

# CSCI 548

## Information Integration on the Web

### Spring 2013

**Instructors:** Prof. Jose-Luis Ambite (ambite@isi.edu)

Prof. Pedro Szekely (pszekely@isi.edu )

**Meeting Time:** Monday and Wednesday 3:30-4:50pm

**Location:** THH 114

**Office Hours:** Immediately before or after class, or by appointment

**Teaching Assistant/Grader:** George Konstantinidis (konstant@usc.edu)

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

This course will focus on foundations and techniques for Information Extraction, Modeling and Integration. Topics covered include semantic web (RDF, OWL, SPARQL), linked data and services, mash-ups, theory of data integration, schema mappings, record/entity linkage, data cleaning, source modeling, and information extraction. The class will be run as a lecture course with significant hands-on experience. Students will work on in 2-3 person groups to develop integrated Web applications using the research and tools covered in the class.

**Prerequisites:**

CSCI561 -- Introduction to AI

CSCI585 – Database System

**Recommended Course:**

CSCI571— Web Technologies

**Grading:**

Course project -- 40%

Homeworks – 15%

Quizzes – 20%

Final Exam -- 25%

**Books:** There is no required textbook. We will read technical papers on each topic. There is a recommended textbook: *Principles of Data Integration*, by AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. Morgan Kaufmann 2012.

**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:** Students will work on in 2-3 person groups to develop integrated Web applications using the research and tools covered in the class.

**Academic Integrity:** All homeworks, quizzes, and tests must be solved and written independently. Students who violate University standards of academic integrity are subject to disciplinary sanctions, including failure in the course and suspension from the University. Since dishonesty in any form harms the individual, other students, and the University, policies on academic integrity will be strictly enforced. We expect you to familiarize yourself with the Academic Integrity guidelines found in the current SCampus. Violations of the Student Conduct Code will be filed with the Office of Student Conduct, and appropriate sanctions will be given.

## Course Syllabus and Schedule

Date	Topic	Instructor
Jan 14	<b>Introduction and course overview</b>	JLA, PS
	<b>Semantic Web</b>	PS
Jan 16	▪ RDF, graph data model	
Jan 23	▪ RDFS, inference	PS
Jan 28	▪ SPARQL query language	PS
Jan 30	▪ Linked Data, common vocabularies/ontologies	PS
Feb 4	▪ OWL2: Description Logics, Inference	JLA
Feb 6	▪ OWL2 Profiles: QL, EL, RL	JLA
Feb 11	▪ Common semantic web sources (dbpedia, data.gov, ...)	PS
Feb 13	▪ Linked Services	MT
Feb 20	<b>Geospatial Data Integration</b>	CK
Feb 25	<b>Mashups</b>	PS
	<b>Data Integration Fundamentals</b>	
Feb 27	▪ Database theory basics: queries, query containment, Datalog	JLA
	▪ Data Warehousing, ETL	JLA
Mar 4	▪ Query unfolding (Global-as-View)	JLA
Mar 6	▪ Answering queries using views (Local-as-View)	JLA, GK
	<b>Spring break</b>	
Mar 11	▪ Data integration under constraints, binding patterns	JLA
Mar 13	▪ Ontology-based data access and integration	JLA

	<b>Mappings, Transformations and Modeling</b>	
Mar 25	▪ String Matching	PS
Mar 27	▪ Schema Mapping	JLA
Apr 1	▪ RDF mappings	PS
Apr 3	▪ Record Linkage	PS
Apr 8	▪ Data Cleaning	PS
Apr 10	▪ Semi-automatic source modeling (modeling for the rest of us)	PS
Apr 15	▪ Automatic Source Modeling	JLA
	<b>Information Extraction</b>	
Apr 17	▪ Unstructured: Extracting entities and relations from text	ZK
Apr 22	▪ Semistructured data: Wrapper Learning	JLA
Apr 24	<b>Project Presentations</b>	JLA, PS
Apr 29	<b>Project Presentations</b>	JLA, PS
May 1	<b>Course Review</b>	JLA, PS
May 10	<b>Final Exam (2-4pm)</b>	JLA, PS

Instructors:

JLA: Prof. Jose Luis Ambite

PS: Prof. Pedro Szekely

ZK: Prof. Zornitsa Kozareva

CK: Prof. Craig Knoblock

GK: Mr. George Konstantinidis

MT: Mr. Mohsen Taheriyani