Hill has uploaded Ag4219.pdf containing **5 phrases** 13 days ago

schedule 7 (Page 5) | schedule 7 (Page 5) | schedule 7 (Page 6) | schedule 7 (Page 7) | schedule 7 (Page 8)

North Sydney has uploaded Minutes.pdf containing **3 phrases** 17 days ago

urbis (Page 2) | urbis (Page 2) | urbis (Page 2)

New South Wales has uploaded dendrobium-public-hearing_final-speaker-schedule.pdf containing **1 phrases** 19 days ago

Goodman (Page 3)

New South Wales has uploaded 15122020-dendrobium-submission.pdf containing **1 phrases** 21 days ago

uplift (Page 5)

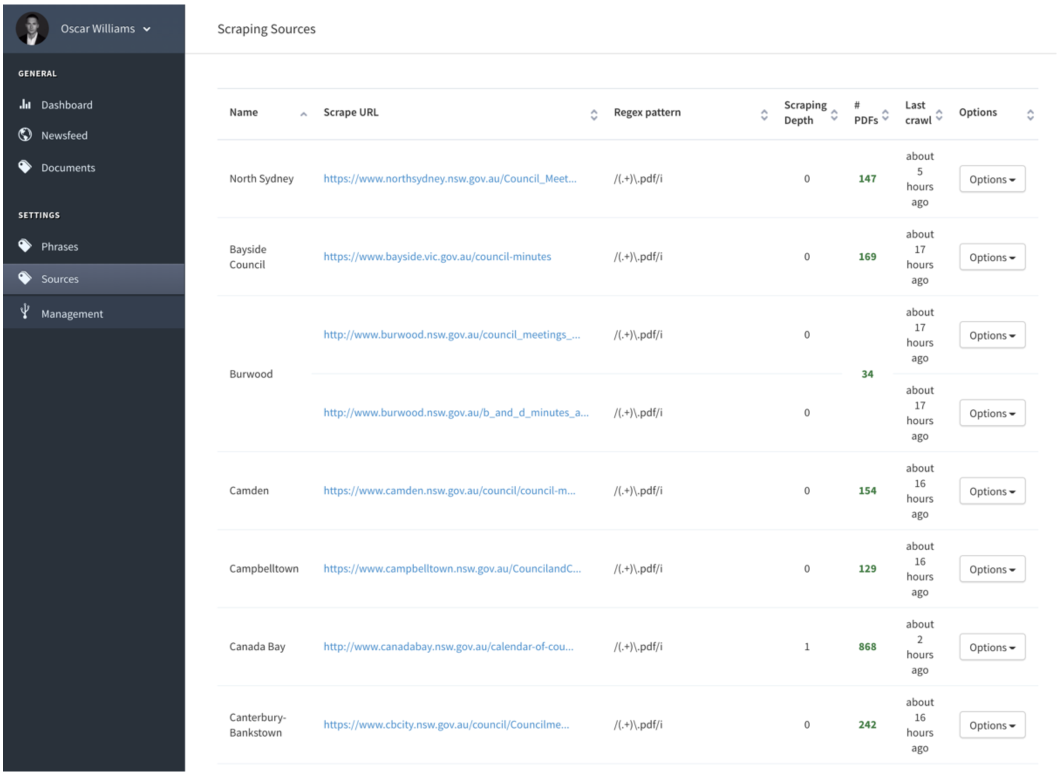
New South Wales has uploaded 201215-peter-dupen.pdf containing **2 phrases** 21 days ago

schedule 7 (Page 1) | schedule 7 (Page 3)

New South Wales has uploaded 201215-environmental-defenders-office.pdf containing **1 phrases** 21 days ago

discussion paper (Page 13)

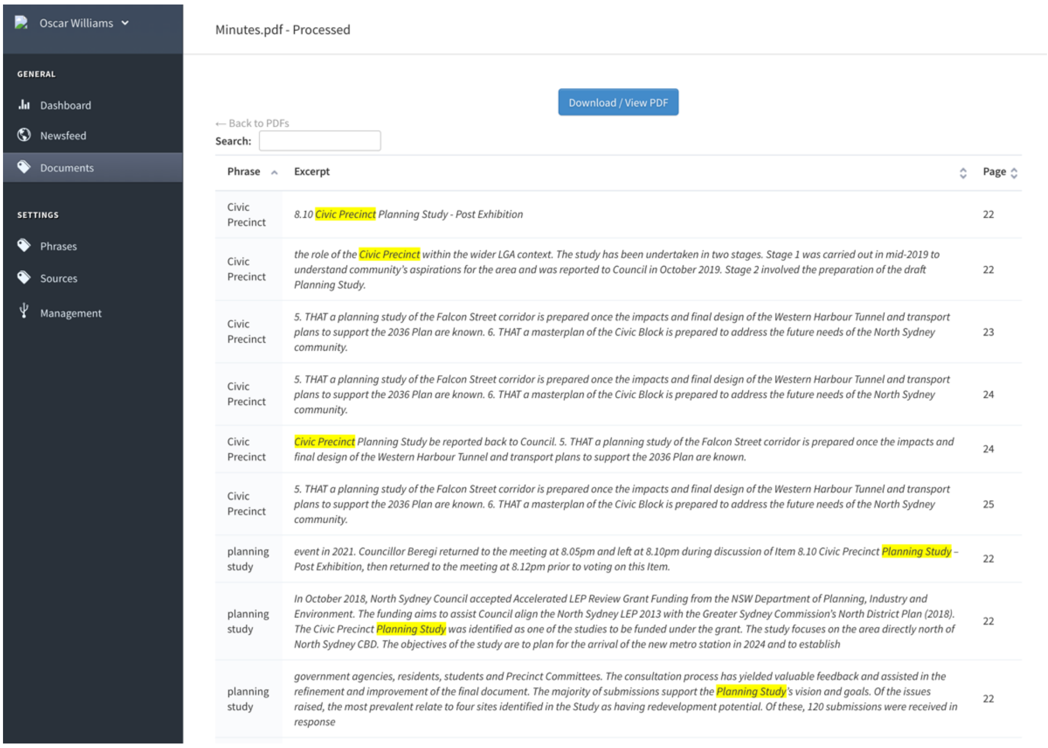
Figure 4.4: Newsfeed sample where retrieved data is displayed



