Forward Data Lab Summer Research Participation

Nandika Vuyyuri

My task for the summer was to integrate grant data and sponsor data with OpenAlex in PostgresQL, specifically I incorporate sponsors data (sponsors.xml) and grants data (grants.xml) into the OpenAlex database.

Starting off, I created a xml_parser that parses grants.xml and sponsors.xml using the ElementTreelibrary, and establishes a PostgreSQL database. I then created tables for the sponsor's data to create the local database schema.

I received the OpenAlex database from a fellow researcher, Patrick, and created a script that finds the matches from the local funders tables to the OpenAlex tables. Originally I determined the matches by checking if the display names and homepage urls were identical. With this, I was able to get 5027 funders matched by name and 304 funders matched by URL, leaving 21,885 funders unmatched.

To combat the low matching rate, I looked towards machine learning models for name matching such as trigram matching which compares sets of three consecutive characters in two strings, metaphone matching which matches similar sounding names by converting them into a standardized phonetic code and then comparing the code of the two strings, and also partial name matching which checks if a smaller string exists in the longer string. I was able to match quite more with the machine learning models, however, this was not fully accurate. I had to fine tune the three functions to be able to be sensitive enough to consider reasonable matches but not too sensitive to contain incorrect matches.

Here I included some results from the Machine Learning Name Matching functions:

Correct Examples of Trigram:

- ***Matched by trigram : Cereals & Grains Association -> Cereals and Grains Association***
- ***Matched by trigram : Alternatives Research & Development Foundation -> Alternatives Research and Development Foundation***
- ***Matched by trigram : American Fibromyalgia Syndrome Association, Inc. -> American Fibromyalgia Syndrome Association***

Incorrect Examples of Trigram:

- ***Matched by trigram : AGC Education and Research Foundation -> AADE Education and Research Foundation***
- ***Matched by trigram : Miami University -> University of Miami***
- ***Matched by trigram : Keller Family Foundation -> Kelly Family Foundation***

Correct Examples of Metaphone matching:

- ***Matched by metaphone : Beckman Institute for Advanced Science and Technology -> Beckman Institute for Advanced Science and Technology, University of Illinois, Urbana-Champaign***
- ***Matched by metaphone : Hawai'i People's Fund -> Hawaii People's Fund***
- ***Matched by metaphone : Sjogren's Foundation -> Sjögren's Foundation***

Incorrect Examples of Metaphone matching:

- ***Matched by metaphone : Cigna Foundation -> SCAN Foundation***
- ***Matched by metaphone : PNC Foundation -> Nike Foundation***
- ***Matched by metaphone : SRI Foundation -> CIRI Foundation***

Correct Examples of Partial Name matching:

- ***Matched by partial name : Ohio Academy of Family Physicians -> Ohio Academy of Family Physicians Foundation***
- ***Matched by partial name : Office of Research Infrastructure Programs -> Office of Research Infrastructure Programs, National Institutes of Health***
- ***Matched by partial name : International Association of Fire Fighters -> International Association of Fire Fighters Charitable***

Possible Incorrect Examples of Partial Name matching:

- ***Matched by partial name : College of Arts and Sciences -> College of Arts and Sciences, Boston University***
- ***Matched by partial name : College of Liberal Arts -> College of Liberal Arts and Human Sciences, Virginia Tech***
- ***Matched by partial name : Department of Anthropology -> Department of Anthropology, University of California, Los Angeles***

After this, I decided to use another attribute in the database such as country code. After a matching occurs through one of the machine learning functions, the country code from the local database is checked with the OpenAlex database. If the country codes differ, that match is deemed incorrect. Most of the data had the country code US so this function wasn’t as helpful as expected.

I implemented another matching function that checks the local database titles to see if there will be a matching with any of the alternate titles from the OpenAlex database. This was a successful function as there is a significant amount of sponsor data that has multiple alternate_names that do match with the name from the local database. I was able to accurately capture many more matches with this. Following the alternate_titles attribute, I similarly looked at other attributes such as image_url and description however implementing matching using these attributes wasn’t feasible with the data I was given. The final matching results include:

Matched by Name :5027

Matched by Alternate Name :621

Matched by Url :304

Matched by Trigram :765

Matched by Metaphone :188

Matched by Partial Name :291

Unmatched : 20020

These results indicate a decent portion of accurate matching. Fine tuning the sensitivity of the machine learning matching functions can vastly change the difference in matches; however, the current settings show to be a good fit for accurate matching. As a significant portion of the matched data comes from the machine learning matching functions, I believe that the results obtained were positive as these functions cannot guarantee an accurate match. One suggestion I have for the data preparation is to add another key code attribute to the sponsors data as this code can be used to guarantee accurate matching.

Throughout the summer, I frequently checked the status of grants in OpenAlex as OpenAlex doesn’t contain grants as of this summer. Due to this, I was not able to complete one of my tasks to integrate grants data into the OpenAlex database. However, I believe that my initial parser file that I created for the grants file will be of help whenever OpenAlex does contain grants. I also believe that the tools and technologies I have used to incorporate the sponsors data should be of significant help for when integrating the grants data as I believe that similar name matching functions will be required.

I learned a lot from the project and I greatly appreciated the feedback I received every week during the update meetings. I believed that the questions I was asked in the meetings allowed me to understand the bigger picture of the project. The only feedback I can provide is to have 2 weekly meetings instead of one so that if someone has a question or is stuck, they don't have to wait for 5-6 days to ask that question. For future work, I can incorporate the grants data to the OpenAlex database when the OpenAlex grants data is published. I will also continue on improving both the matching process as well as the querying process as I believe the queries can be done quicker.