New Homework Project

**Database parser**

The goal here is to make a parsing module for JSON formatted picture archives. Overall, the input will be a JSON file and the output will be a consistent data structure which allows easy down stream parsing (say into a postgress-importable CSV file). Instead of directly uploading the JSON files into postgres, the idea is that a parsing module may allow for further formatting. For example, parsing may allow one to normalize the JSON data into further atomized tables, or into a structure consistent with a predefine entity relationship diagram.

Useful URLs:

* The github page for the database pictures.
  + <https://github.com/Smithsonian/OpenAccess/tree/master/metadata/objects>
* General information on the repository release.
  + <https://www.si.edu/openaccess>
* OpenAccess on github
  + <https://github.com/Smithsonian/OpenAccess>
* API for OpenAccess
  + <https://www.si.edu/openaccess/devtools>

Major headers in the JSON file (at least for NPG):

id

version

unitCode

linkedId

type

content – contains many sub dirs. – making unique table

url

hash

docSignature

timestamp

lastTimeUpdated

title

multiple values in some: (make into separate table or multi-valuee attribute?)

'name': [{'label': 'Artist', 'content': 'Max Rosenthal, 1833 - 1918'}, {'label': 'Copy after', 'content': 'Henry Louis Stephens, 1824 - 1882'}, {'label': 'Sitter', 'content': 'James S. Wallace'}], 'topic': [{'label': 'Topic', 'content': 'Exterior'}, {'label': 'Topic', 'content': 'Caricature'}, {'label': 'Topic', 'content': 'Nature & Environment\\Animal\\Bird'}, {'label': 'Topic', 'content': 'Book'}, {'label': 'Topic', 'content': 'James S. Wallace: Male'}, {'label': 'Topic', 'content': 'Portrait'}]

Output needed for arrays: <https://stackoverflow.com/questions/11170099/copy-import-data-into-postgresql-array-column>

Merging tables: <https://pandas.pydata.org/pandas-docs/stable/user_guide/merging.html>

**Importing into postgres**

The goal for importing into postgres is to take a csv file structure and import the data into a postgres database.

Wrapping psql into bash: <https://www.manniwood.com/postgresql_and_bash_stuff/>

Creating table in psql: <https://stackoverflow.com/questions/2987433/how-to-import-csv-file-data-into-a-postgresql-table>

Tested on psql using:

CREATE TABLE zip\_codes (zip VARCHAR[]);

\copy zip\_codes FROM '~/Desktop/RICESPRING2020/COMP533TA/new\_material\_project/table.csv' DELIMITER ',' CSV

SELECT \* FROM zip\_codes;

SELECT zip[3] FROM zip\_codes;