Scraping Sources

Name	Scrape URL	Regex pattern	Scraping Depth	# PDFs	Last crawl	Options
North Sydney	https://www.northsydney.nsw.gov.au/Council_Meet...	/(.+)\.pdf/i	0	147	about 5 hours ago	Options ▾
Bayside Council	https://www.bayside.vic.gov.au/council-minutes	/(.+)\.pdf/i	0	169	about 17 hours ago	Options ▾
Burwood	http://www.burwood.nsw.gov.au/council_meetings_...	/(.+)\.pdf/i	0	34	about 17 hours ago	Options ▾
	http://www.burwood.nsw.gov.au/b_and_d_minutes_a...	/(.+)\.pdf/i	0		about 17 hours ago	Options ▾
Camden	https://www.camden.nsw.gov.au/council/council-m...	/(.+)\.pdf/i	0	154	about 16 hours ago	Options ▾
Campbelltown	https://www.campbelltown.nsw.gov.au/CouncilandC...	/(.+)\.pdf/i	0	129	about 16 hours ago	Options ▾
Canada Bay	http://www.canadabay.nsw.gov.au/calendar-of-cou...	/(.+)\.pdf/i	1	868	about 2 hours ago	Options ▾
Canterbury-Bankstown	https://www.cbcity.nsw.gov.au/council/Councilme...	/(.+)\.pdf/i	0	242	about 16 hours ago	Options ▾

Figure 4.5: Sources page sample listing source URLs

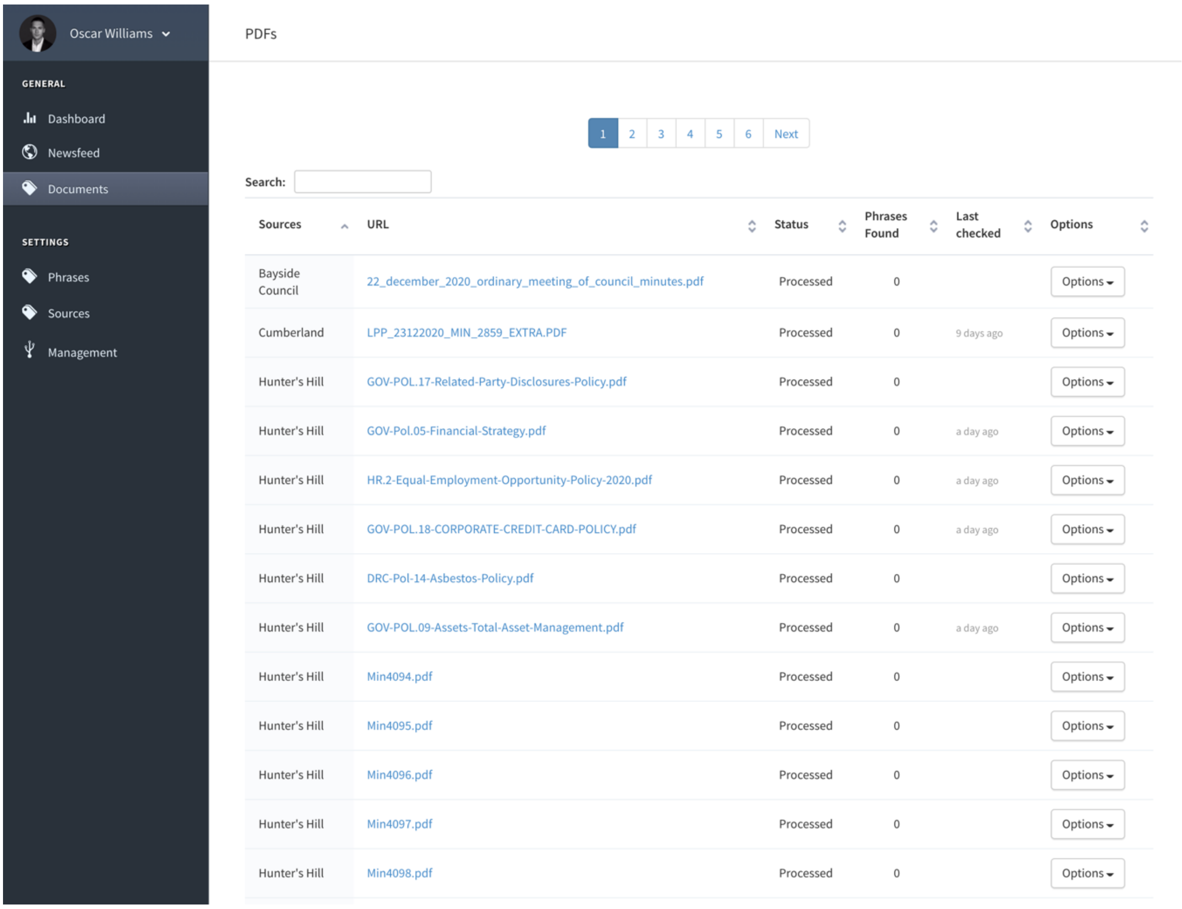


Minutes.pdf - Processed

Search:

Phrase	Excerpt	Page
Civic Precinct	*8.10 Civic Precinct Planning Study - Post Exhibition*	22
Civic Precinct	*the role of the Civic Precinct within the wider LGA context. The study has been undertaken in two stages. Stage 1 was carried out in mid-2019 to understand community's aspirations for the area and was reported to Council in October 2019. Stage 2 involved the preparation of the draft Planning Study.*	22
Civic Precinct	*5. THAT a planning study of the Falcon Street corridor is prepared once the impacts and final design of the Western Harbour Tunnel and transport plans to support the 2036 Plan are known. 6. THAT a masterplan of the Civic Block is prepared to address the future needs of the North Sydney community.*	23
Civic Precinct	*5. THAT a planning study of the Falcon Street corridor is prepared once the impacts and final design of the Western Harbour Tunnel and transport plans to support the 2036 Plan are known. 6. THAT a masterplan of the Civic Block is prepared to address the future needs of the North Sydney community.*	24
Civic Precinct	*Civic Precinct Planning Study be reported back to Council. 5. THAT a planning study of the Falcon Street corridor is prepared once the impacts and final design of the Western Harbour Tunnel and transport plans to support the 2036 Plan are known.*	24
Civic Precinct	*5. THAT a planning study of the Falcon Street corridor is prepared once the impacts and final design of the Western Harbour Tunnel and transport plans to support the 2036 Plan are known. 6. THAT a masterplan of the Civic Block is prepared to address the future needs of the North Sydney community.*	25
planning study	*event in 2021. Councillor Beregi returned to the meeting at 8.05pm and left at 8.10pm during discussion of Item 8.10 Civic Precinct Planning Study – Post Exhibition, then returned to the meeting at 8.12pm prior to voting on this Item.*	22
planning study	*In October 2018, North Sydney Council accepted Accelerated LEP Review Grant Funding from the NSW Department of Planning, Industry and Environment. The funding aims to assist Council align the North Sydney LEP 2013 with the Greater Sydney Commission's North District Plan (2018). The Civic Precinct Planning Study was identified as one of the studies to be funded under the grant. The study focuses on the area directly north of North Sydney CBD. The objectives of the study are to plan for the arrival of the new metro station in 2024 and to establish*	22
planning study	*government agencies, residents, students and Precinct Committees. The consultation process has yielded valuable feedback and assisted in the refinement and improvement of the final document. The majority of submissions support the Planning Study's vision and goals. Of the issues raised, the most prevalent relate to four sites identified in the Study as having redevelopment potential. Of these, 120 submissions were received in response*	22

Figure 4.6: Documents page sample displaying keywords found within a PDF



Sources	URL	Status	Phrases Found	Last checked	Options
Bayside Council	22_december_2020_ordinary_meeting_of_council_minutes.pdf	Processed	0		Options
Cumberland	LPP_23122020_MIN_2859_EXTRA.PDF	Processed	0	9 days ago	Options
Hunter's Hill	GOV-POL.17-Related-Party-Disclosures-Policy.pdf	Processed	0		Options
Hunter's Hill	GOV-Pol.05-Financial-Strategy.pdf	Processed	0	a day ago	Options
Hunter's Hill	HR.2-Equal-Employment-Opportunity-Policy-2020.pdf	Processed	0	a day ago	Options
Hunter's Hill	GOV-POL.18-CORPORATE-CREDIT-CARD-POLICY.pdf	Processed	0	a day ago	Options
Hunter's Hill	DRC-Pol-14-Asbestos-Policy.pdf	Processed	0		Options
Hunter's Hill	GOV-POL.09-Assets-Total-Asset-Management.pdf	Processed	0	a day ago	Options
Hunter's Hill	Min4094.pdf	Processed	0		Options
Hunter's Hill	Min4095.pdf	Processed	0		Options
Hunter's Hill	Min4096.pdf	Processed	0		Options
Hunter's Hill	Min4097.pdf	Processed	0		Options
Hunter's Hill	Min4098.pdf	Processed	0		Options

