In [1]:

```
import json
import pandas as pd
import numpy as np
import os
import pprint
import openpyxl
```

In [2]:

```
os.listdir()
```

Out[2]:

```
['.DS_Store',
 '.ipynb_checkpoints',
 'annotations.json',
 'annotations_1.xlsx',
 'annotations_10.xlsx',
 'annotations_11.xlsx',
 'annotations_12.xlsx',
 'annotations_13.xlsx',
 'annotations_14.xlsx',
 'annotations_15.xlsx',
 'annotations_16.xlsx',
 'annotations_17.xlsx',
 'annotations_18.xlsx',
 'annotations_19.xlsx',
 'annotations_2.xlsx',
 'annotations_20.xlsx',
 'annotations_21.xlsx',
 'annotations_22.xlsx',
 'annotations_3.xlsx',
 'annotations_4.xlsx',
 'annotations_5.xlsx',
 'annotations_6.xlsx',
 'annotations_7.xlsx',
 'annotations_8.xlsx',
 'annotations_9.xlsx',
 'merged_data.xlsx',
 'Merger.ipynb',
 'Regiatery_Complete.csv',
 'submatcheds.txt',
 'Untitled.ipynb',
 'Untitled1.ipynb',
 '~$merged_data.xlsx']
```

In [ ]:

```

```

In [3]:

```
registry = pd.read_csv("Regiatery_Complete.csv", low_memory=False)
softcite = pd.read_excel("merged_data.xlsx")
```

In [4]:

```
registry.head()
```

Out[4]:

	rid	scr_id	original_id	type	parent_organization_id	Resource_Name	Defini
0	1	SCR_000001	nlx_152482	Organization	(null)	TransGenic	
1	2	SCR_000002	nlx_152901	Resource	SCR_001373	monarch-ontologies	
2	3	SCR_000003	nlx_158000	Resource	(null)	Sarah Cannon Research Institute; Tennessee; USA	
3	4	SCR_000004	nlx_152368	Organization	(null)	GE Healthcare	
4	5	SCR_000005	nif-0000-00023	Resource	(null)	Neuroshare - Open data specifications and soft...	

5 rows × 51 columns

In [5]:

```
softcite_Name = []
for i in list(softcite["sn"]):
    if type(i) == str:
        softcite_Name.append(i)
```

In [6]:

```
registry_Name = []
source_Name = []
for i in list(registry["Resource_Name"]):
    if type(i) == str:
        registry_Name.append(i)

for i in list(registry["resources_names"]):
    if type(i) == str:
        source_Name.append(i)
```

In [7]:

```
matching = []
submatch = []
for i in softcite_Name:
    for j in registry_Name:
        if i == j:
            matching.append(i)
        if i in j:
            submatch.append(i)
```

In [8]:

```python
resourcematch = []
resmatchedSUB = []

for i in softcite_Name:
    for j in source_Name:
        if i == j:
            resourcematch.append(i)
        if i in j:
            resmatchedSUB.append(i)
```

In [18]:

```python
resmatchedSUB
```

```
 'Python',
 'Python',
 'Python',
 'Python',
 'Python',
 'Python',
 'Python',
 'Python',
 'Python',
 'Python',
 'Python',
 'Python',
 'Python',
 'Python',
 'Python',
 'Python',
 'Python',
 'Python',
 'Python',
 'Python',
```

In [9]:

```python
submatch
```

```
 'RStudio',
 'dplyr',
 'CyTOF',
 'CyTOF',
 'CyTOF',
 'openxlsx',
 'tidyr',
 'Enrichr',
 'HISAT2',
 'mixOmics',
 'featureCounts',
 'EdgeR',
 'GalaxyRefine',
 'SCRATCH',
 'SnapGene',
 'SnapGene',
 'Living Image',
 'Gen5',
 'Nanopolish',
 ...]
```

In [10]:
```
matching
```

```
 'TFPGA',
 'LeadIT',
 'VIPERdb',
 'MariaDB',
 'HEM',
 'Grinder',
 'dbNSFP',
 'MetaCyc',
 'GAS',
 'CLAIRE',
 'CCP',
 'Brainstorm',
 'Orange Data Mining',
 'CopyCaller',
 'NetSurfP',
 'OpenPose',
 'DPABI',
 'Protege',
 'FASTSLINK',
 'ImmuneSpace',
```

In [11]:
```
len(matching)
```

Out[11]:
```
1952
```

In [12]:
```
len(submatch)
```

Out[12]:
```
185147
```

In [13]:
```
new_list = list(set(submatch))
```

In [14]:
```
new_list
```

```
 'Count',
 'MetaCore',
 'GeneCard',
 'GROMACS',
 'Pati',
 'Analysis)',
 'MFDp',
 'pybedtools',
 'Meta-IDB',
 'CS',
 'MuTect',
 'SilkDB',
 'Epik',
 'ModBase',
 'SeqPrep',
 'Mouse Phylogeny Viewer',
 'MobiDB',
 'IMS',
 'UMI-tools',
 'JASPAR',
```

In [15]:

```
len(new_list)
```

Out[15]:

```
5383
```

In [16]:

```
with open("submatcheds.txt", "w", encoding="utf-8") as f:
    f.write('\n'.join(new_list))
```

In [ ]:

```

```