## 



Handover and Completion Report

29.05.2020

Zachary Herberstein, Ashley Foley & Supriya Mayuri

**Western Sydney University**Second Ave,   
Kingswood NSW 2747

# Table of Contents

[**Table of Contents**](#_heading=h.3znysh7) **[1](#_heading=h.3znysh7)**

[**Executive Summary**](#_heading=h.2et92p0) **[2](#_heading=h.2et92p0)**

[**Introduction**](#_heading=h.3dy6vkm) **[3](#_heading=h.3dy6vkm)**

[**Test Results**](#_heading=h.1t3h5sf) **[3](#_heading=h.1t3h5sf)**

[**Identify word-wrap and multiline cells**](#_heading=h.4d34og8) **[3](#_heading=h.4d34og8)**

[**Handle word-wrap cell**](#_heading=h.2s8eyo1) **[4](#_heading=h.2s8eyo1)**

[**Handle multiline cell**](#_heading=h.17dp8vu) **[5](#_heading=h.17dp8vu)**

[**Use machine learning model to identify tables**](#_heading=h.26in1rg) **[6](#_heading=h.26in1rg)**

[**Variations Report**](#_heading=h.lnxbz9) **[7](#_heading=h.lnxbz9)**

[**From development and bug fixing point of view outstanding issues**](#_heading=h.35nkun2) **[7](#_heading=h.35nkun2)**

[**Conclusion**](#_heading=h.2jxsxqh) **8**

# Executive Summary

Insightus is a team of volunteer data scientists who work on public policy projects such as well water contamination and voter access. Multiple PDF reports contain data that need to be extracted and the tool Tabula was used previously to extract the data inside tables in the PDF files. However, it had limitations due to which the Tabula Enhancement Project at NCSU was initiated. After many limitations were overcome, the next step was to handle word wrap in a cell and to handle multi-line cells. The Data Liberation Project focuses on extending these capabilities as well as providing a wider objective on new technologies like machine learning that can provide a more robust solution.

The aim of this document is to present variations such as data, functional and screen. Additionally, the test results indicate the type of testing that was conducted and if the expected results successfully matched the actual results. Further improvements that could be made to the project is also mentioned.

# 

# Introduction

The purpose of this report is to highlight the changes and variations that have been made to the design as a result of constant testing in the development process of the Data Liberation Project. Test results were gathered from each phase of the process and issues were constantly being identified. Known issues that could be useful for future developments will be discussed in the report as well.

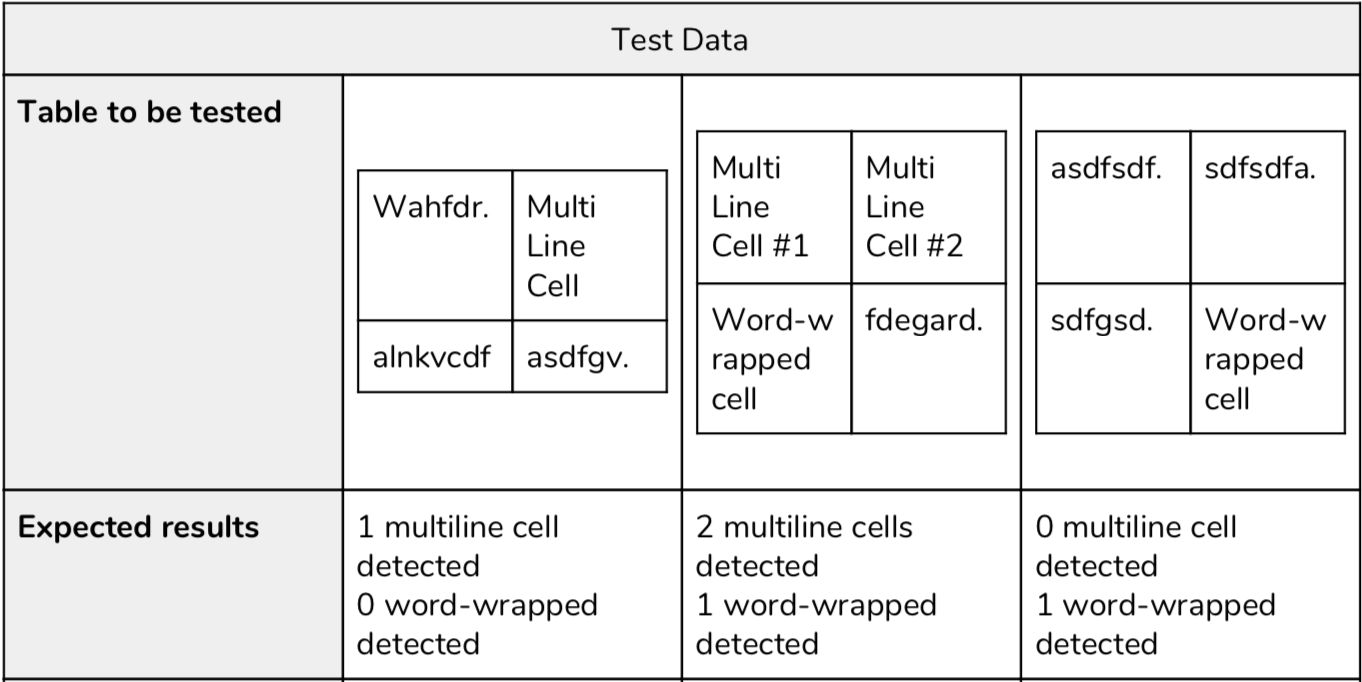
# Test Results

| **Features to be tested** | **Type of testing** | **Personnel** | **Method** |
| --- | --- | --- | --- |
| Identify word-wrap and multi line cells | User testing | Development Team | Lattice |
| Handle word-wrap cell | User testing | Development Team | Lattice |
| Handle multiline cell | User testing | Development Team | Lattice |
| Use machine learning model to identify tables | User testing | Development Team | Python and Pytorch |

# Identify word-wrap and multiline cells

| **Features to be tested** | **Test Purpose** | **Risk** | **Changes made to existing Tabula** | **Success/**  **Failure** | **Date**  **Tested** |
| --- | --- | --- | --- | --- | --- |
| Cell data when word-wrap occurs and multiline occurs | To recognize the correct set of data when these functions occur in order to accurately apply methods to extract data from cells in the upgraded Tabula system | Cells not being identified correctly | Adding capability to handle word-wrap and multiline cells in PDF documents by using the flag method to identify each cell as word-wrap or multiline | Success | 09/06/20 |