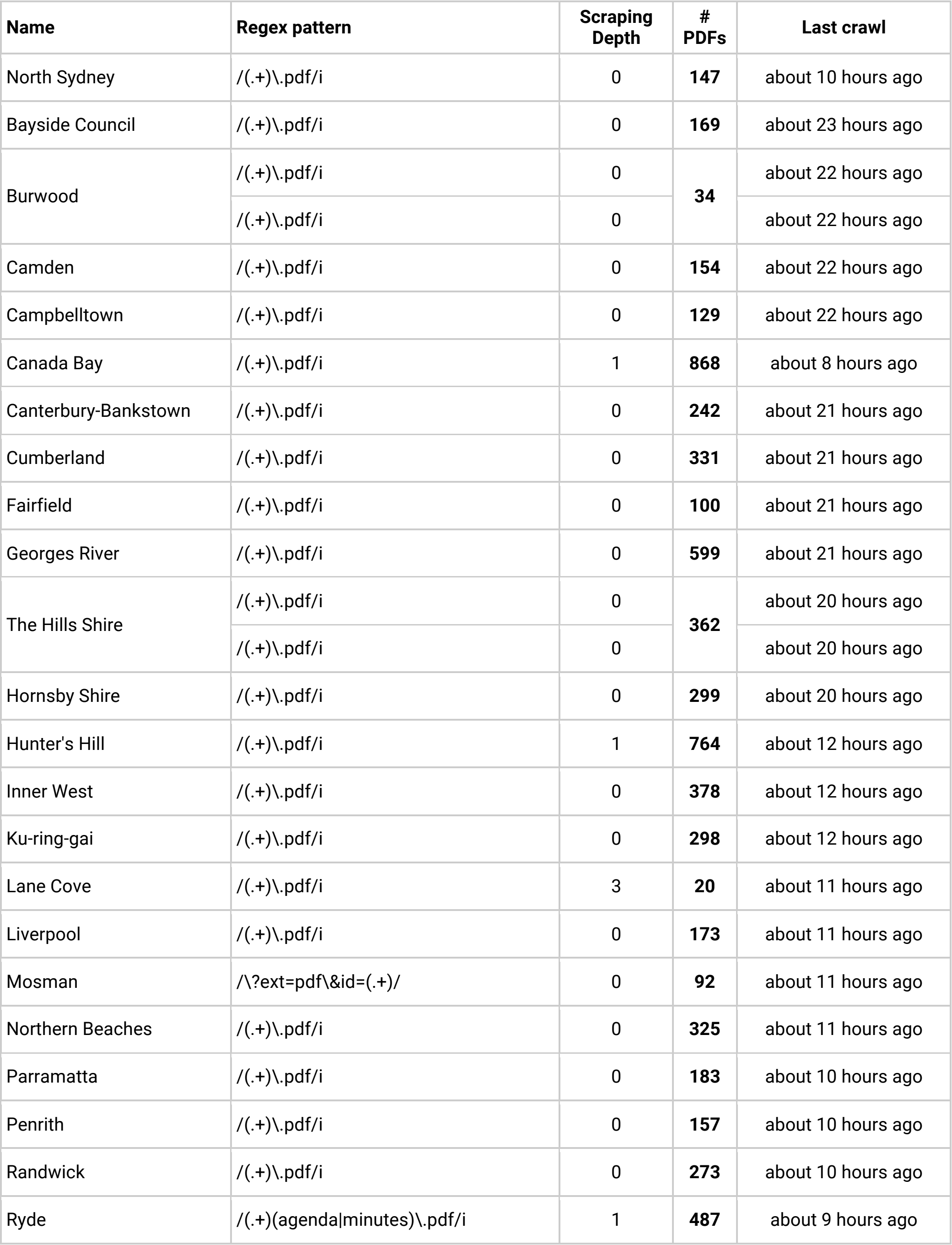
Figure 4.7: Documents page sample displaying retrieved PDFs

CHAPTER 5: OUTPUT & FINDINGS

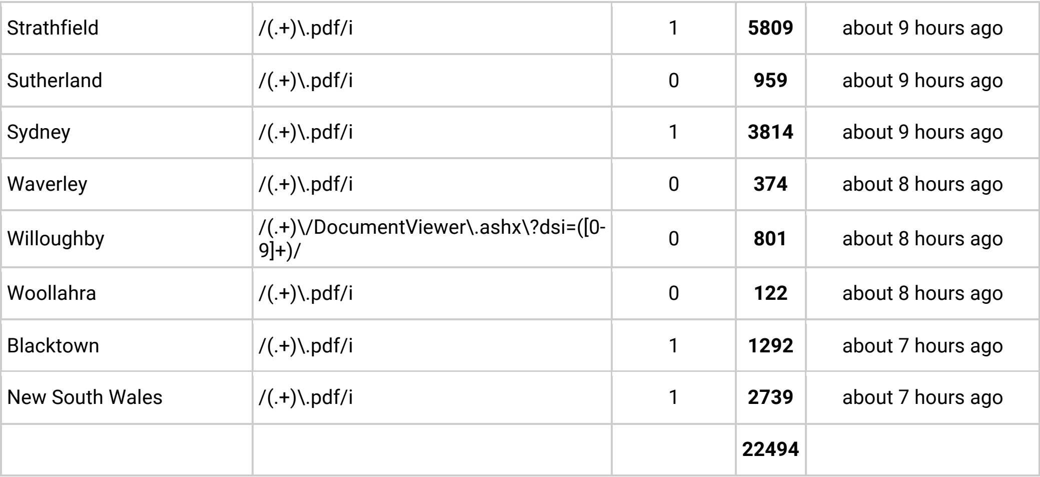
5.1 INTRODUCTION

Chapter 5 presents a discussion of the findings and the output of the algorithms. The algorithms searched 31 LGAs and parsed 22,548 PDF documents with date ranges between January 1st 2018 and December 31st 2020. Within these 22,548 documents, keywords were detected 10,014 times. The output summary of the document searches and occurrences is provided below.

