Study Abroad Programs Database

Rohan Dalal, Nithika Menon, Kshitija Banala, Akhilesh Nimse

DS 320

Professor Huang

December 14, 2024

Abstract

This project aims to integrate study abroad program data from Penn State University (PSU) and Ohio State University (OSU) into a unified and queryable schema. By web scraping public program tables, cleaning inconsistencies, and structuring the data into an SQLite database, we developed a comprehensive database that enables efficient querying and analysis. Queries demonstrated insights such as program distributions by region, college, and location. This work highlights the power of data integration and the importance of standardizing schema to support multi-source data analysis.

Background

Study abroad programs provide students with transformative educational experiences. However, data about these programs is often siloed within individual colleges, making it challenging to compare or analyze offerings across universities. This project addresses the problem by combining publicly available study abroad program data from The Pennsylvania State University and The Ohio State University, creating a single mediated schema for cross-university analysis. The project emphasizes methods of data integration, cleaning, and visualization to derive actionable insights from combined datasets.

Methods

Web Scraping

To begin, we sourced the study abroad program data from the publicly accessible websites of Penn State and Ohio State. Both websites presented their program information in structured tables, making them ideal for web scraping. Using Python's Beautiful Soup library, we extracted the data directly from the HTML content of these tables. Each program's details, such as the program name, city, country, and region, were parsed and saved into a structured format. Pandas was then used to organize and inspect the extracted data for completeness. This step gave us raw datasets for both universities, which needed further refinement.

Data Cleaning

The raw datasets contained several inconsistencies that required careful cleaning. For example, some cities and countries were stored as comma-separated lists within a single entry, creating challenges for accurate analysis. To address this:

- Multi-city and multi-region entries were aggregated into generic labels like "multiple cities" or "multiple regions."
- Entries with multiple countries were split into individual rows to maintain a clean mapping between cities and countries.
- Additional formatting issues, such as leading spaces, commas, and colons, were removed to standardize the data. We also added a "College" column to identify the source of each program (The Pennsylvania State University or The Ohio State

University. This step ensured that the data was clean, consistent, and ready for integration

Mediated Schema Design

To make the data more usable and query-friendly, we designed a mediated schema that unified the attributes from both datasets into a single structure. The schema included the following fields: Program Name, City, Country, Region, and College.

The primary goal was to create a consistent framework for querying across both universities' datasets. For example, programs spanning multiple cities were grouped under "multiple cities," and multi-country entries were normalized into separate rows. Below is an example of our mediated schema

Index	Program.Name	City	Country	Region	College
1	Amsterdam, Netherlands: Dutch Criminal and Social Justice (Summer)	multiple cities	Netherlands	Europe	The Pennsylvania State University
2	Arusha, Tanzania: Biology of Eco-Health (Summer)	multiple cities	Tanzania	Africa	The Pennsylvania State University
3	Athens, Greece: History, Culture, and Archaeology of Greece	multiple cities	Greece	Europe	The Pennsylvania State University
4	Barcelona, Spain: Landscape Architecture, Dept LARCH (Summer)	multiple cities	Spain	Europe	The Pennsylvania State University
5	Besancon, France: Intensive French Language and Cultural Immersion (Summer)	Besançon	France	Europe	The Pennsylvania State University
6	Brno, Czech Republic: Business in Central Europe (Summer)	multiple cities	Austria	Europe	The Pennsylvania State University
7	Brno, Czech Republic: Business in Central Europe (Summer)	multiple cities	Czech Republic	Europe	The Pennsylvania State University
8	Brno, Czech Republic: Business in Central Europe (Summer)	multiple cities	Hungary	Europe	The Pennsylvania State University
9	Brno, Czech Republic: Business in Central Europe (Summer)	multiple cities	Poland	Europe	The Pennsylvania State University
10	Cadiz, Spain: Spanish Language and Culture (Summer)	multiple cities	Spain	Europe	The Pennsylvania State University

This schema served as the foundation for integrating the data into an SQL database, enabling streamlined queries and analyses.

Data Integration

Once the data was cleaned and the schema finalized, we merged the datasets from PSU and OSU into a single, unified table. Using SQLite, we created a database named study_abroad_psu_osu and loaded the combined dataset into it. The integration process allowed us to consolidate program data from both universities while preserving each program's source and details.

SQL was then used to validate the integration and test the schema. For instance, queries were run to verify data integrity, such as checking programs offered in specific cities or regions. This ensured that the integration process was successful and the data was ready for analysis

Results

SQL Queries and Analysis

After integrating the cleaned and unified dataset into an SQLite database named study_abroad_psu_osu, we conducted several SQL queries to analyze the study abroad programs and draw meaningful insights. The queries allowed us to explore various aspects of the data, such as program distribution by city, region, and country, as well as comparisons between PSU and OSU. Below, we highlight some key queries and their results, which are visualized in the accompanying screenshots.