**Expected Results:**

****

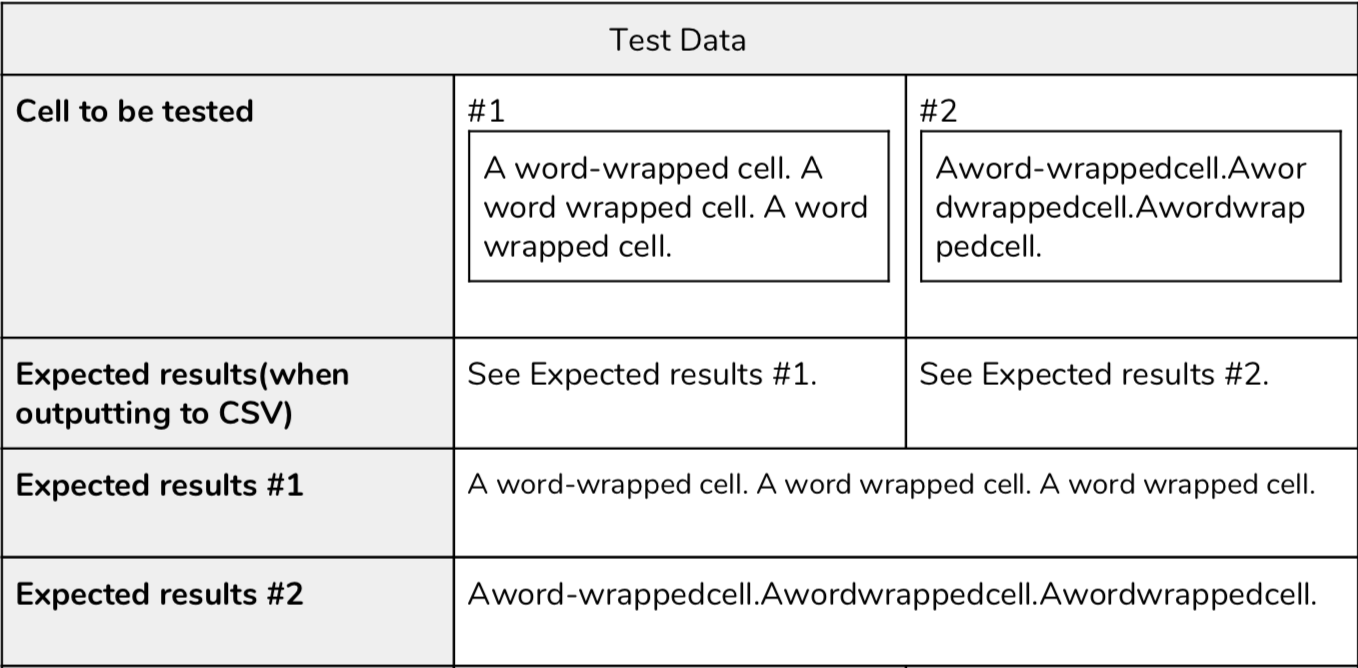
**Actual Results:**

Same as expected results.

# Handle word-wrap cell

| **Features to be tested** | **Test Purpose** | **Risk** | **Changes made to existing Tabula** | **Success/**  **Failure** | **Date**  **Tested** |
| --- | --- | --- | --- | --- | --- |
| Cell data when word-wrap occurs in PDF document | Once word-wrap in the table is detected, cell data need to be correctly merged into one cell to be accurately outputted to CSV format | Data incorrectly handled which lead to wrong CSV output | Adding capability to handle word-wrap in PDF documents by removing the white spaces caused by word-wrap and new lines, and forming data into one line of text | Success | 09/06/20 |

**Expected Results:**

****

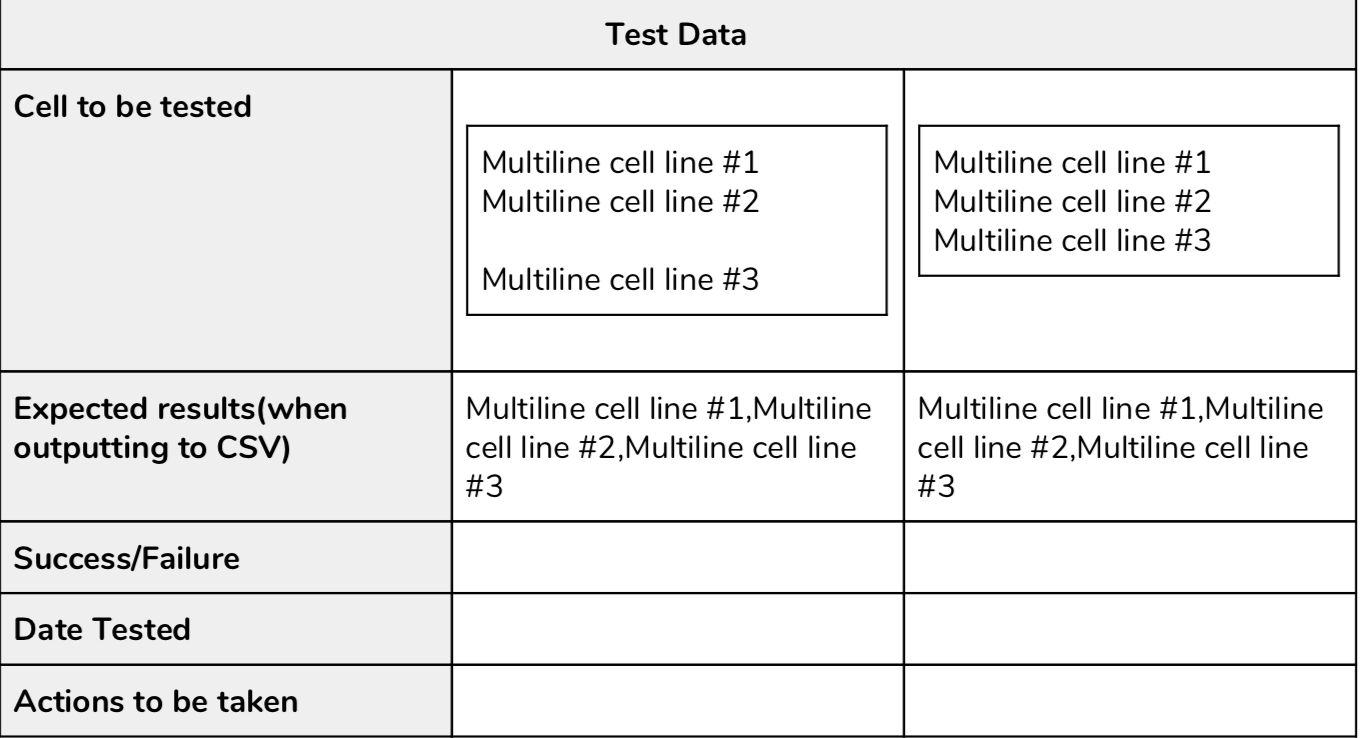
**Actual Results:**

Same as expected results.

# Handle multiline cell

| **Features to be tested** | **Test Purpose** | **Risk** | **Changes made to existing Tabula** | **Success/**  **Failure** | **Date**  **Tested** |
| --- | --- | --- | --- | --- | --- |
| Cell data when multiline occurs in PDF document | Once multiline in the table is detected, cell data need to be correctly divided into separate cells to be accurately outputted to CSV format | Data incorrectly handled which lead to wrong CSV output | Adding capability to handle multiline cells in PDF documents by separating data into multiple cells where a line break occurs in a cell | Success | 09/06/20 |

# Expected Results:

****

**Actual Results:**

Multiline cell line #1

Multiline cell line #2

Multiline cell line #3

# Use machine learning model to identify tables

| **Features to be tested** | **Test Purpose** | **Risk** | **Changes made to existing Tabula** | **Success/**  **Failure** | **Date**  **Tested** |
| --- | --- | --- | --- | --- | --- |
| Treating PDF files as an image and recognizing cell data by shapes on PDF | To be able extract data from PDF with fast response time and with better performance | Wrong pattern picked up and stored in memory | No changes were made as this is a new addition | Success | 09/06/20 |

**Expected Results:**

Tables and cells are recognized as rectangular shapes and stored in memory as patterns.

**Actual Results:**

As our proof of concept model does not recognise tables as of yet this could not be tested, although tests relating to detecting shapes(Circles, Rectangles) were successful.

# Variations Report

| **Type** | **Variation** | **Initiated By** | **Implemented or not** |
| --- | --- | --- | --- |
| Functional | * Add a machine learning model to visually represent recommendations | Client | Implemented |
| Data | * Use lattice method instead of stream | Team | Implemented |
| Screen | * Add visual indicators to GUI to give options for multiline and word-wrap | Team | To be included as further development |

# From development and bug fixing point of view outstanding issues

Currently there are no known bugs relating to the work conducted on this project. However, bugs relating to previous Tabula development as listed in the code comments and on the Github are still present. Additionally, implementation of the new algorithm into the GUI is considered as a further improvement to the project.

# 

# Conclusion

This document presented details of the processes that were conducted throughout the project. It emphasizes on testing features such as identifying word-wrap and multiline cells, handling word-wrapped cell data, and handling multiline cell data, and how test cases were carried out. Additionally, variations made to the project are included as well as known issues that were identified along the way.