5.2 FINDINGS & COLLECTED DATA



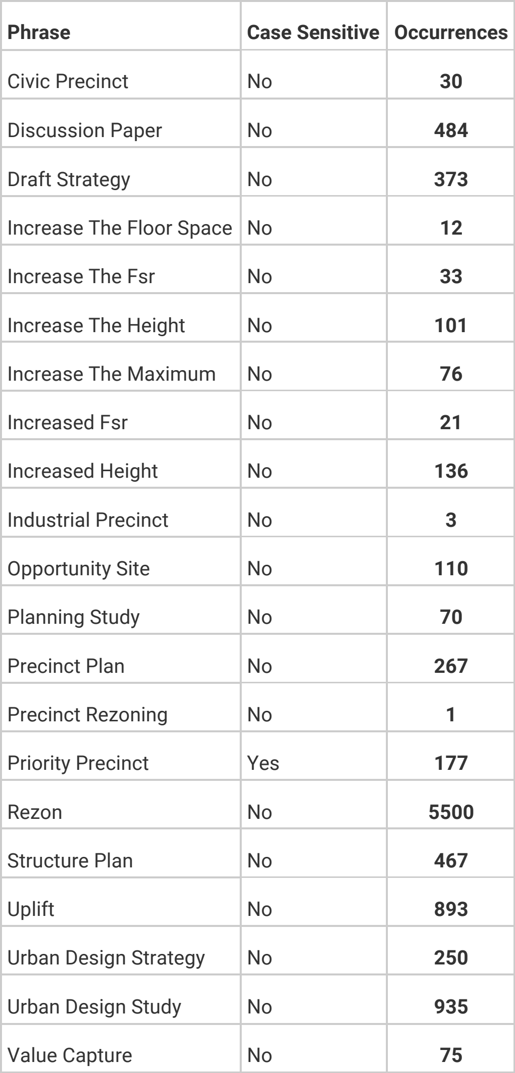
Name	Regex pattern	Scraping Depth	# PDFs	Last crawl
North Sydney	/(.+)\.pdf/i	0	**147**	about 10 hours ago
Bayside Council	/(.+)\.pdf/i	0	**169**	about 23 hours ago
Burwood	/(.+)\.pdf/i	0	**34**	about 22 hours ago
	/(.+)\.pdf/i	0		about 22 hours ago
Camden	/(.+)\.pdf/i	0	**154**	about 22 hours ago
Campbelltown	/(.+)\.pdf/i	0	**129**	about 22 hours ago
Canada Bay	/(.+)\.pdf/i	1	**868**	about 8 hours ago
Canterbury-Bankstown	/(.+)\.pdf/i	0	**242**	about 21 hours ago
Cumberland	/(.+)\.pdf/i	0	**331**	about 21 hours ago
Fairfield	/(.+)\.pdf/i	0	**100**	about 21 hours ago
Georges River	/(.+)\.pdf/i	0	**599**	about 21 hours ago
The Hills Shire	/(.+)\.pdf/i	0	**362**	about 20 hours ago
	/(.+)\.pdf/i	0		about 20 hours ago
Hornsby Shire	/(.+)\.pdf/i	0	**299**	about 20 hours ago
Hunter's Hill	/(.+)\.pdf/i	1	**764**	about 12 hours ago
Inner West	/(.+)\.pdf/i	0	**378**	about 12 hours ago
Ku-ring-gai	/(.+)\.pdf/i	0	**298**	about 12 hours ago
Lane Cove	/(.+)\.pdf/i	3	**20**	about 11 hours ago
Liverpool	/(.+)\.pdf/i	0	**173**	about 11 hours ago
Mosman	/\?ext=pdf\&id=(.+)/	0	**92**	about 11 hours ago
Northern Beaches	/(.+)\.pdf/i	0	**325**	about 11 hours ago
Parramatta	/(.+)\.pdf/i	0	**183**	about 10 hours ago
Penrith	/(.+)\.pdf/i	0	**157**	about 10 hours ago
Randwick	/(.+)\.pdf/i	0	**273**	about 10 hours ago
Ryde	/(.+)(agenda|minutes)\.pdf/i	1	**487**	about 9 hours ago



Strathfield	/(.+)\.pdf/i	1	**5809**	about 9 hours ago
Sutherland	/(.+)\.pdf/i	0	**959**	about 9 hours ago
Sydney	/(.+)\.pdf/i	1	**3814**	about 9 hours ago
Waverley	/(.+)\.pdf/i	0	**374**	about 8 hours ago
Willoughby	/(.+)\/DocumentViewer\.ashx\?dsi=([0-9]+)/	0	**801**	about 8 hours ago
Woollahra	/(.+)\.pdf/i	0	**122**	about 8 hours ago
Blacktown	/(.+)\.pdf/i	1	**1292**	about 7 hours ago
New South Wales	/(.+)\.pdf/i	1	**2739**	about 7 hours ago
			22494	