To begin, we queried the database for all programs offered in London. This query highlighted six programs, with four provided by PSU and two by OSU. These programs are all based in the United Kingdom, within the European region, showcasing the universities' strong presence in this global city.

```
SELECT * FROM study_abroad_psu_osu WHERE city = 'London';
```

Index	Program.Name	City	Country	Region	College
186	IES: London, Queen Mary University	London	United Kingdom	Europe	The Pennsylvania State University
187	IES: London, Study London	London	United Kingdom	Europe	The Pennsylvania State University
188	IES: London, University College London	London	United Kingdom	Europe	The Pennsylvania State University
227	London: FSU, Theatre Academy London	London	United Kingdom	Europe	The Pennsylvania State University
307	London, United Kingdom: Adulthood and Families ...	London	United Kingdom	Europe	The Ohio State University
310	London, United Kingdom: Theatre Studies Program ...	London	United Kingdom	Europe	The Ohio State University

Next, we filtered the database to retrieve all programs located in Asia. The results demonstrate the variety of programs offered in countries such as Cambodia, Singapore, Thailand, Japan, and South Korea. This query emphasizes the diverse opportunities available for students interested in studying in this region.

```
SELECT * FROM study_abroad_psu_osu WHERE region = 'Asia';
```

Index	Program.Name	City	Country	Region	College
246	SFS: Cambodia	Siem Reap	Cambodia	Asia	The Pennsylvania State University
255	Singapore: National University of Singapore...	Singapore	Singapore	Asia	The Pennsylvania State University
266	University of Minnesota LAC: Thailand, Studies in ...	Chiang Mai	Thailand	Asia	The Pennsylvania State University
303	Korea/Japan: Summer Study Abroad for Design Majors ...	multiple cities	Japan	Asia	The Ohio State University
304	Korea/Japan: Summer Study Abroad for Design Majors ...	multiple cities	South Korea	Asia	The Ohio State University
322	Seoul, Korea: Sociology and Culture in Korea (Maymester)	multiple cities	South Korea	Asia	The Ohio State University

Focusing further, we analyzed programs specifically offered in Japan. The database revealed entries for cities like Tokyo, Mito, and Sendai, as well as multi-city programs. This query

highlights the range of study abroad opportunities available in Japan, catering to different academic and cultural interests.

```
SELECT * FROM study_abroad_psu_osu WHERE country = 'Japan';
```

Index	Program.Name	City	Country	Region	College
212	IES: Tokyo, Language and Culture	Tokyo	Japan	Asia	The Pennsylvania State University
233	Mito: Ibaraki University...	Mito	Japan	Asia	The Pennsylvania State University
303	Korea/Japan: Summer Study Abroad for Design Majors ...	multiple cities	Japan	Asia	The Ohio State University
371	CIEE: Open Campus	multiple cities	Japan	multiple regions	The Ohio State University
393	CRCC Asia: Tokyo Internship (Summer)	Tokyo	Japan	Asia	The Ohio State University
401	Dept ENGR & EMS: Sendai, Tohoku University...	Sendai	Japan	Asia	The Ohio State University

To understand the overall distribution of study abroad programs, we aggregated the data by region. The results indicate that Europe hosts the majority of programs (300), followed by Asia (64) and Latin America (34). This insight underscores Europe's popularity as a study abroad destination and provides a high-level view of program availability across regions.

```
SELECT region, COUNT(*) as program_count
FROM study_abroad_psu_osu
GROUP BY region;
```

Region	program_count
Africa	30
Asia	64
Europe	300
Latin America	34
Middle East	14
North America	2

Finally, we examined the distribution of programs by college and country. The analysis shows that OSU has a significant number of programs in the United Kingdom, while PSU offers several in Australia and Argentina. This comparative view highlights the unique regional focuses of the two universities

```
SELECT College, country, COUNT(*) as program_count
FROM study_abroad_psu_osu
GROUP BY College, country;
```

College	Country	program_count
The Ohio State University	Turks and Caicos Islands	1
The Ohio State University	United Kingdom	30
The Ohio State University	Various	1
The Ohio State University	Virtual	2
The Pennsylvania State University	Argentina	7
The Pennsylvania State University	Australia	11

Key Insights

These queries demonstrate the power and flexibility of the mediated schema and SQL integration. The database enables detailed analysis of program distributions, comparisons between colleges, and exploration of location-based opportunities. By integrating the datasets into a single query-friendly structure, we created a tool that provides valuable insights into study abroad opportunities for both Penn State and Ohio State.

Discussion

This project addresses several important challenges in international education, particularly in the areas of accessibility, standardization, and the broader impact of integrated databases. By consolidating information from multiple sources into a single unified database, we have significantly reduced the time and effort students need to research study abroad opportunities. Instead of navigating through multiple websites with varying structures, students can now access a centralized source for program details.

The use of a mediated schema ensured that program information was presented in a standardized format. This not only facilitates easier comparison and analysis but also enhances the ability to extract meaningful insights. Universities can use this structured data to better understand their offerings and identify areas for improvement or expansion.

Looking ahead, the database has immense potential for growth. It could be expanded to include study abroad programs from more universities, fostering greater collaboration and cultural exchange opportunities. Additionally, creating a user-friendly and interactive interface for students and administrators to query the database would make the system more

accessible and impactful. This feature could include filters, search capabilities, and visualizations to further simplify the process of exploring programs.

Conclusion

This project demonstrates the feasibility and value of creating a unified database for study abroad programs. By integrating data from PSU and OSU, we have streamlined the research process for students and provided universities with a powerful analytical tool. The system simplifies the discovery of study abroad opportunities, empowering students to make informed decisions and promoting international education.

Beyond its immediate benefits, this project lays the foundation for a more interconnected and efficient approach to managing study abroad programs. As global learning continues to grow in importance within higher education, tools like this database will play a critical role in fostering engagement and enhancing the reach of international programs. Ultimately, this database not only promotes accessibility and standardization but also encourages greater participation in global learning experiences, enriching the academic and cultural journeys of students worldwide