CONTRIBUTING DATA AND DESIGN TO THE MAPPING & VISUALIZATION OF ACADEMIC COLLABORATIONS PROJECT

To analyze equity amongst academic researchers, we are actively developing a system design to accurately construct collaboration networks, represented by collaborative experiences such as publications. Data gathered for this project includes processes such as web scraping to create an accurate construction of the network, and technologies such as SQL and NetworkX are used to represent and analyze networks of collaborations. We describe our work in the summer of 2021 contributing to the data acquisition and database design for this project.

AUTHORS

Theresa Migler, Zoë Wood, Maya Zeng

AFFILIATIONS & ACKNOWLEDGEMENTS

We're also proud of the institutions that we are with and support our research. Computing Research Associate DREU/ Cal Poly SLO

We would also like to thank the following student researchers who took the time to mentor along this process: Brandon Tat, Logan McNichols, and Leticia Siqueira.

INTRODUCTION

We aim to discover key insights into academic equity, for example, analyzing the variance of female researcher's academic collaboration networks. For this project, the network consists of nodes representing researchers and edges representing the collaboration between authors of a shared publication.

Starting with lists of researchers, the Scopus database is used to aggregate publication information as the database provides one of the largest abstract and citation database of peer-reviewed literature [1]. The data can then be visualized using the Google Maps API. Python and NetworkX are used to understand aspects of the network.

OBJECTIVE

To create an accurate representation of the network, we web-scrapped faculty data such as first name and last name from university websites. We then clean and compile this data to be inputted to the database.

From there, we discovered key findings that reflect the accuracy of the data provided and future considerations towards improving the database.

METHODOLOGY

Technologies used in this research include but not limited to:

— Python
— SQL
— Beautiful Soup Python Package
— NetworkX Python Package

Due to the variety of design schemas used by different universities, a human element is needed to extract pertinent data from these websites. Once we analyzed the HTML code, we would adjust the web-scrapping program created using Python and the Beautiful Soup package to obtain the data.

Once data acquisition is completed, we would add the data to the MySQL database.

KEY OBSERVATIONS

Leveraging the Scopus database is a good starting point for our data exploration; however, initial and subsequent cleaning is needed to create an accurate representation of the network for our use case. A few observations have been made such as:

- Categorization

Some schools have two fields in one department. For example, "Department of Electrical Engineering and Computer Science". How would this be categorized? Would we consider professors under this field in both categories or to a general one?

- Missing Data

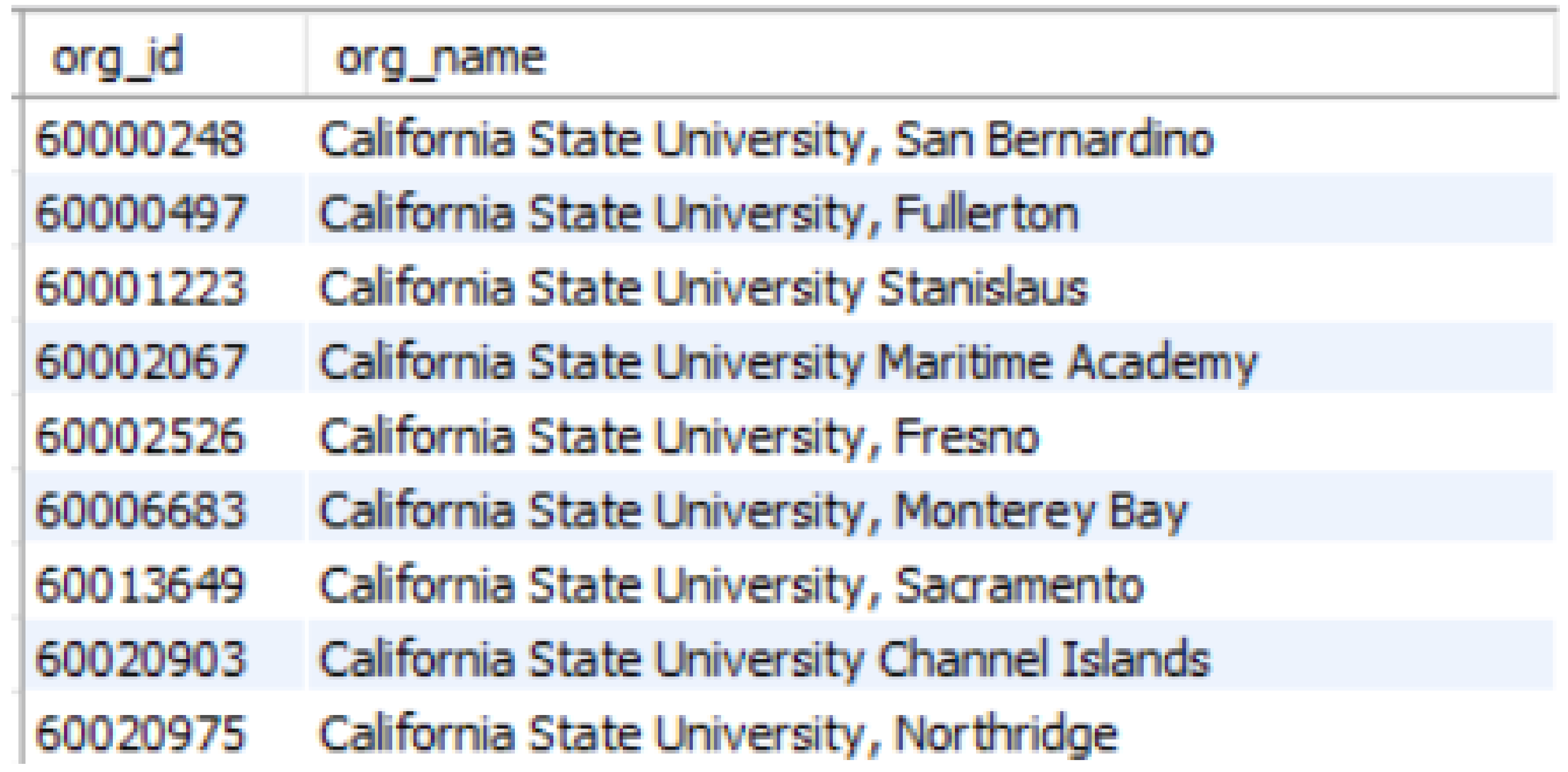
Data acquisition could have missing values. In our case, some professors have retired, and we are unable to obtain their email addresses to create accurate network constructions