Table 5.1: PDFs searched by the algorithms

5.3 KEYWORD OCCURRENCES

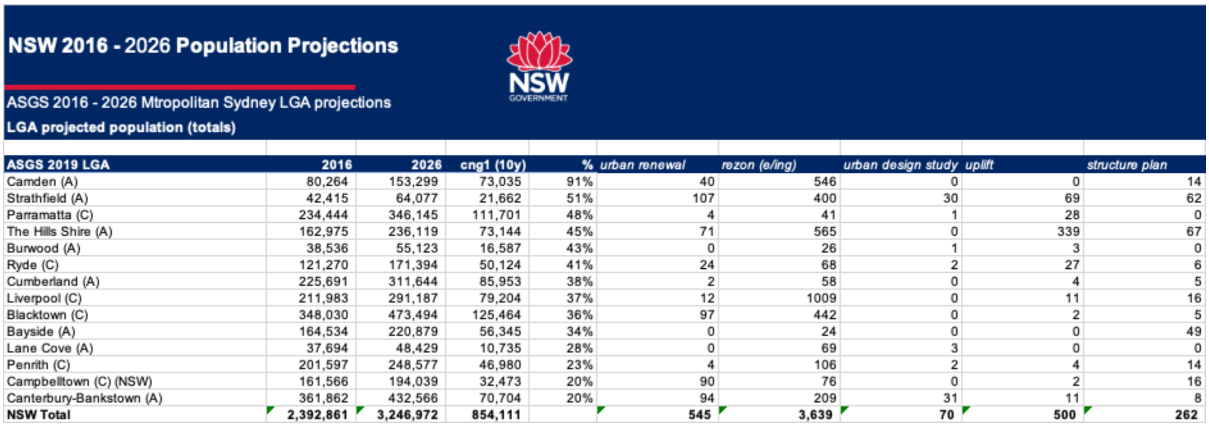


Phrase	**Case Sensitive**	**Occurrences**
Civic Precinct	No	**30**
Discussion Paper	No	**484**
Draft Strategy	No	**373**
Increase The Floor Space	No	**12**
Increase The Fsr	No	**33**
Increase The Height	No	**101**
Increase The Maximum	No	**76**
Increased Fsr	No	**21**
Increased Height	No	**136**
Industrial Precinct	No	**3**
Opportunity Site	No	**110**
Planning Study	No	**70**
Precinct Plan	No	**267**
Precinct Rezoning	No	**1**
Priority Precinct	Yes	**177**
Rezon	No	**5500**
Structure Plan	No	**467**
Uplift	No	**893**
Urban Design Strategy	No	**250**
Urban Design Study	No	**935**
Value Capture	No	**75**

Table 5.2: Occurrences of keywords

5.4 CORRELATION & SIGNIFICANCE

In order assess the significance of the findings, a regression analysis was run between five keywords and the forecast population changes for each LGA between 2016 and 2026 as published by the Australian Bureau of Statistics (ABS). The five keywords selected were “urban renewal”, “rezon(e/ing)”, “urban design study”, “uplift” and “structure plan”. These keywords were selected for two key reasons (1) they had the highest number of occurrences across all of the LGA’s; (2) they are generic and don’t relate to a specific location, developer or consultant. The rationale behind selecting generic keywords is that they would likely be more useful and widely applicable than niche keywords. Further research is required to determine which keywords are most valuable to the real estate development industry professionals. The data used to run the regression analysis is summarized below:



NSW 2016 - 2026 Population Projections

NSW GOVERNMENT

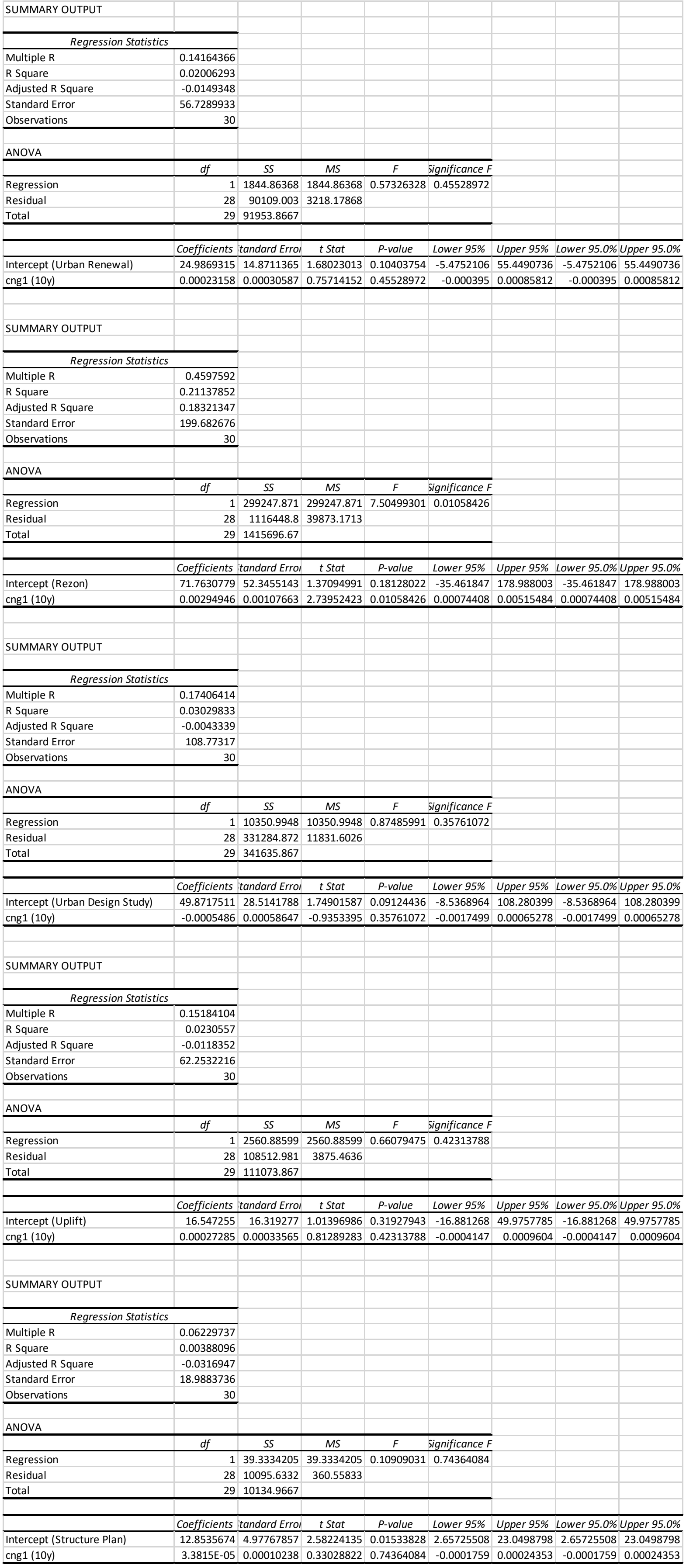
ASGS 2016 - 2026 Mtropolitan Sydney LGA projections

