Katherine Noack, Aaron Butler, Bipin Karki, and Mahmoud Yousefi

University of South Australia

Hansard Data Mining Installation and Deployment Guide

Contents

[1 Introduction 2](#_Toc23533347)

[1.1 Purpose 2](#_Toc23533348)

[1.2 Intended audience and reading suggestions 2](#_Toc23533349)

[1.3 Technical project stakeholders 2](#_Toc23533350)

[2 Installation Requirements 3](#_Toc23533351)

[2.1 Software Prerequisites 3](#_Toc23533352)

[2.1.1 Software 3](#_Toc23533353)

[2.1.2 Python Packages 3](#_Toc23533354)

[2.1.3 R Packages 4](#_Toc23533355)

[2.2 SSIS Package Requirements 4](#_Toc23533356)

[2.2.1 Chrome Driver 4](#_Toc23533357)

[2.2.2 Python function files 4](#_Toc23533358)

[2.2.3 Variable setup 4](#_Toc23533359)

[2.2.4 Connection Manager 5](#_Toc23533360)

[2.2.5 Code Paths 5](#_Toc23533361)

[3 Running the Project 5](#_Toc23533362)

[3.1 Configuration 5](#_Toc23533363)

[3.2 Automate Process 5](#_Toc23533364)

[4 Configuration Post-Run 6](#_Toc23533365)

[4.1 Updating Key Terms 6](#_Toc23533366)

[4.2 Updating Clients 6](#_Toc23533367)

[4.3 Dashboards 7](#_Toc23533368)

[4.3.1 Update data extract with new records from the database 7](#_Toc23533369)

[4.3.2 Update fixed axis in Top Clients chart 7](#_Toc23533370)

[4.3.3 Publish dashboards to AGD 8](#_Toc23533371)

[5 Adding Additional Proceeding Types 8](#_Toc23533372)

[Appendix A: Data Dictionary 10](#_Toc23533373)

[Appendix B: SSIS Variable Table 0](#_Toc23533374)

# Introduction

## Purpose

The purpose of this is to describe in technical terms the steps necessary to install the Hansard Data Mining Project software and make it operational.

## Intended audience and reading suggestions

This is intended to be used by technical stakeholders of the project who will be responsible for planning, performing, or maintaining the installation or deployment, such as the Systems Administrator, Analysts, or Developers.

It is intended that stakeholders and software support personnel can read this document and coordinate their efforts in the installation/deployment of the application.

## Technical project stakeholders

This section provides a list of all known technical stakeholders with an interest in the project.

| Name | E-mail address | Role |
| --- | --- | --- |
| Katherine Noack | noaky001@mymail.unisa.edu.au | Developer |
| Mahmoud Yousefi | youmy022@mymail.unisa.edu.au | Developer |
| Bipin Karki | karby008@mymail.unisa.edu.au | Developer |
| Aaron Butler | butaj01p@mymail.unisa.edu.au | Developer |

# Installation Requirements

The following prerequisites and requirements must be satisfied for the project to be setup successfully.

## Software Prerequisites

### Software

|  |  |  |
| --- | --- | --- |
| Software | Version | Description |
| Tableau Desktop | 2019.2 | Visualisation software required for changes to developed dashboards |
| Tableau Server | 2019.2 | Required to publish dashboards at AGD |
| Python | Minimum 3.7 | Required for running text analytics code and web scraper |
| R | Minimum 3.6.1 | Required for running text analytics code |
| Java | 8 | Required to run jar files when testing XSLT. Not required for running automated ETL process. |
| SQL Server |  |  |
| SQL Management Studio |  | Configuring, managing, and administering all components within Microsoft SQL Server |
| Google Chrome |  | Web Browser required by web scraper |
| Chrome Driver | Must be same version as Google Chrome | Required to run web scraper |

### Python Packages

Python packages required to run any of the code within the GitHub project are listed below. [Pip](https://packaging.python.org/tutorials/installing-packages/) can be used to install the required Python libraries using the command line:

pip install "SomeLibrary"

|  |  |
| --- | --- |
| Package | Description |
| xlrd | Extract data from Excel spreadsheets |
| numpy | Support for large, multi-dimensional arrays and matrices |
| pyodbc | ODBC database connectivity |
| pandas | Data manipulation and analysis |
| openpyocl | Read/Write Excel spreadsheets |
| gensim | Unsupervised topic modeling and natural language processing |
| nltk | Symbolic and statistical natural language processing |
| networkx | Graphs and networks |
| matplotlib | Plotting library |
| selenium | Web-based automation tool |
| bs4 | Library for pulling data out of HTML and XML files |

