

Portfolio Task 1

Sandy Pullen

2022-05-03

Portfolio Task 1

Introduction

The Tasmanian Land Records (1832-1935) dataset contains over 23,000 records relating to deeds of land grants under General Law, which is a system of land title based on common law used up to 1862 in Tasmania.

The Land records dataset, each record contains links to digital objects available at the Tasmanian Names Index held and maintained at Libraries Tasmania. The nature of these links is as follows:

1. The URL of the person in the Names Index, which then links to the next two URLs listed here,
2. The URL to the index page containing the index record,
3. The URL of the digitised image of the deed of land grant relating to that record.

The area of interest for this project is the location of Kempton, a historical town which was previously known as Green Ponds, in the southern Midlands of Tasmania.

Method

The Tasmanian Land Records (1832 - 1935) dataset was downloaded in CSV format.

The dataset was opened in Open Refine and the following steps were performed.

1. Check for the location of Kempton, and verify that the earlier name of Green Ponds was not used, and that there were no misspellings for Kempton using 'text facets'.
2. Filter for Location=Kempton using 'text facets'. This reduced the dataset to 55 records.
3. Check that each column had data, using 'text facets' to check for blanks.
4. The URL fields contained combined links, so 'Edit Column > Split into several columns' was used to separate the columns into meaningful data as follows:

Original Column	Separator Character	New Columns
DIGITAL_OBJECT - URL_TEXT		DIGITAL_OBJECT - URL_TEXT 1 DIGITAL_OBJECT - URL_TEXT 2
DIGITAL_OBJECT - URL		DIGITAL_OBJECT - URL 1 DIGITAL_OBJECT - URL 2
NAME	,	LASTNAME FIRSTNAME

Original Column	Separator Character	New Columns
FIRSTNAME	<space>	FIRSTNAME1 FIRSTNAME2

Three records contained two firstnames, however, since this did not aid in identification of duplicates, the column was split using the space as a separator, and the second firstname was ignored.

The column DIGITAL_OBJECT - URL 2 contained a link to a URL on the Libraries Tasmania website which showed an image viewer, and the relevant image embedded into the viewer. For example the URL for the land grant for Thomas Croxton is:

[https://stors.tas.gov.au/RD1-1-12\\$init=RD1-1-12P110JPG](https://stors.tas.gov.au/RD1-1-12$init=RD1-1-12P110JPG)

Inspection of the html code using Firefox revealed that the direct link to the image file without the image viewer frame was:

<https://stors.tas.gov.au/fetch/RD1-1-12P110JPG>

So this column was split, using a separator of '=' to isolate the image name required (i.e. RD1-1-12P110JPG). The column was renamed 'IMAGE'.

The data file was saved as

A Python script was developed to iterate (loop over) the rows of the dataset, and for each image listed, web scraping was used to retrieve the image file and save it with a file extension of .jpg to a subfolder named 'captured'. The LASTNAME and FIRSTNAME1 were added to the filename created in the python script to aid readability of the image list.

Web sources used to develop the script are noted here: ¹ ² ³

An example of the digitised land record is shown in Figure 1.

Results

For the 55 records, there were 54 unique images. One image was repeated due to the land being owned in partnership. (i.e, Thomas Croxton and Edward Randall were joint owners, so there is one record for each person, linking to the same image.)

One record had blank year information, this was visually checked on the deed image, and confirmed as missing. 'missing' was entered into this cell in the data entry phase.

Reference List

¹<https://www.rstudio.com/blog/three-ways-to-program-in-python-with-rstudio>

²https://www.machinelearningplus.com/pandas/pandas-read_csv-completed/

³<https://www.educative.io/edpresso/how-to-locally-save-an-image-using-urllib>

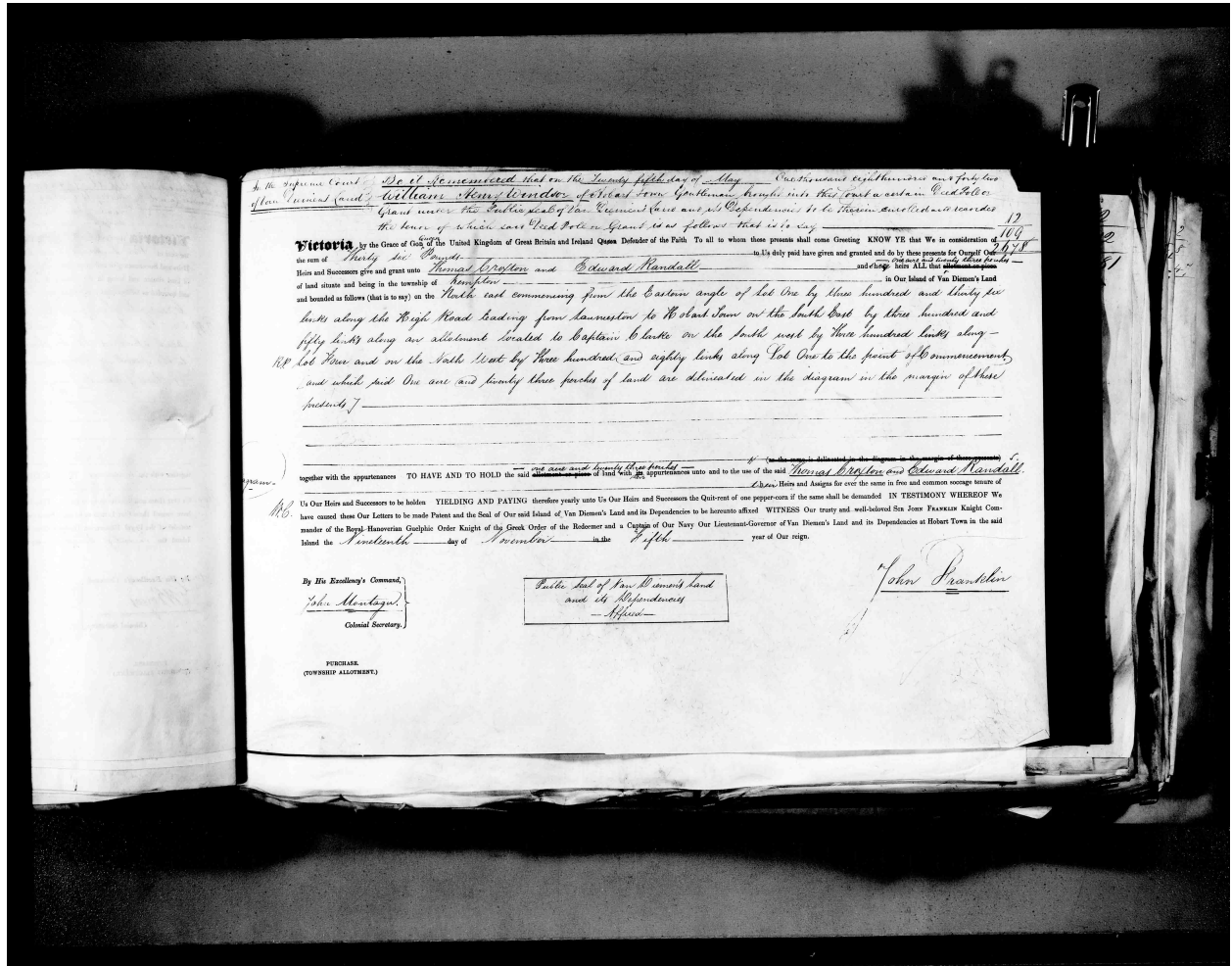


Figure 1: Deed of Land Grant for Thomas Croxton