ReadMe

Sandy Pullen

2022-05-26

Portfolio Task 1 - Data Schema

Data Set Description

The Tasmanian Land Records (1832-1935) dataset contains over 23,000 records relating to deeds of land grants under General Law.

The Land records dataset contains links to digital objects available at the Tasmanian Names Index held and maintained at Libraries Tasmania.

Data Sources

The data sources used were:

1. Land records dataset - https://data.gov.au/dataset/ds-dga-4940b8e0-a2b4-49e6-b658-b124d98e2082/details?q=

(Note: Original metadata is held at https://data.gov.au/data/dataset/tasmanian-land-records-1832-1972)

This dataset contains the following fields:

- DIGITAL_OBJECT - URL_TEXT
- DIGITAL_OBJECT - URL
- LAND_DATE
- RECORD_URL
- REMARKS
- LOCATION
- NAME
- YEAR
- DESC_SR

The datset was filtered to select only those records with a LOCATION containing 'Kempton' resulting in 55 records.

2. The digital objects

The field DIGITAL_OBJECT - URL contains a link to a digitised image in the Tasmanian Archives of the original land grant record. The collection of images relevant to this dataset was downloaded using a Python script and saved in to a sub-directory called '/captured'.

The data in the added fields listed below has been transcribed from the digitised images.

The added fields are:

- BOUNDARY
- SUM_POUNDS
- SUM_SHILLINGS
- SUM_PENCE
- DIAGRAM

Files used in this project

Filename	Description
land.csv	raw data of 23,000 + records
refined-land-csv-May-1.csv	filtered and refined dataset of 55 records for Kempton
data-entry.xlsx	refined dataset with added columns for data entry, validation rules included.
portfolio-task-01-data.csv	exported file from Excel after data entry completed on data-entry.xlsx
capture-images.py	Python script to download images from Libraries Tasmania website

Variables and Rules

The following variables were used for data entry, transcribing written words from the digital image relating to the land record.

Column Heading	Rules
SUM_POUNDS	Contains the value in integers of the written word amount for pounds after the text 'the Sum of'. Value = 'np' if no monetary value present
SUM_SHILLINGS	Contains the value in integers of the written word amount for shillings after the text 'the Sum of'. Value= 0 if pounds has a value but there are no shillings. Value = 'np' if no monetary value present
SUM_PENCE	Contains the value in integers of the written word amount for pence after the text 'the Sum of'. Value= 0 if pounds and/or shillings have a value but there are no pence. Value = 'np' if no monetary value present
BOUNDARY	Transcribed text of the paragraph following the text 'bounded as follows (that is to say) on the' on the digital image. This will be used to accurately locate the land grant on a map.
DIAGRAM	Indicates whether a plan is drawn in the margins Value = 'yes' or 'no'

Reference List

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

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

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