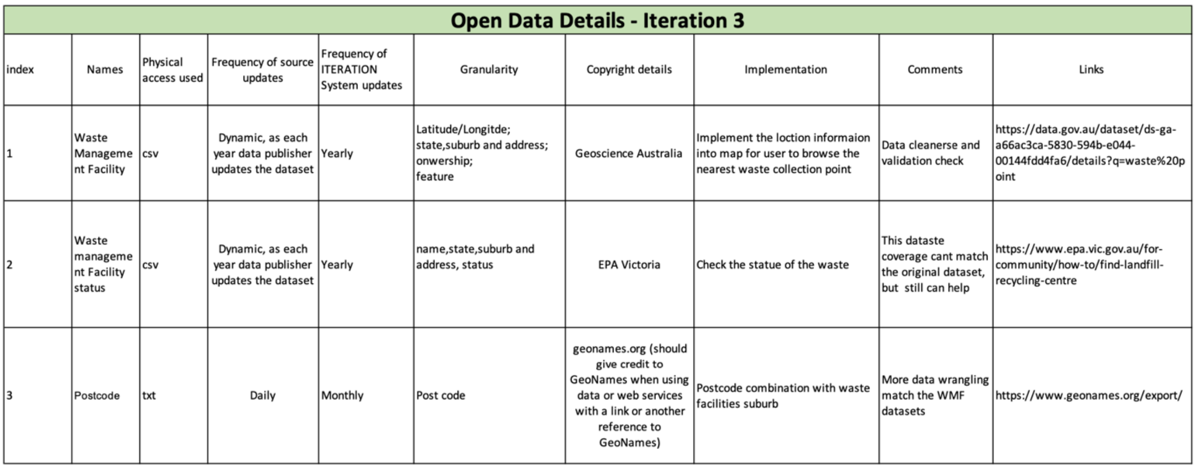
MA15 Data Plan – Iteration 3

For iteration 3, the datasets are mainly for the feature that allows the user to find the nearby waste management facilities. Where the datasets should be presented by the location (address and coordinator), name, and the phone number of each waste facilities.

The raw datasets consist of three parts, the facilities information, postcode information, the statue of the facilities. Because the original dataset is first published in 2012, which there could be some errors and changes in the dataset. To obtain the most updated information and the accuracy of the datasets, the data processing will involve the information checking stage with the help of google map API.



Open Data Details - Iteration 3

index	Names	Physical access used	Frequency of source updates	Frequency of ITERATION System updates	Granularity	Copyright details	Implementation	Comments	Links

1	Waste Management Facility	csv	Dynamic, as each year data publisher updates the dataset	Yearly	Latitude/Longitde;
					state,suburb and address;
					onwership;
					feature
						Geoscience Australia	Implement the loction informaion into map for user to browse the nearest waste collection point	Data cleanerse and validation check	https://data.gov.au/dataset/ds-ga-a66ac3ca-5830-594b-e044-00144fdd4fa6/details?q=waste%20point

2	Waste management Facility status	csv	Dynamic, as each year data publisher updates the dataset	Yearly	name,state,suburb and address, status	EPA Victoria	Check the statue of the waste	This dataste coverage cant match the original dataset, but still can help	https://www.epa.vic.gov.au/for-community/how-to/find-landfill-recycling-centre
3	Postcode	txt	Daily	Monthly	Post code	geonames.org (should give credit to GeoNames when using data or web services with a link or another reference to GeoNames)	Postcode combination with waste facilities suburb	More data wrangling match the WMF datasets	https://www.geonames.org/export/

Step 1 and Insights: Waste management facilities dataset EDA, it shows there are some duplicate records and the detail information like street name and house number are unclear.

- There are total over 2300 recorded waste mananged facilities in Australia wide. NSW has the largest number of registered waste manangment facilities, the second place is Victoria. Which follow the same pattern of population of each state.
- The address attribute is this dataset is misplace with a lot of records, main issue are duplicate and vague information. Which will be fixed by the
- Most of waste management facilities are landfill, the second is transfer station.
- In the datasets, most of the waste management facilities are controlled by local governments. And there are still some large environment cooperations participant in the waste management business.

Step 2 and Insights: Filter the Victoria state data, it shows the suburb attribute has some wrong records.

- In Victoria region, the dataset shows some errors in suburb name and address detail. It also shows a pattern that in the same address, there are multiple facilities where they do have difference in functionality but belong to the same cooperation.
- For that situation, leave it to the following step to process.

Step 3 and Insights: Combine the postcode dataset and waste management facilities datasets through the suburb. When comes to multiple postcodes from the same suburb, manually check those records.

- Since the raw dataset did not involving the postcode, combine the postcode based on the suburb to add this attribute.
- Postcode can privide information to when processing the full address of each records and assist to the potential search function.
- It may happen that for one suburb, it could be multiple postcode, leave the multiple postcode records out to manual check.

Step 4 and Insights: Waste facilities statue check. With the help of waste facilities statue dataset, it recorded some closed waste facilities and due to close facilities. Could help to reduce and update the raw dataset.

- Waste management facilities dataset was created in 2012, which is the latest possible dataset could find and consider as open data.
- Waste facilities statue dataset only has about 500 records, over 400of them as closed satue. Which can only as a reference to double check the quility of the final implemention dataset.
- Waste facilities statue dataset and Waste management facilities dataset have different attributes, facilities name is the only attribute can help to find the common records. But it still could be some variance.
- There are total 80 records in Waste management facilities dataset considered as closed statue in 2020.

Step 5 and Insights: Remove duplicated records. The duplicated records may be formatted as similar name, address. Using the difflib library to detect the similar name of each location.

- As mentioned above, the first process is find the duplicated records, as it may just similar name but not exactly, difflib library privide the funtion to check the similarity of each records. If it is truly consisted as the same, check the coordinator to find out the if those records are refering to the same one.
- There are total over 16 records considered as duplicated from the aspect that they have the similar name or address, and their coordinators show that nearby only could have one possible waste management facility.

Step 6 and Insights: Add detail information of each record. Using google map API, from google place to obtain the phone and the full address, setting the coordinator as the searching key. Due to the quality of the dataset, this process cannot match the raw dataset records. For these unsearchable records, manually check and add value to the table.

- Google map API and google map place funtion help to find the detailed information of each records. Assuming the coordinator is correct, with certain ratio starte from the coordinator, searhing with the key words as ‘waste’/ ‘ landfill’ / ‘transfer’ to have the information like full address and phone number.

- Almost 130 records can not be searched by google map, which turn into manual check.

Step 7 and Insights: Revise the precessed data, check again if there would be misrecorded information and duplicated information.

Data Implementation:

- Consider the project will not using user real-time location, like GPS, we perfer to have a database of waste management facilities. Though the dataset can not cover the whole actual facilities, but can trust on most of the records are been double check the statue and information.
- The database for this feature contains those attributes (name, latitude, longitude, suburb, postcode, address, fullAddress phone). Planing to using one table to achieve that.
- User will enter suburb name to find the nearby waste facilities. After the searching, it will show a map of the nearby facilites and the detial information of each presented facility.