CSCI 586 Project Report
Fall 2013

Group #6

Viterbi Semantic Guide [P3]

Group Members:
Anamika Mukherji
Lohith Ram
Shubham Sharma
Varad Pathak
1. Problem Statement

The purpose of this project is to create a curriculum and research guide for the Viterbi School of Engineering. It is common that there arises semantic heterogeneity when different entities try to model the same/similar set of data according to their priorities. This causes non-interoperability of data. We are trying to solve this problem through the example of Viterbi Semantic Guide. We are creating a single interface for a person to get the data of different departments of Viterbi, which may have inherent semantic heterogeneity.

2. Scope of the Project

We have decided to limit our scope to Computer Science, Electrical Engineering, and Civil Engineering department. We are collecting the data of these departments in following areas:

- Courses
- Faculty
- PhD Students
- PostDoc Students
- Research Institutes, Labs and groups

3. Steps Involved

The diagrammatic representation of the steps involved is shown in figure 1. As shown in the figure we collected the data cleaned it and created the ontology for it, independently. Out of the four team members, three of us worked on construction of individual department ontologies to ensure that there is some inherent semantic heterogeneity in the collected data and corresponding ontology. The fourth team member created meta-ontology and integrated the data and ensured the interoperability of the same is achieved.
1) Data Collection and Cleaning:

We first collected the data by scraping it from the web sites of each department. This was done using a Google Chrome extension called Scraper. The scraped data was exported to Excel file format. The collected data was cleaned to remove some common semantic heterogeneity related to string formatting. This was done manually using Karma and Excel.

2) Ontology Creation:

The second step involved was the creation of ontologies for each of the department. The tool used for creation of ontologies was Protégé. Ontologies for each department were created independently to ensure the heterogeneity. All the classes, data properties and object properties were defined. The consistency of the ontology was verified by using Hermit Logic Reasoner provided by Protégé.

3) Data Modeling:

We modeled the data based on the created ontologies using Karma, a tool developed by Information Integration group at USC-ISI. The Semantic classes were mapped to the corresponding columns in the data and their relations were established using the data and object properties.

4) RDF Generation of Data:

We then exported the RDF generated by Karma in Turtle format (.ttl format) which was then imported in Protégé for further processing.
5) Create Meta-Ontology:

A Viterbi Meta-Ontology was created which has all the generic classes and data and object properties. It was taken care that meta-ontology has all possible classes and properties that are present in the individual department ontologies.

4. Design and Implementation
   a) Ontology:

To resolve the issue of semantic heterogeneity there are two basic approaches namely, Top-down and Bottom-up approach. Each approach has some pros and cons. We discussed and debated over both the approaches and finally used a hybrid approach which suits best for our problem statement.

**Top-Down Approach:** Create a single standardized ontology (Top Level - Viterbi) first, then ask individual entities to create their own local/private ontology which would follow the standard ontology.

**Bottom-Up Approach:** Let individual private parties create their ontologies (Departments) and then create Top-level ontology (Viterbi) which has all the required classes and properties.

**Hybrid Approach:** Construct the meta-ontology first with basic classes and properties. Then construct individual private ontologies which can use the top level ontology classes and properties. This approach avoids the problem of classifying the entities in an incorrect class because of unavailability of the proper class structure, especially for the entities which are outside the scope of the private ontology. In the end the meta-ontology is refined to include all the possible classes of the bottom level ontologies.

We created individual department ontologies for Computer Science, Electrical Engineering and Civil Engineering Department. These ontologies were then mapped to the corresponding classes and properties of meta/federated ontology of Viterbi school of Engineering using subclass and subproperty relationship.
Civil Department Ontology Hierarchy:

Following diagrams show various classes of individual department ontologies and their relationships with complete Viterbi Ontology:

**Fig 2. Research Labs and Institutes**
Fig 3. Department and Courses

Fig 4. Research Areas
Fig 5. Student and Faculty Classes and their relations for Civil Department
b) Modeling of data:

The following screenshot shows how the data was modeled using the Karma Tool. As shown in the figure the boxes represent the instances of the classes and arrows represent the data and object property mapping. The blue labeled property, advisor in this case, is an object property and the rest are data property. The star mark on a name suggests that the field is used to create a unique URI for the entity. In this case we have assumed that the name of person is unique as there was no other field which was universally available for each person.

