

Semantic Data Integration: University Domain Use Case

Team 21: Sanaz Bayat, Shirin Shoghli, Shahrzad Torabi, Navid Hadipour Limouei

May 28, 2025

1 Week 01 (05.05 – 11.05)

Datasets selection

In the first week, we selected “**Universities**” as our use case for semantic data integration. Our goal was to unify diverse data sources related to higher education institutions and develop a mediated schema that can answer meaningful competency questions. We chose the topic of “Universities” because it is easy to relate to and has many available datasets from different sources like rankings, campus information, and student data. These datasets often have similar information but in different formats, which makes it a good example for learning how to bring everything together in one clear and useful structure. Also, since we are students ourselves, it felt like a familiar and interesting topic to work with.

We collected **three datasets** representing this domain:

- **Colleges and Universities Dataset**
- **National University Rankings**
- **University and College Campuses**

All datasets were uploaded to the FIM GitLab repository as .xlsx files.

2 Week 02 (12.05 – 18.05)

Competency Questions and Conjunctive Queries

In the second week, we developed ten competency questions that an end-user or analyst might ask about the university domain. The questions were designed to capture key aspects such as location, ranking, tuition, and contacts, and were then formalized using Conjunctive Query.

Ten Competency Questions are Uploaded in The GIT folder of the group. Also described CQ in Conjunctive Query form.

1. Which universities are located in the US?

Q(Name) :-

CollegesUniversities(Name, Country), Country = "US"

2. What is the URL for the University of UCLA?

Q(UniversityName, Url) :-

CollegeUniversityCampuses(UniversityName, Url), UniversityName = "UCLA"

3. What is the contact name for the University of ULM?

Q(ContactName) :-

CollegeUniversityCampuses(UniversityName, ContactName), UniversityName = "ULM"

4. Which universities are located in Pulaski County?

Q(UniversityName) :-

CollegeUniversityCampuses(UniversityName, County), County = "Pulaski"

5. Which universities have ranking > 10 and are located in New York?

Q(UniversityName) :-

CollegesUniversities(Name,...), CollegeUniversityCampuses(UnieiversityName),
NationalUniversitiesRankings(Uni, Rank, Location), Name=UnievrstityName, UnievrstityName=Uni,
Rank > 10, Location = "New York"

6. How many students are enrolled in bachelor programs at Harvard and Stanford?

Q(Uni, UndergradEnrollment) :-

NationalUniversitiesRankings(Uni, UndergradEnrollment), Uni="Harvard" or Uni="Stanford"

7. Which universities are located in Latitude="125217.7396" and Longitude="34.75930829"?

Q(Name):-

CollegesUniversities(Name, Latitude, Longitude), Latitude="125217.7396",
Longitude="34.75930829"

8. What is the tuition and fees for Georgetown University?

Q(Uni, TuitionFees) :-

NationalUniversitiesRankings(Uni, TuitionFees), Uni = "Georgetown University"

9. What is the telephone number for American College of Healthcare Sciences?

Q(Telephone) :-

CollegesUniversities(Name, Telephone), Name = "American College of Healthcare Sciences"

10. Show the number of students and employees at the University of Passau.

Q(StudentCount, EmployeeCount) :-

CollegesUniversities(Name, StudentCount), CollegeUniversityCampuses(UniversityName,
EmployeeCount), Name=UniversityName, Name="University of Passau"

3 Week 03 (19.05 – 25.05)

Mediated Schema

During week 3, we analyzed the schemas of the three source datasets and designed a mediated schema that can unify the different structures into a common representation. The mediated schema provides a simplified and normalized view, enabling query answering across all data sources.

Source Schemas Overview:

CollegesUniversities('latitude', 'Longitude', 'NAME', 'Address', 'address2', 'City', 'ZIPCode', 'Telephone', 'StudentCount', 'County', 'Country', 'OtherInfo', 'EstablishDate', 'Website', 'TotalEnrollment', 'HasDormitory', 'CapacityDormitory', 'TotalEmployee')

CollegeUniversityCampuses('UniversityName', 'Url', 'Address', 'ZIP', 'County', 'ContactName', 'EmployeeCount', AverageGPA, NumberOfGraduatedStudents, AnnualScholar, ShipGranted, CountStudentWorkInUni, NumberOfDisabledStudent)

NationalUniversitiesRankings('Uni', 'Location', 'Rank', 'Description', 'TuitionFees', 'UndergradEnrollment', 'Date', 'Country', 'ContactName', 'City', 'PostalCode')

Mediated Schema:

CollegesUniversities(NAME, Address, address2, City, ZIPCode, County, Country, Website, TotalEmployee)

Student(UniversityName, AverageGPA, NumberOfGraduatedStudents, AnnualScholar, ShipGranted, CountStudentWorkInUni, NumberOfDisabledStudent)

Employee(Uni, EmployeeCount, Location, Country, City, PostalCode)