### R Packages

R packages required to run any of the code within the GitHub project are listed below.

|  |  |
| --- | --- |
| Package | Description |
| readxl | Read Excel Spreadsheet files |
| tidyr | Data manipulation |
| dplyr | Data manipulation |
| syuzhet | Sentiment analysis |
| lubridate | Working with dates and times |
| sentimentr | Sentiment analysis |
| ggplot2 | Data visualisation |
| tm | Text modelling |
| RODBC | ODBC database connectivity |
| openair | Analysis of air pollution data |
| scales | Functions for data visualisation |
| reshape2 | Data manipulation |
| topicmodels | Topic modelling |
| stringr | String functions |
| SnowballC | Snowball stemmers |
| plyr | Data manipulation |

## SSIS Package Requirements

### Chrome Driver

A chrome driver is required for the Python web scraper to run. This has been included in the project in the “scraper/ssis” directory. If a different chrome driver is used the path needs to be updated in the project's configuration file (config.ini). The Chrome Driver must be compatible with the version of Google Chrome installed.

### Python function files

The following Python files are required by text analytics Python code (ssis\_text\_analytics.py) and need to be stored in the same directory. This is currently located in the “analytics/ssis” directory of the GitHub project.

* document\_summary.py
* term\_search.py
* text\_summary\_statistics.py

### Variable setup

The variables below are all folder paths that need to be updated (from Variable Settings) when the SSIS package is moved to the new server.

|  |  |
| --- | --- |
| Name | Description of the variable |
| AchiveFolder | Archive folder path, which contains all the laded XMLs |
| RejectFolder | Rejected folder path, which contains all the rejected XMLs (duplicates) |
| DirBill | The source folder path for Bill XMLs before they were loaded |
| DirQA | The source folder path for Answers to Questions XMLs before they were loaded |
| DirQT | The source folder path for Question Time XMLs before they were loaded |

Table 1: SSIS Package Variable Setup

### Connection Manager

The following connection managers needed to be established before SSIS package can run correctly:

* To XSLT file which is called: HANSARD\_General.xsl. This file will come with the package
* To one XML of each source folders
* To the HANSARD database in SQL Server Management Studio (SSMS)

### Code Paths

The project path for R files will need to be updated to the current path of the GitHub project. The table below describes the files that will need to be updated.

|  |  |  |
| --- | --- | --- |
| File | Directory within project | Line number |
| AGDclients.R | analytics/ssis | 8 |
| topic\_modelling\_visualisation.R | analytics | 16 |
| topic\_modelling.R | analytics | 16 |

# Running the Project

## Configuration

The project contains a configuration file (config.ini) which will need to be updated with the appropriate directory paths and how many days the web scraper should scrape Hansard. For example, the project was initially setup to run automatically weekly and the configuration file set to scrape the last 12 days of records to ensure all new data was retrieved. Duplicate records are handled by the SSIS package and will not be added to the database.

## Automate Process

The SSIS package has been deployed to SQL Server Management Studio. As per the client’s request, the package has been scheduled to run automatically (weekly) at 12AM Saturday in SQL Server Agent. The agent job activity is called SSIS Job and from the Job Property > Schedules, SSIS Schedule can be retrieved and adjusted to update the Job Schedule Property as shown in Figure 1.

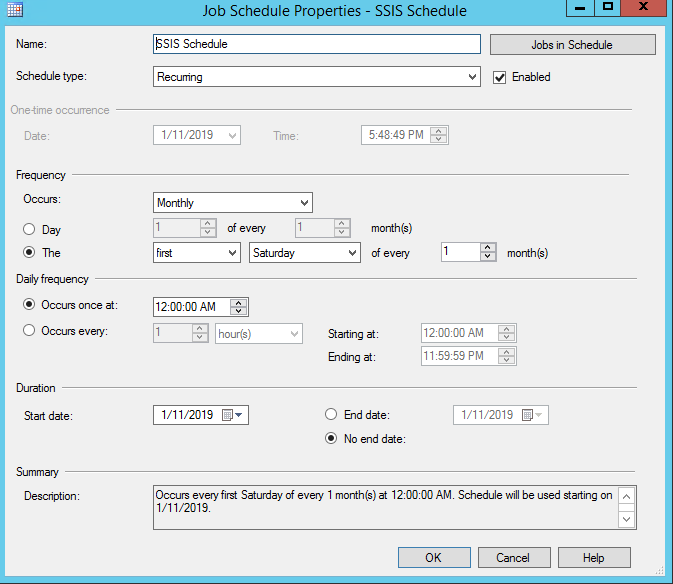


Figure 1: Job Schedule Property

# Configuration Post-Run

## Updating Key Terms

The key terms searched for in Hansard records are described in the spreadsheet AuditTeamTerms.xslx. This file is located in the data directory of the project. Each sheet in this spreadsheet is named for an audit team and new teams can be added by adding a new appropriately named sheet. This new sheet should follow the same structure of previous sheets (Term and Alternate named columns).