Fig 6. Data modeling Using Karma
5. Properties:
   a. Object Properties

The relations between different entities were created using following set of object properties

b. Data Properties

Data Properties were created add details about the individual entities
c. Merged Ontology:

All the data and ontologies were merged using Protégé. Classes were mapped using sub-class relationship with top-level ontology. Data and Object properties were mapped using sub-property relationship.

We used “Hermit Reasoner” to verify the correctness and consistency of the ontologies. The Reasoner was also used to find all the possible inferences. All the inferred triples were exported to form the final merged ontology. The size of the complete data after getting all the possible inferences was 49145 lines!!

6. Tools used:

a) Protégé:

Protégé is a free, open source ontology editor and knowledge-base framework. The Protégé platform supports modeling ontologies via a web client or a desktop client. Protégé ontologies can be developed in a variety of formats including OWL, RDF(S), and XML Schema.

Protégé is based on Java, is extensible, and provides a plug-and-play environment that makes it a flexible base for rapid prototyping and application development.

b) Karma:

Karma is an information integration tool that enables users to quickly and easily integrate data from a variety of data sources including databases, spreadsheets, and delimited text files, XML, JSON, KML and Web APIs. Users integrate information by modeling it according to ontology of their choice using a graphical user interface that automates much of the process. Karma learns to recognize the mapping of data to ontology classes and then uses the ontology to propose a model that ties together these classes. Users then interact with the system to adjust the automatically generated model. During this process, users can transform the data as needed to normalize data expressed in different formats and to restructure it. Once the model is complete, users can published the integrated data as RDF or store it in a database.

c) Scraper:

Scraper is a Google Chrome extension which uses X-Path queries to collect data from web pages. It stores the result in the form of Excel sheets in Google drive. All the data was collected by writing simple X-path queries.
7. Sample SPARQL queries:

Query 1: Advisor

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX viterbi: <http://www.viterbi.edu#>

SELECT (str(?Advisor) as ?Advisor) (str(?Student) as ?Student) ?EmailOfStudent
WHERE {
  ?faculty viterbi:advisorOf ?student;
  viterbi:name ?Advisor.
  FILTER (str(?Advisor) = "Dennis McLeod").
  ?student viterbi:name ?Student;
  viterbi:mbox ?EmailOfStudent.
}

ORDER BY ?Student

Results:

<table>
<thead>
<tr>
<th>Advisor</th>
<th>Student</th>
<th>EmailOfStudent</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Dennis McLeod&quot;</td>
<td>&quot;Ali-Ghanmi Rami&quot;</td>
<td>&quot;<a href="mailto:alghanmi@usc.edu">alghanmi@usc.edu</a>&quot;</td>
</tr>
<tr>
<td>&quot;Dennis McLeod&quot;</td>
<td>&quot;Jun Jong-Eun&quot;</td>
<td>&quot;<a href="mailto:jongeun@usc.edu">jongeun@usc.edu</a>&quot;</td>
</tr>
<tr>
<td>&quot;Dennis McLeod&quot;</td>
<td>&quot;Lim Jong Woo&quot;</td>
<td>&quot;<a href="mailto:jonglim@usc.edu">jonglim@usc.edu</a>&quot;</td>
</tr>
<tr>
<td>&quot;Dennis McLeod&quot;</td>
<td>&quot;Xu Kaijian&quot;</td>
<td>&quot;<a href="mailto:lccu@usc.edu">lccu@usc.edu</a>&quot;</td>
</tr>
</tbody>
</table>
Query 2: Units

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX viterbi: <http://www.viterbi.edu#>

SELECT (str(?id) as ?CourseID) (str(?name) as ?Course) (str(?InstructorName) as ?Instructor) ?Prerequisite ?Corequisite ?units WHERE {
  ?subject a viterbi:course;
  viterbi:id ?id;
  viterbi:units ?units;
  viterbi:name ?name.
  OPTIONAL {?subject viterbi:prerequisite ?Prerequisite}
  OPTIONAL {?subject viterbi:corequisite ?Corequisite}
  OPTIONAL {?subject viterbi:instructor ?instr.
    ?instr viterbi:name ?InstructorName.
  }
  FILTER(?units>3 || ?units<3)
} ORDER BY DESC(?units) ?subject

Results:

<table>
<thead>
<tr>
<th>CourseID</th>
<th>Course</th>
<th>Instructor</th>
<th>Prerequisite</th>
<th>Corequisite</th>
<th>units</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;CE 467L&quot;</td>
<td>&quot;Geotechnical Engineering&quot;</td>
<td>Amy Rechmannacher</td>
<td></td>
<td></td>
<td>&quot;4&quot;</td>
</tr>
<tr>
<td>&quot;CE 653&quot;</td>
<td>&quot;Urban Transportation Planning&quot;</td>
<td>Sandipan Bhattacharjee</td>
<td></td>
<td></td>
<td>&quot;4&quot;</td>
</tr>
<tr>
<td>&quot;CSCI 104L&quot;</td>
<td>&quot;Data Structures and Object Oriented Programming&quot;</td>
<td>Mark Redekopp</td>
<td>CSCI103</td>
<td>CSCI170</td>
<td>&quot;4&quot;</td>
</tr>
<tr>
<td>&quot;CSCI 106L&quot;</td>
<td>&quot;Data Structures and Object Oriented Programming&quot;</td>
<td>Mark Redekopp</td>
<td>CSCI109</td>
<td>CSCI170</td>
<td>&quot;4&quot;</td>
</tr>
<tr>
<td>&quot;CSCI 170&quot;</td>
<td>&quot;Discrete Methods in Computer Science&quot;</td>
<td>Aaron Cote</td>
<td>CSCI103</td>
<td></td>
<td>&quot;4&quot;</td>
</tr>
<tr>
<td>&quot;CSCI 201L&quot;</td>
<td>&quot;Principles of Software Development&quot;</td>
<td>David Wilczynski</td>
<td>CSCI109</td>
<td></td>
<td>&quot;4&quot;</td>
</tr>
<tr>
<td>&quot;CSCI 270&quot;</td>
<td>&quot;Introduction to Algorithms and Data Structures&quot;</td>
<td>Aaron Cote</td>
<td>CSCI170</td>
<td></td>
<td>&quot;4&quot;</td>
</tr>
<tr>
<td>&quot;CSCI 280&quot;</td>
<td>&quot;Introduction to Algorithms and Data Structures&quot;</td>
<td>Aaron Cote</td>
<td>CSCI104</td>
<td></td>
<td>&quot;4&quot;</td>
</tr>
<tr>
<td>&quot;CSCI 360&quot;</td>
<td>&quot;Video Game Programming&quot;</td>
<td>Sanjay Madhavi</td>
<td>ITP355</td>
<td></td>
<td>&quot;4&quot;</td>
</tr>
<tr>
<td>&quot;CSCI 380&quot;</td>
<td>&quot;Video Game Programming&quot;</td>
<td>Sanjay Madhavi</td>
<td>CSCI104</td>
<td></td>
<td>&quot;4&quot;</td>
</tr>
<tr>
<td>&quot;CSCI 402&quot;</td>
<td>&quot;Operating Systems&quot;</td>
<td>William Cheng</td>
<td>CSCI455</td>
<td></td>
<td>&quot;4&quot;</td>
</tr>
<tr>
<td>&quot;CSCI 402&quot;</td>
<td>&quot;Operating Systems&quot;</td>
<td>William Cheng</td>
<td>EE352</td>
<td></td>
<td>&quot;4&quot;</td>
</tr>
</tbody>
</table>
Query 3: Research Area

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX viterbi: <http://www.viterbi.edu#>
PREFIX fn: <http://www.w3.org/2005/xpath-functions#>

SELECT (str(?ResearchArea) as ?Work) (str(?Faculty) as ?Researcher) (str(?Phone) as ?Phone) (str(?Office) as ?Office) WHERE {
  ?subject a viterbi:faculty;
    viterbi:researchFocus ?Research.
  ?subject viterbi:name ?Faculty;
    viterbi:phone ?Phone;
  FILTER (str(fn:lower-case(?ResearchArea)) = "machine learning").
}

ORDER BY ?Researcher

Results:

<table>
<thead>
<tr>
<th>Work</th>
<th>Researcher</th>
<th>Phone</th>
<th>Office</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Machine learning&quot;</td>
<td>&quot;Anton Leski&quot;</td>
<td>&quot;(210) 448-210&quot;</td>
<td>&quot;ICT&quot;</td>
</tr>
<tr>
<td>&quot;Machine learning&quot;</td>
<td>&quot;Craig Knoblock&quot;</td>
<td>&quot;(310) 448-8768&quot;</td>
<td>&quot;BST&quot;</td>
</tr>
<tr>
<td>&quot;Machine learning&quot;</td>
<td>&quot;Fei Sha&quot;</td>
<td>&quot;(213) 740-5924&quot;</td>
<td>&quot;RTH 403&quot;</td>
</tr>
<tr>
<td>&quot;Machine learning&quot;</td>
<td>&quot;Louis-Philippe Morency&quot;</td>
<td>&quot;(310) 448-5323&quot;</td>
<td>&quot;ICT&quot;</td>
</tr>
<tr>
<td>&quot;Machine Learning&quot;</td>
<td>&quot;Lorne Soibelman&quot;</td>
<td>&quot;(213) 740-0609&quot;</td>
<td>&quot;KAP 268A&quot;</td>
</tr>
<tr>
<td>&quot;Machine learning&quot;</td>
<td>&quot;Shrikant Narayanan&quot;</td>
<td>&quot;(213) 740-6432&quot;</td>
<td>&quot;EEB 430&quot;</td>
</tr>
<tr>
<td>&quot;Machine learning&quot;</td>
<td>&quot;Steve Chien&quot;</td>
<td>&quot;(512) 393-5320&quot;</td>
<td>&quot;JPL&quot;</td>
</tr>
<tr>
<td>&quot;Machine learning&quot;</td>
<td>&quot;Wei-Min Shen&quot;</td>
<td>&quot;(310) 448-8710&quot;</td>
<td>&quot;BST&quot;</td>
</tr>
<tr>
<td>&quot;Machine learning&quot;</td>
<td>&quot;Yu-Han Chang&quot;</td>
<td>&quot;(210) 448-8494&quot;</td>
<td>&quot;BST&quot;</td>
</tr>
</tbody>
</table>
Query 4: Research Institutes

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX viterbi: <http://www.viterbi.edu#>
PREFIX fn: <http://www.w3.org/2005/xpath-functions#>

SELECT (str(?Name) as ?Name) (str(?WebPage) as ?URL)
WHERE {
  ?subject a viterbi:organization;
  viterbi:name ?Name;
}

ORDER BY ?subject

Results:

<table>
<thead>
<tr>
<th>Name</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Aerosol Laboratory”</td>
<td>“<a href="http://www.usc.edu/dept/civil_eng/aerosol/%E2%80%9D">http://www.usc.edu/dept/civil_eng/aerosol/”</a></td>
</tr>
<tr>
<td>“Earthquake Engineering – Strong Motion Group”</td>
<td>“<a href="http://www.usc.edu/dept/civil_eng/Earthquake_eng/%E2%80%9D">http://www.usc.edu/dept/civil_eng/Earthquake_eng/”</a></td>
</tr>
<tr>
<td>“Foundation for Cross-Connection Control and Hydraulic Research”</td>
<td>“<a href="http://www.usc.edu/dept/fccchyr/%E2%80%9D">http://www.usc.edu/dept/fccchyr/”</a></td>
</tr>
<tr>
<td>“Hewlett Packard Geotechnical Laboratory”</td>
<td>“<a href="http://gees.usc.edu/soilab/%E2%80%9D">http://gees.usc.edu/soilab/”</a></td>
</tr>
<tr>
<td>“Leroy-Crandall Geotechnical Laboratory”</td>
<td>“<a href="http://www.usc.edu/dept/civil_eng/structural_lab/sc_struct.html%E2%80%9D">http://www.usc.edu/dept/civil_eng/structural_lab/sc_struct.html”</a></td>
</tr>
<tr>
<td>“Structural Research Laboratory”</td>
<td>“<a href="http://www.usc.edu/dept/civil_eng/structural/welcome.html/%E2%80%9D">http://www.usc.edu/dept/civil_eng/structural/welcome.html/”</a></td>
</tr>
<tr>
<td>“Center for Systems and Software Engineering”</td>
<td>“<a href="http://cse.usc.edu/%E2%80%9D">http://cse.usc.edu/”</a></td>
</tr>
<tr>
<td>“Collaboratory for Advanced Computing and Simulations”</td>
<td>“<a href="http://cacs.usc.edu/%E2%80%9D">http://cacs.usc.edu/”</a></td>
</tr>
<tr>
<td>“Computational Behavior Group”</td>
<td>“<a href="http://cb.usc.edu/%E2%80%9D">http://cb.usc.edu/”</a></td>
</tr>
<tr>
<td>“Computational Learning and Motor Control Laboratory”</td>
<td>“<a href="http://www-dlm.usc.edu/%E2%80%9D">http://www-dlm.usc.edu/”</a></td>
</tr>
<tr>
<td>“Computational Linguistics”</td>
<td>“<a href="http://cluesc.usc.edu/%E2%80%9D">http://cluesc.usc.edu/”</a></td>
</tr>
</tbody>
</table>
References: