Milestone 1: Project Proposal and Data Selection/Preparation

Step 1: Preparing for Your Proposal

Client selected is Tanzania Ministry of Water. They need us to predict which water pumps are faulty.

The data is taken from Taarifa which aggregates data from the Tanzania Ministry of Water.

These are the data we are looking at:

In [1]:
```
import pandas as pd
```

In [2]:
```
df = pd.read_csv("train.csv", low_memory=False)
```

In [3]:
```
df
```

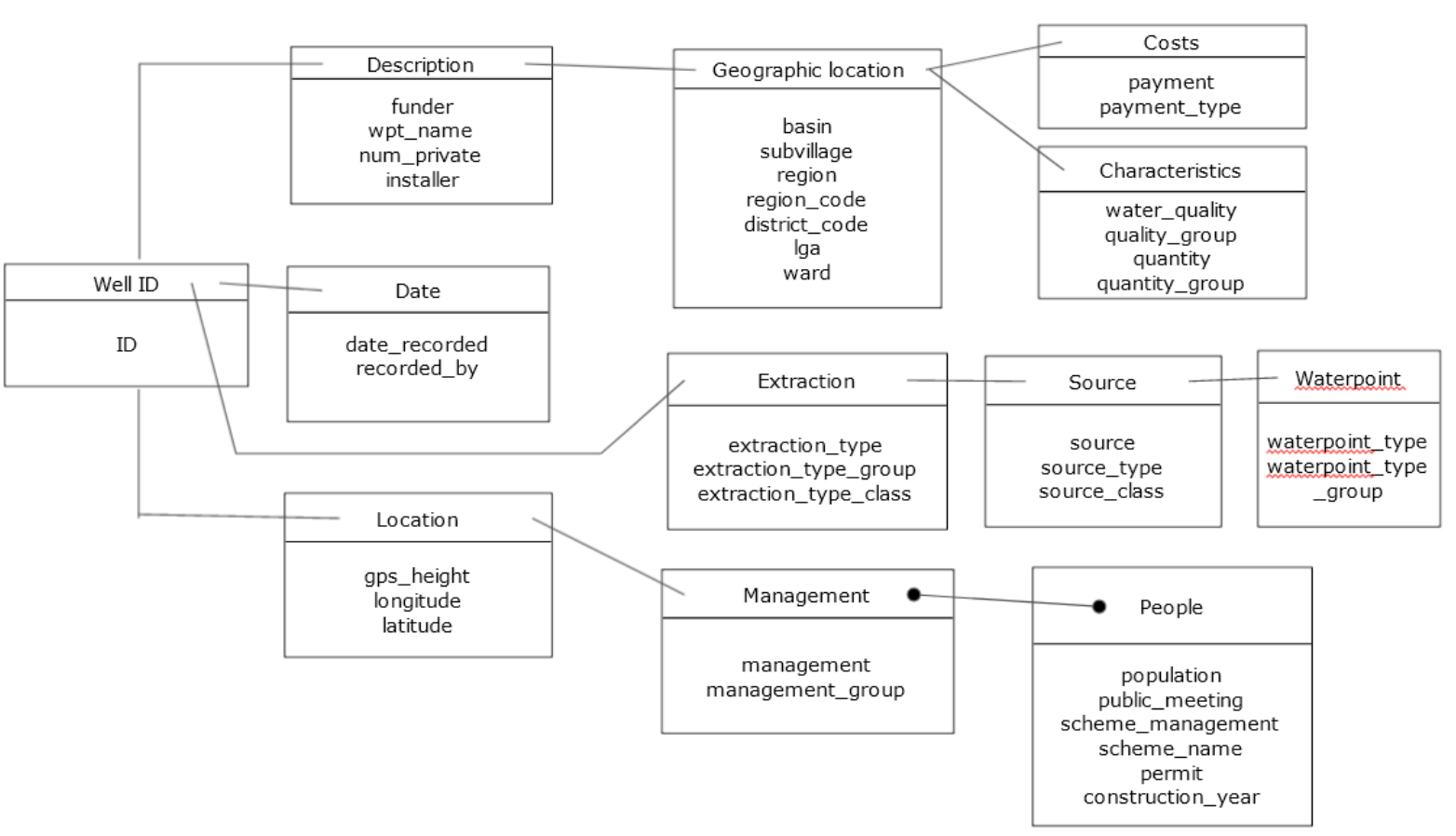
Out[3]:

	id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_private	...	water_qua
0	69572	6000.0	14/3/2011	Roman	1390	Roman	34.938093	-9.856322	none	0	...	
1	8776	0.0	6/3/2013	Grumeti	1399	GRUMETI	34.698766	-2.147466	Zahanati	0	...	
2	34310	25.0	25/2/2013	Lottery Club	686	World vision	37.460664	-3.821329	Kwa Mahundi	0	...	
3	67743	0.0	28/1/2013	Unicef	263	UNICEF	38.486161	-11.155298	Zahanati Ya Nanyumbu	0	...	
4	19728	0.0	13/7/2011	Action In A	0	Artisan	31.130847	-1.825359	Shuleni	0	...	
...	...	...	...	...	...	...	...	...	...	...	...	
59395	60739	10.0	3/5/2013	Germany Republi	1210	CES	37.169807	-3.253847	Area Three Namba 27	0	...	
59396	27263	4700.0	7/5/2011	Cefa-njombe	1212	Cefa	35.249991	-9.070629	Kwa Yahona Kuvala	0	...	
59397	37057	0.0	11/4/2011	NaN	0	NaN	34.017087	-8.750434	Mashine	0	...	fluo
59398	31282	0.0	8/3/2011	Malec	0	Musa	35.861315	-6.378573	Mshoro	0	...	
59399	26348	0.0	23/3/2011	World Bank	191	World	38.104048	-6.747464	Kwa Mzee Lugawa	0	...	s

59400 rows × 41 columns

We are interested in finding status_group which is the water pump condition:

Proposed ERD



Step 2: Develop Project Proposal

Description

The Tanzania Water Ministry would like to analyse and report findings on all water pumps throughout the country, They also would like a prediction model to be developed based on data recorded. A smart understanding of which waterpoints will fail can improve maintenance operations and ensure that clean, potable water is available to communities across Tanzania.

Questions that you want to answer with the data:

1. Are there any locations which has more breakdowns?
2. Does Management issues affect the pump maintainences?
3. How about water extraction, source and water points affect the pumps?

Hypothesis assumptions

1. Locations like Basin, Subvillage and Region will determine pump operations
2. Populations are more or less similarly in quantity
3. Water source, types, quantities, quality are consistent in all water pumps

Approach

We will use Geographic locations, Management, and Extraction features to do data exploration.

We will use proper graphs to illustrate any relationships exists.

Metrics will be used are accuracy, precision, recall and F1 scores.