LGA projected population (totals)

ASGS 2019 LGA	2016	2026	cng1 (10y)	%	urban renewal	rezon (e/ing)	urban design study	uplift	structure plan
Camden (A)	80,264	153,299	73,035	91%	40	546	0	0	14
Strathfield (A)	42,415	64,077	21,662	51%	107	400	30	69	62
Parramatta (C)	234,444	346,145	111,701	48%	4	41	1	28	0
The Hills Shire (A)	162,975	236,119	73,144	45%	71	565	0	339	67
Burwood (A)	38,536	55,123	16,587	43%	0	26	1	3	0
Ryde (C)	121,270	171,394	50,124	41%	24	68	2	27	6
Cumberland (A)	225,691	311,644	85,953	38%	2	58	0	4	5
Liverpool (C)	211,983	291,187	79,204	37%	12	1009	0	11	16
Blacktown (C)	348,030	473,494	125,464	36%	97	442	0	2	5
Bayside (A)	164,534	220,879	56,345	34%	0	24	0	0	49
Lane Cove (A)	37,694	48,429	10,735	28%	0	69	3	0	0
Penrith (C)	201,597	248,577	46,980	23%	4	106	2	4	14
Campbelltown (C) (NSW)	161,566	194,039	32,473	20%	90	76	0	2	16
Canterbury-Bankstown (A)	361,862	432,566	70,704	20%	94	209	31	11	8
NSW Total	**2,392,861**	**3,246,972**	**854,111**		**545**	**3,639**	**70**	**500**	**262**

Figure 8: NSW 2016 - 2026 Population Projection & Aiden Data

The regression analysis was designed to determine the correlation between the prevalence of keywords and the forecast population change in each LGA. The hypothesis being testing is that there should be a correlation between the prevalence of keywords that indicate a rezoning and the LGA’s that are forecast to experience the largest increase in net population over the course of ten years. Firstly, the population data from the ABS has to be organized and summarized. The experiment adopts the numerical change in population over a ten-year period from 2016 to 2026. This timeframe was selected as it provides a meaningful variation between the population forecasts whilst not being so long as to lose credibly due to the uncertainly of population forecasting. The population change was selected as the independent x-variable as the zoning changes in NSW are determined by mandates provided by State Government to Local Government (LGA’s) that, in turn, are based on the population forecasts used in this experiment. The regression analysis identified a weak to moderation correlation between the keywords and the population forecasts in the LGA’s, with the strongest correlation being for the keyword “rezon(e/ing)” at 0.46. The other keywords produced the following correlation coefficients: “urban renewal” (0.14), “urban design study” (0.17), “uplift” (0.15) and “structure plan” (0.06).



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.14164366
R Square	0.02006293
Adjusted R Square	-0.0149348
Standard Error	56.7289933
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1844.86368	1844.86368	0.57326328	0.45528972
Residual	28	90109.003	3218.17868		
Total	29	91953.8667			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept (Urban Renewal)	24.9869315	14.8711365	1.68023013	0.10403754	-5.4752106	55.4490736	-5.4752106	55.4490736
cng1 (10y)	0.00023158	0.00030587	0.75714152	0.45528972	-0.000395	0.00085812	-0.000395	0.00085812

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.4597592
R Square	0.21137852
Adjusted R Square	0.18321347
Standard Error	199.682676
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	299247.871	299247.871	7.50499301	0.01058426
Residual	28	1116448.8	39873.1713		
Total	29	1415696.67			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept (Rezon)	71.7630779	52.3455143	1.37094991	0.18128022	-35.461847	178.988003	-35.461847	178.988003
cng1 (10y)	0.00294946	0.00107663	2.73952423	0.01058426	0.00074408	0.00515484	0.00074408	0.00515484

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.17406414
R Square	0.03029833
Adjusted R Square	-0.0043339
Standard Error	108.77317
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	10350.9948	10350.9948	0.87485991	0.35761072
Residual	28	331284.872	11831.6026		
Total	29	341635.867			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept (Urban Design Study)	49.8717511	28.5141788	1.74901587	0.09124436	-8.5368964	108.280399	-8.5368964	108.280399
cng1 (10y)	-0.0005486	0.00058647	-0.9353395	0.35761072	-0.0017499	0.00065278	-0.0017499	0.00065278

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.15184104
R Square	0.0230557
Adjusted R Square	-0.0118352
Standard Error	62.2532216
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2560.88599	2560.88599	0.66079475	0.42313788
Residual	28	108512.981	3875.4636		
Total	29	111073.867			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept (Uplift)	16.547255	16.319277	1.01396986	0.31927943	-16.881268	49.9757785	-16.881268	49.9757785
cng1 (10y)	0.00027285	0.00033565	0.81289283	0.42313788	-0.0004147	0.0009604	-0.0004147	0.0009604

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.06229737
R Square	0.00388096
Adjusted R Square	-0.0316947
Standard Error	18.9883736
Observations	30

