

CSCI 548

Information Integration on the Web

Fall 2013

Instructors: Prof. Craig Knoblock (knoblock@isi.edu)

Prof. Pedro Szekely (pszekely@isi.edu)

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

Location: VKC 100

Teaching Assistant: Bo Wu (bowu@isi.edu)

Course Grader: Varad Pathak (vspathak@usc.edu)

Office Hours: Professor Knoblock: Monday & Wednesday 2:30-3:20 AHF B55D

Professor Szekely: Monday & Wednesday 4:50-6:00 max After class in VKC

Bo Wu: Wednesdays 1-3pm SAL 211

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 in 2 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 – 20%

Quizzes – 20%

Final Exam -- 20%, Dec 16, 2-4pm (Check for conflicts!)

Books: There is one required textbook: *Principles of Data Integration*, by AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. Morgan Kaufmann 2012. The book is available online from the USC library at:

<<http://proquest.safaribooksonline.com.libproxy.usc.edu/book/databases/9780124160446>>.

We will also read technical papers on each topic.

Class Project: Students will work on in two person groups to develop integrated Web applications using the research and tools covered in the class.

Quizzes: There will be a quiz every Monday unless it is a holiday, in which case the quiz will be on Wednesday. The quiz will have two parts. One part with conceptual questions that test your understanding of the topics covered in the previous week (the required readings for the two previous lectures). The other part with easy questions that test whether you read the required reading for the day of the quiz.

Academic Integrity: All homework, 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
Aug 26	▪ Introduction and Course Overview	CK, PS
Aug 28	▪ RDF, graph data model	PS
Sep 4	▪ RDFS, inference	PS
Sep 9	▪ SPARQL query language	PS
Sep 11	▪ Linked Data, common vocabularies/ontologies	PS
Sep 16	▪ Karma, Semi-automatic source modeling	PS
Sep 18	▪ Common semantic web sources (dbpedia, data.gov, ...), REST	PS
Sep 23	▪ Mashups	CK
Sep 25	▪ Mashups: Yahoo Pipes & YQL	CK
Sep 30	▪ Record Linkage – String Matching	PS

Oct 2	▪ Record Linkage – Record Matching	CK
Oct 7	▪ Unstructured: Extracting entities and relations from text	ZK
Oct 9	▪ Unstructured: Extracting entities and relations from text	ZK
Oct 14	▪ Data cleaning	BW
Oct 16	▪ Automatic Source Modeling	CK
Oct 21	▪ RDF mapping tools	PS
Oct 23	▪ Linked Services	MT
Oct 28	▪ Database theory basics: queries, query containment, Datalog	CK
Oct 30	▪ Answering queries using views (Global-as-View, Local-as-View)	CK
Nov 4	▪ OWL2: Description Logics, Inference	PS
Nov 6	▪ OWL2 Profiles: QL, EL, RL	PS
Nov 11	▪ Schema Mapping	CK
Nov 13	▪ Semi-structured Data: Wrapper Generation	CK
Nov 18	▪ Semi-structured Data: Wrapper Learning	CK
Nov 20	▪ Intellectual Property	CK
Nov 25	▪ Project Presentations	CK, PS, BW
Dec 2	▪ Project Presentations	CK, PS, BW
Dec 4	▪ Project Presentations	CK, PS, BW
Dec 16	▪ Final Exam (2-4pm)	CK, PS, BW

Instructors:

CK: Prof. Craig Knoblock

PS: Prof. Pedro Szekely

ZK: Prof. Zornitsa Kozareva

BW: Mr. Bo Wu

MT: Mr. Mohsen Taheriyani