When this file has been updated with new key terms the associated database table needs to be recreated.

To recreate the table run the AuditTeamTerms\_Changed.bat file in the data directory. This can take several hours to run depending on the number of terms searched for and number of records in the database.

## Updating Clients

The clients searched for within record text and stored in the database can be updated by updating the Excel spreadheet “ClientNames.xlsx” located in the data directory. Each client should have its own row.

## Dashboards

When new records have been added to the database the Tableau dashboards need to be updated as this is not currently automated. Steps that need to be completed are:

### Update data extract with new records from the database

### Update fixed axis in Top Clients chart

This chart is created in the “Top Subjects” worksheet of the Tableau workbook called HansardAudit.twb. The two axes are fixed to the same number to allow easier comparison and may need to have the axis Fixed end increased to include the additional records.

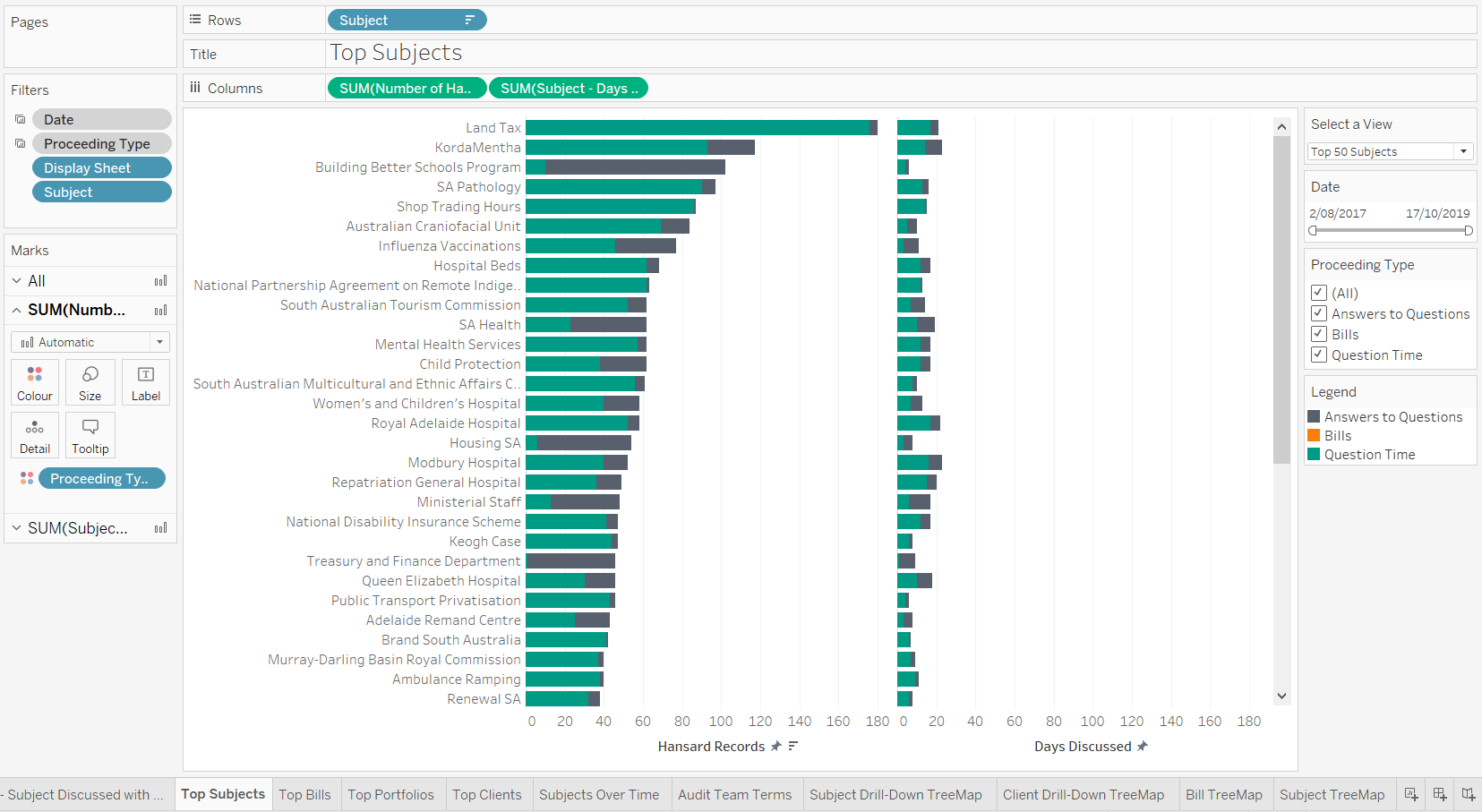


Figure 2: Top Subjects Worksheet that needs to be modified

To update the two axes (Hansard Records and Days Discussed) double click on the axis to bring up a dialog to edit the axis. This is shown in Figure 2. Both axes should have the same number for “Fixed end.”

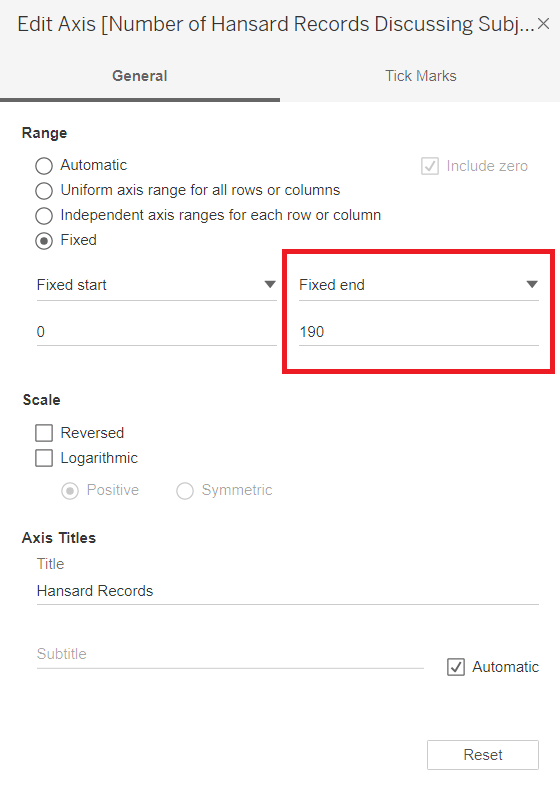


Figure 3: Update Fixed end for Axis (surrounded by red)

### Publish dashboards to AGD

# Adding Additional Proceeding Types

Currently this project only scrapes and stores Hansard records of the proceeding types: Bills, Question Time, and Answers to Questions. To add additional proceeding types to the project changes need to be made to the following files and processes:

|  |  |  |
| --- | --- | --- |
| File/Process | Location | Modification required |
| ssis\_scrape\_hansard.py | scraper/ssis directory in GitHub project | Additional calls to the scrape\_hansard function with appropriate inputs |
| config.ini | GitHub Project | Additional paths to directories that will contain the XML files scraped from the Hansard website. Each proceeding type must have its own directory. |
| SSIS changes | SSIS Package | 1. Source directory needs to be created in the same path as the others 2. Two SSIS variables need to be created: source directory and FileSourcePath 3. A foreach Loop Container (with all the sub containers) is required for each proceeding as the others (e.g. Bill). It needs to be setup and connected to the diagram before the Sequence Container for Target Tables. |

# Appendix A: Data Dictionary

|  |  |  |
| --- | --- | --- |
| Variable | Data Type | Description |
| HansardID | Integer | Unique identifier for Hansard record. Same as Hansard table id |
| Name | String | Name of Hansard record |
| Parliament Name | String | Parliament name (e.g. Parliament of South Australia) |
| Parliament Num | Integer | Parliament number |
| Review Stage | String | Review stage of record (e.g. Published) |
| Session Name | String | Session name |
| Session Num | Integer | Session number |
| Venue | String | Venue proceeding was held |
| Date | Date | Date proceeding was held |
| Date Modified | Date/Time | Date and time proceeding record was last modified |
| House | String | House (e.g. Legislative Council) |
| Proceeding Type | String | Type of proceeding (e.g. Bills, Motions, Question Time) |
| Subject | String | Subject of proceeding |