ANOVA

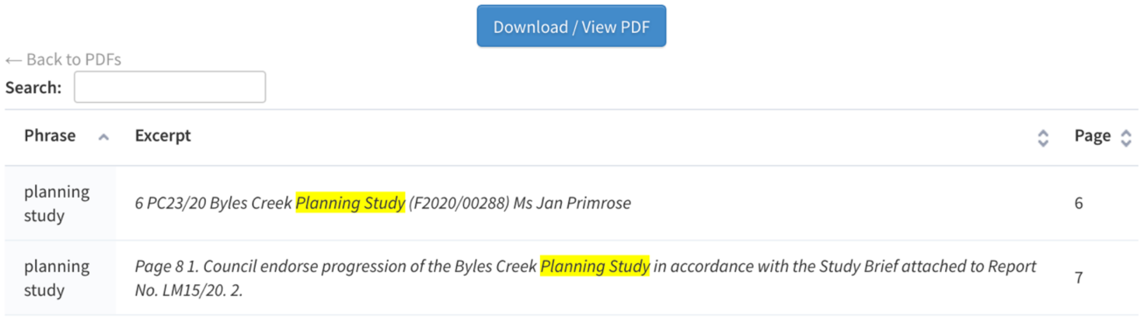
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	39.3334205	39.3334205	0.10909031	0.74364084
Residual	28	10095.6332	360.55833		
Total	29	10134.9667			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept (Structure Plan)	12.8535674	4.97767857	2.58224135	0.01533828	2.65725508	23.0498798	2.65725508	23.0498798
cng1 (10y)	3.3815E-05	0.00010238	0.33028822	0.74364084	-0.0001759	0.00024353	-0.0001759	0.00024353

Figure 9: Regression Output Data

5.5 RELEVANCE OF FINDINGS

As is evident in the findings, the *Aiden* platform is able to parse source data and identify a keyword in a PDF document. Furthermore, the algorithms allow this to be executed at scale on a daily basis in an automated manner. This enables the user to conduct work on a daily basis that would otherwise require >200 hours of manual labor. When a keyword is found, the user receives and email notifying them. In addition to identifying a given keyword, the web-scraper is designed to retrieve the sentence, paragraph, page number and a link to the entire original document. The user can continue to access the data and the original PDF file at their convenience. A sample of the user interface is provided below.



Phrase	Excerpt	Page
planning study	*6 PC23/20 Byles Creek Planning Study (F2020/00288) Ms Jan Primrose*	6
planning study	*Page 8 1. Council endorse progression of the Byles Creek Planning Study in accordance with the Study Brief attached to Report No. LM15/20. 2.*	7

Table 5.3: Sample of a keyword detected in a PDF

The figure above is also an example of information that would likely be useful to a real estate development professional who is interested in new rezoning areas that may provide investment opportunities. The keyword 'Planning Study' typically indicates that a planning investigation has taken place and if supported, a rezoning will occur. In this instance, it is clear that the Council endorses the "progression of the Byles Creek Planning Study". Information of this nature would likely be of interest and relevant to a real estate development professional and would merit further investigation.

Although the platform is highly effective in analyzing, identifying and retrieving information, there is insufficient data to evaluate the usefulness of the platform based on the keywords selected. The weak to moderate correlation between the keywords and forecast population growth is consistent with the assertion that it is difficult to evaluate the usefulness of the platform as the keywords cannot be tested in a quantitative way. Testing and analyzing the utility of each of the keywords is beyond the scope of this research however it is an important area for further investigation. Platforms like the one described in this paper are typically built 'in-house' by hedge funds, private equity groups and other private organizations. This fact that there is no scalable technology product similar to this may suggest that the keywords or other inputs are niche and highly specialized. If this is in fact true, the platform would likely be very useful to professionals who have a clear understanding of what there are looking for and of little use to those who are unable to provide specialized inputs or keywords.

CHAPTER 6: CONCLUSION & RECOMMENDATIONS

6.1 CONCLUSION

As noted at in the Introduction, the purpose of this research is to evaluate the effectiveness and usefulness of web-scraping technology for real estate development professionals. Within this narrow scope, it is clear that there is use for this technology. However, it is also clear that keyword selection is critical and the value of the information requires further investigation and research.

The development and implementation of a platform like *Aiden* is highly specialized and would require a developer to build to application based on the target data sources. Furthermore, there may be legal considerations that need to be made when accessing data using web-scraping technology.

Given that the keywords were identified over 10,000 times by the platform and there were multiple instances that suggesting that a rezoning is imminent, it is clear that the platform and underlying web-scraping technology would be of use to the real estate development industry. In order to understand the full potential of this technology, further research is required to optimize the keywords and to identify new data sources.

REFERENCES

CLAUSSEN, J. & PEUKERT, C. 2019. Obtaining Data from the Internet: A Guide to Data Crawling in Management Research. *Available at SSRN 3403799*.

CRESWELL 2009. *Research design : qualitative, quantitative, and mixed methods approaches (3rd ed.),* Thousan Oaks Calif., Sage Publications.

FAYYAD, U., PIATETSKY-SHAPIRO, G. & SMYTH, P. 1996. From data mining to knowledge discovery in databases. *AI magazine,* 17**,** 37-37.

MOHANRAM, P. S. 2020. A Brave New World: The Use of Non-traditional Information in Capital Markets. *World Scientific Book Chapters***,** 217-237.

PATEL, J. M. 2020. Introduction to Web Scraping. *Getting Structured Data from the Internet.* Springer.

ROBSON, C. 2002. *Real world research : a resource for social scientists and practitioner-researchers (2nd ed.).* Oxford U.K., Blackwell Publishers.

SANTOS, J. M. A. 2018. Real Estate Market Data Scraping and Analysis for Financial Investments.

SNELL, J. & MENALDO, N. 2016. Web scraping in an era of big data 2.0. *Bloomberg Law News*.

WINSON-GEIDEMAN, K., KRAUSE, A., LIPSCOMB, C. A. & EVANGELOPOULOS, N. 2017. *Real estate analysis in the information age: techniques for big data and statistical modeling, Routledge.*