- Data Relevant/Current

We came across professors that were not in the Scopus database, but they were in the faculty directory of university websites. Faculty change institutions and even names (due to marriage or other life events). How do we guarantee our data is representational?

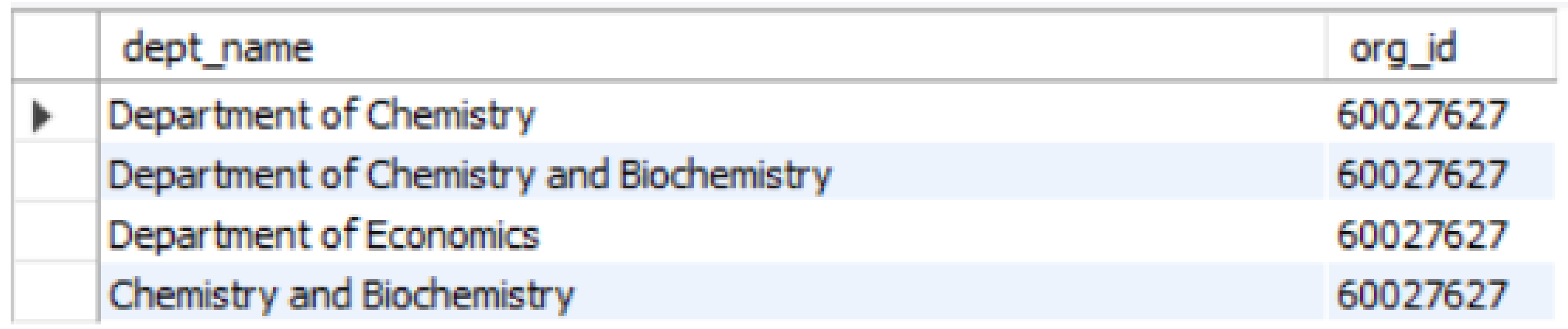
- Conventions/Standardization

IIf one does not use a keyword, the entry would not pop up. For example, the California State Universities generally follow this notation: California State University, [location of university]. However, there are exceptions to this case as shown to the left.



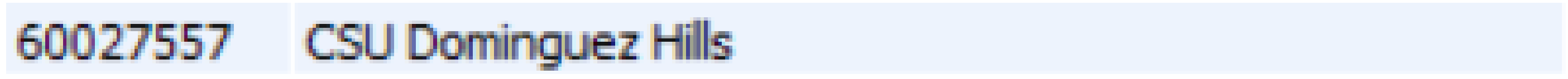
org_id	org_name
60000248	California State University, San Bernardino
60000497	California State University, Fullerton
60001223	California State University Stanislaus
60002067	California State University Maritime Academy
60002526	California State University, Fresno
60006683	California State University, Monterey Bay
60013649	California State University, Sacramento
60020903	California State University Channel Islands
60020975	California State University, Northridge

Some of the CSUs would not follow a standardized format as they do not have a comma before their location



	dept_name	org_id
▶	Department of Chemistry	60027627
	Department of Chemistry and Biochemistry	60027627
	Department of Economics	60027627
	Chemistry and Biochemistry	60027627

There would be cases where there are duplicates of multiple like or same departments, leading to further discussion on what and how department categories should be created



60027557	CSU Dominguez Hills

We had to use the LIKE clause in SQL in order to CSU Dominguez Hills since it did not follow the same notation

CONCLUSIONS

After further discussion, the database redesign includes the following:

- Establish general field of study in accordance to USDA guidelines
- Add additional department fields depending on if the department covers multiple fields
- A checklist was created with universities of our interests to check for the following:
  - Database already has the data
  - Data has been web scrapped for that university

Database development and design is foundational to data science related projects. With that, a human element is heavily needed to find and compensate errors that is in the data provided.

Real world data is noisy; hence we need to be attentive to the types of data being created and infer avenues where data can be misrepresented.

Dataset cleaning is also a time-consuming, exhaustive, yet very important task in data science. So how to do we streamline this process and deliver key findings fast?

Further discussion is needed as data scientists and stakeholders that provide the data need to collaborate with one another to provide and standardize data creation and development.

RELATED LITERATURE

McNichols L. , Pineda S., Sauerborn E., Tat B., Yoo K., Lehr J., Wood W., Migler T. (2021) MAVAC: Mapping and Visualization of Academic Collaborations with a Focus on Diversity. In: Teixeira A.S., Pacheco D., Oliveira M., Barbosa H., Gonçalves B., Menezes R. (eds) Complex Networks XII. CompleNet-Live 2021. Springer Proceedings in Complexity. Springer, Cham. https://doi.org/10.1007/978-3-030-81854-8_8