*Table 3: Final Header Table*

|  |  |  |
| --- | --- | --- |
| Variable | Data Type | Description |
| ID | Integer | Unique identifier for Hansard record |
| File Name | String | Name of XML file. Includes file extension (.xml) |
| URL | String | URL to original record on Hansard webpage |
| ProcessedOn | Date/Time | Date and time that Hansard record was processed and added to database |
| Sentiment | Float | Sentiment of Hansard record text |
| Summary | String | Three-sentence summary of Hansard record using the Page Rank algorithm |
| Key Words | String | Maximum of 10 most ranked words identified from Hansard record text |
| Record Text | String | Entire record text |

*Table 4: HANSARDFilesInfo Table*

|  |  |  |
| --- | --- | --- |
| Variable | Data Type | Description |
| Bill ID | String | Unique identifier for Bill |
| Hansard ID | Integer | Unique identifier for Hansard record |
| Bill Name | String | Name of Bill |

*Table 5: Final Bill Table*

|  |  |  |
| --- | --- | --- |
| Variable | Data Type | Description |
| Text ID | String | Unique identifier for text |
| Talker ID | String | Unique identifier for talker |
| Hansard ID | Integer | Unique identifier for Hansard record |
| Kind | String | Kind of text (e.g. speech, question, interjection, proceeding) |
| Text | String | Transcribed speech or proceeding text |
| Word Count | Integer | Number of words in the text |

*Table 6: Final Text Table*

|  |  |  |
| --- | --- | --- |
| Variable | Data Type | Description |
| Talker ID | String | Unique identifier for talker |
| House | String | House talker belongs to (e.g. Legislative Council) |
| Talker Name | String | Talkers name or position in proceeding |
| Role | String | Role of talker (e.g. office, member) |
| Electorate | String | Electorate of talker |

*Table 7: Distinct Talker Table*

|  |  |  |
| --- | --- | --- |
| Variable | Data Type | Description |
| Hansard ID | Integer | Unique identifier for Hansard record |
| Question ID | Integer | Unique identifier for question |
| Talker ID | String | Unique identifier for talker that asked the question |
| Qon Num | Integer |  |
| Question | String | Subject of question |
| Date | Date | Date question was asked |

*Table 8: Final Question Table*

|  |  |  |
| --- | --- | --- |
| Variable | Data Type | Description |
| Portfolio ID | Integer | Unique identifier for portfolio |
| Hansard ID | Integer | Unique identifier for Hansard record |
| Talker ID | String | Unique identifier for talker |
| Name | String | Name of portfolio (e.g. Minister for Health and Wellbeing) |

*Table 9: Final Portfolio Table*

|  |  |  |
| --- | --- | --- |
| Variable | Data Type | Description |
| Hansard ID | Integer | Unique identifier for Hansard record |
| Text ID | String | Unique identifier for text |
| Term | String | Key term of interest to audit team |
| Audit Team | String | Name of audit team at Auditor-General's Department |

*Table 10: Key Terms Table*

|  |  |  |
| --- | --- | --- |
| Variable | Data Type | Description |
| File Name | String | Name of XML file. Includes file extension (.xml) |
| Text ID | String | Unique identifier for text |
| AGD Client | String | Common name of client |
| AGD Formal | String | Formal name of client |
| Client Type | String | Type of client (e.g. Local councils) |

*Table 11: Client Mentions Table*

# 

# Appendix B: SSIS Variable Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Description of the variable | Name | Scope | Data Type | The value needs to be changed if the package moves? |
| Archive folder path, which contains all the laded XMLs | AchiveFolder | Package | String | Yes |
| Rejected folder path, which contains all the rejected XMLs (duplicates) | RejectFolder | Package | String | Yes |
| Source folder path for Bill XMLs before they being loaded | DirBill | Package | String | Yes |
| Source folder path for Questions & Answers XMLs before they being loaded | DirQA | Package | String | Yes |
| Source folder path for Question Time XMLs before they being loaded | DirQT | Package | String | Yes |
| Renaming variable after they moved to the Archive folder | FileDestPath | Package | String | No |
| Name of the XML files | FileName | Package | String | No |
| Renaming variable after the rejected XMLs are moved to the Rejected folder | FileRejectedPath | Package | String | No |
| Full path and file name of Bill XMLs | FileSourcePathBill | Package | String | No |
| Full path and file name of Question & Answers XMLs | FileSourcePathQA | Package | String | No |
| Full path and file name of Question Time XMLs | FileSourcePathQT | Package | String | No |
| URL | URL | Package | String | No |
| XMLs after XSLT transformation | varXMLOutput | Package | String | No |