In this project, I use the dataset of Wuhan city in China, file size is 64 MB uncompressed.

Here are some challenges encountered during the wrangling:

1. Not all ‘name’ tag are in the same language.
2. There are a lot of duplicated tags.
3. Some tags as "bench", "shelter", "atm" always appear with the same value ‘yes’. It seems to me, they can be noted in a list, as facilities provided.
4. Sometimes, the tag ‘name’ isn’t written in Chinese, I replace it with value of tag ‘name:zh’.

```
def get_type(elem):
```

This function takes an element, try to split key with colon. It returns string and array.

```
def shape_element(element):
```

This function takes an xml element, transfer it to a JSon dictionary.

```
def audit_chinese(s):
```

This function audit if parameter s is pure Chinese character, returns True or False.

Here are ideas for additional improvements.

As there are a lot of duplicated nodes with same name tag, provided by different users. It could be more efficient, if we come up with a neutral template, which combines those tags all together. This approach can massively decrease duplication in this giant dataset.

The downside is that we might miss out those unique tags, that only some users noted.

```
def id_query(id):
```

```
def id_query(id):
```

Query of id, is used to find a specific node in collection.

```
def user_query(user):
```

Query of user, is used to find all child user of a created parent.

```
def service_query(srv):
```

Query of service, is used to find nodes that contains facility.