CSCI 548
Information Integration on the Web
Spring 2012

Instructors: Prof. Craig Knoblock (knoblock@isi.edu)
Prof. Pedro Szekely (szekely@isi.edu)
Prof. Jose-Luis Ambite (ambite@isi.edu)

Meeting Time: Tuesday and Thursday 3:30-4:50pm

Location: ZHS 163

Office Hours:

Professor Knoblock
- Immediately before or after class
- Or by appointment (ISI 922 or by phone: 310-448-8786)

Professor Szekely
Immediately before or after class
Or by appointment (ISI 921 or by phone 310-448-8641)

Professor Ambite
- Immediate before or after class
- Or by appointment (ISI 935 or by phone: 310-448-8472)

Teaching Assistant/Grader:

Mohsen Taheriyan (mohsen@isi.edu)

Course Web Page: USC Blackboard (blackboard.usc.edu)

This course will focus on the basic foundations and techniques in Information Extraction and Integration. There has been a great deal of interest and research on this topic and the course will cover the research and tools for addressing the technical problems. The topics covered will include data integration techniques, machine learning techniques for information extraction and wrapper construction, high-performance query execution systems based on streaming dataflow, constraint-based integration systems, approaches to record linkage for resolving naming inconsistencies across sites, and the challenges of accessing and integrating information from online social networking sites.

The class will be run as a lecture course with lots student participation and hands-on experience. As an integral part of the course each student will develop and build an integrated Web application using the research and tools covered in the class.

Prerequisites:

CSCI561 -- Introduction to AI
CSCI585 – Database Systems

Recommended Course:

CSCI571— Web Technologies

Grading:

Course project -- 50%
Homeworks – 10%
Quizzes – 20%
Final Exam -- 20%

Books: There is no required textbook. We will read technical papers on each topic.
Lab: There is no lab for this course. Students should contact the instructor if they do not have access to a computer where they can install their own software.
Class Project: The course project this year will be to build a system to compete in the ESWC 2012 AI Mashup Challenge (https://sites.google.com/site/aimashup12/). You are not required to submit your project, but task and deadlines will still be used.

Course Syllabus and Schedule

- **January 10**
  - **Topic: Introduction (Professors Knoblock, Szekely, and Ambite)**
- **January 12**
  - **Topic: Semantic Web (Professor Ambite)**
- **January 17**
  - **Topic: Semantic Web (Professor Ambite)**
- **January 19**
  - **Topic: Linked Data (Professor Szekely)**
- **January 24**
  - **Topic: Linked Data (Professor Szekely)**
- **January 26**
  - **Topic: Linked Services (Mohsen Taheriyan & Professor Knoblock)**
- **January 31**
  - **Topic: Geospatial Data Integration (Professor Knoblock)**
  -

- **February 2**
  - **Topic: Mashups (Professor Szekely)**

- **February 7**
  - **Topic: Mashups (Professor Szekely)**

- **February 9**
  - **Topic: Data Integration (Professor Ambite)**

- **February 14**
  - **Topic: Data Integration (Professor Ambite)**

- **February 16**
  - **Topic: Record Linkage (Professor Knoblock)**

- **February 21**
  - **Topic: Record Linkage (Professor Knoblock)**

- **February 23**
  - **Topic: Schema Mapping (Professor Knoblock)**

- **February 28**
  - **Topic: Source Modeling (Professor Knoblock)**

- **March 1**
  - **Topic: Information Extraction (Professor Szekely)**

- **March 6**
  - **Topic: Information Extraction (Professor Szekely)**

- **March 8**
  - **TBD – Guest Lecture**

- **March 12-16**
  - **Spring recess!**

- **March 20**
  - **Topic: Data Cleaning (Professor Szekely)**

- **March 22**
  - **Topic: Constraint Integration (Professor Ambite)**

- **March 27**
  - **Topic: Wrapper Learning (Professor Knoblock)**

- **March 29**
  - **Topic: Wrapper Generation (Professor Knoblock)**

- **April 3**
  - **Topic: Ontology-based Data Integration (Professor Ambite)**

- **April 5**
  - **Topic: Data Integration Under Constraints (Professor Ambite)**

- **April 10**
  - **Topic: Dataflow Execution (Professor Knoblock)**

- **April 12**
  - **Topic: Dataflow optimization (Professor Knoblock)**

- **April 17**
  - **Topic: Intellectual Property (Professor Knoblock)**

- **April 19**
  - **Project Presentations**

- **April 24**
  - **Project Presentations**

- **April 26**
  - **Course Review (Professors Knoblock, Szekely, and Ambite)**

- **Final Exam (Tuesday, May 8, 2-4pm)**
  - **Location: Classroom**