INFO/CS 4302
Web Information Systems

FT 2012
Week 1: Course Introduction
Who We Are - Instructors

Bernhard Haslhofer
bh392@cornell.edu
Office hours:
TUE / THU 1:30 - 3:00
Availability:
Oct, Nov, Dec 2012

Theresa Velden
tav6@cornell.edu
Office hours:
TUE / THU 1:30 - 3:00
Availability:
Sept, Nov, Dec 2012
Who We Are - TAs

Changchen He
ch627@cornell.edu
Office hours:
WED 4:30 - 6:00

Syed Ishtiaque Ahmed
sa738@cornell.edu
Office hours:
MON / FRI 5:00 - 6:30
Course Website / Piazza

http://www.infosci.cornell.edu/Courses/info4302/2012fa/

https://piazza.com/cornell/fall2012/infocs4302

#info4302
Where you can find us

301 College Avenue
1st office on the left
Open space in IS
Our plan for today

Group-based class activity

What is this course about?

Review of course syllabus

Questions
My Web, Your Web... the web from your perspective

• Form groups of 4 with your immediate neighbours
• Within the next 10 minutes, take turns to introduce yourself to your group and tell your classmates about
  1. What is your first memory of using the web, e.g. how old were you, what were you doing, what device were you using…?
  2. How has the web and your usage of the web evolved since then?
• Take note of similarities and differences in the experiences and preferences represented in your group. Anything that surprised you? Be ready to report back to the class
Our plan for today

Group-based class activity

What is this course about?

Review of course syllabus

Questions
What is this course about?

Web Information Systems
What is this course about?

the "Web" = "World Wide Web" = "WWW" =
"A system of interlinked documents accessed via the Internet"

Web Information Systems

It is now almost two decades since the Web has been invented. Initially motivated by the need to link documents across computer systems to support large collaborations in high energy physics, the Web evolved rapidly. It has reshaped the notion of information systems, and changed our social interactions and cultural development. Decentralization and openness were fundamental design principles in the Web Architecture and enabled the creation of large, community-driven information spaces such as Wikipedia. In recent years, these principles were adopted for publishing structured data on the Web, resulting in efforts such as Linked Data, schema.org, or the Open Graph protocol.

This course will introduce you to technologies for building data-centric information systems on the World Wide Web, show the practical applications of such systems, and discuss their design and their social and policy context by examining cross-cutting issues such as citizen science, data journalism and open government. Course work involves lectures and readings as well as weekly homework assignments, and a semester-long project in which the students demonstrate their expertise in building data-centric Web information systems.
What is this course about?

Web Information Systems

Data

4302, 75, 10, 2

raw, unorganized facts

process, organize, structure, contextualize

Information

course number: 4302
registered students: 75
number of HWs: 10
instructors: 2

"useful" data
What is this course about?

Web Architecture

Web Information Systems

Data representation

Standards

Openness and decentralization

Tools and Frameworks
What is this course about?

Web APIs

Open Data

Global Data Networks

data-centric

Web Information Systems

Publishing Data on the Web

Using Data from the Web
Google Maps API Web Services

This document discusses the Maps API Web Services, a collection of HTTP interfaces to Google services providing geographic data for your maps applications. This guide serves only to introduce the web services and host information common to all of the different services. Individual documentation for each service is located below:

- Directions API
- Distance Matrix API
- Elevation API
- Geocoding API
- Places API

The remainder of this guide discusses techniques for setting up web service requests and parsing the responses. For particular documentation for each service, however, you must consult the appropriate documentation.

Table of Contents

1. What is a Web Service?
2. SSL Access
3. Tracking Usage with the sensor Parameter
4. Building a Valid URL
5. Processing Responses
   a. Processing XML with XPath
   b. Processing JSON with Javascript
data.nytimes.com

For the last 150 years, The New York Times has maintained one of the most authoritative news vocabularies ever developed. In 2009, we began to publish this vocabulary as linked open data.

The Data

As of 13 January 2010, The New York Times has published approximately 10,000 subject headings as linked open data under a CC BY license. We provide both RDF documents and a human-friendly HTML versions. The table below gives a breakdown of the various tag types and mapping strategies on data.nytimes.com.

<table>
<thead>
<tr>
<th>Type</th>
<th>Manually Mapped Tags</th>
<th>Automatically Mapped Tags</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>People</td>
<td>4,978</td>
<td>0</td>
<td>4,978</td>
</tr>
<tr>
<td>Organizations</td>
<td>1,489</td>
<td>1,592</td>
<td>3,081</td>
</tr>
<tr>
<td>Locations</td>
<td>1,910</td>
<td>0</td>
<td>1,910</td>
</tr>
<tr>
<td>Descriptors</td>
<td>498</td>
<td>0</td>
<td>498</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10,467</td>
</tr>
</tbody>
</table>

Browse individual data records:

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

SKOS Files

Download all of the data records as SKOS Files.

- People
- Organizations
- Locations
- Subject Descriptors
Welcome to data.nature.com – the NPG Linked Data Platform

The NPG Linked Data Platform provides access to datasets from NPG published as linked data and made available through SPARQL services. Two different interfaces are provided, a form interface for interactive queries and a service endpoint for remote queries:

/query - form interface (non-streaming)
/sparql - service endpoint (streaming)

Full documentation, demos and data snapshots for downloading are available on the nature.com developers portal.

Triple count: **279,885,352** (279.8 million)

Note that a live updating process is actively adding in triples to the datasets as new articles are published.

What is Available?

NPG is making available a number of datasets for public access as linked data. These datasets include data about articles published by NPG as well as the NPG product and subject ontologies. All datasets are registered on the [Data Hub](http://data.nature.com).

The datasets can be queried with SPARQL and snapshots are also available for downloading.

Data Organization

The datasets are organized by graphs with one graph maintained per object type. A directory graph maintains descriptions for each of the individual graphs with class and property counts, and vocabularies used. Note that an NPG vocabulary has been used for object type properties as well as for certain data type properties:

See: [http://ns.nature.com/terms/](http://ns.nature.com/terms/)
Cornell University contains seven undergraduate colleges plus the College of Veterinary Medicine, the Law School, the Samuel Curtis Johnson Graduate ... Google review

144 East Ave, Ithaca, NY 14850 (607) 254-4836

Admissions
Admissions - Financial Aid - Apply Now - Undergraduate - ...

Jobs at Cornell
Searchable listing provided by the Office of Human Resources ...

College of Veterinary Medicine
Admissions - For Pet Owners - Feline Health Center - Library

Graduate School
The Graduate School is organized into ninety-four major fields of ...

Legal Information Institute
Cornell Law School - U.S. Code - Supreme Court - State - Wex

Visiting
Walking Tours - Getting Here - Life in Ithaca - ...

News for cornell university
Cornell University to install nets on bridges in bid to prevent suicides
Fox News - 2 days ago
Cornell University in New York plans to begin installing nets on several

Cornell University is an American private Ivy League research university located in Ithaca, New York, United States. Wikipedia

Motto: "I would found an institution where any person can find instruction in any study."

Nickname: Big Red
Address: 144 East Ave, Ithaca, NY 14850
Mascot: Big Red Bear
Acceptance rate: 16.2% (2012)
Colors: Carnelian, White
What is this course about?

http://developers.mystartup.com
http://mystartup.com/apis
http://data.mystartup.com
What is this course about?

Cross cutting Issues

- Citizen Science
- Web Information Systems
- Data Journalism
- ...
What is this course not about?

Web site design (INFO 1300)

Web application development (INFO 2300)

Search and Retrieval (INFO 4300)
Our plan for today

Group-based class activity

What is this course about?

Review of course syllabus

Questions
# INFO/CS 4302 Syllabus

<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Day</th>
<th>Lecture</th>
<th>Homework</th>
<th>Project</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8/23</td>
<td>TH</td>
<td>Course Introduction</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8/26</td>
<td>SU</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>8/28</td>
<td>TU</td>
<td>Technical Foundations of The Internet and The Web</td>
<td></td>
<td>release hw1</td>
</tr>
<tr>
<td></td>
<td>8/30</td>
<td>TH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9/2</td>
<td>SU</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>9/4</td>
<td>TU</td>
<td>The Web as an Internet Application</td>
<td></td>
<td>hw1 due &amp; release hw2</td>
</tr>
<tr>
<td></td>
<td>9/6</td>
<td>TH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9/9</td>
<td>SU</td>
<td></td>
<td></td>
<td>hw2 due &amp; release hw3</td>
</tr>
<tr>
<td>4</td>
<td>9/11</td>
<td>TU</td>
<td>Semi-Structured Data: XML and XML Manipulation</td>
<td></td>
<td>hw3 due &amp; release hw4</td>
</tr>
<tr>
<td></td>
<td>9/13</td>
<td>TH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9/16</td>
<td>SU</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>9/18</td>
<td>TU</td>
<td>Semi-Structured Data: JSON and other formats</td>
<td></td>
<td>hw4 due &amp; release hw5</td>
</tr>
<tr>
<td></td>
<td>9/20</td>
<td>TH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9/23</td>
<td>SU</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>9/25</td>
<td>TU</td>
<td>Cross Cutting Issues</td>
<td></td>
<td>hw5 due &amp; release hw6</td>
</tr>
<tr>
<td></td>
<td>9/27</td>
<td>TH</td>
<td>Recap</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9/30</td>
<td>SU</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week</td>
<td>Date</td>
<td>Day</td>
<td>Topic</td>
<td>Notes</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
<td>------</td>
<td>--------------------------------------------</td>
<td>------------------------------</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>10/2</td>
<td>TU</td>
<td>RESTful Webservice APIs</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10/4</td>
<td>TH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10/5</td>
<td>FR</td>
<td></td>
<td>hw6 due (Friday!)</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>10/9</td>
<td>TU</td>
<td>(no class) Fall Break</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10/11</td>
<td>TH</td>
<td>Global Data Networks Intro</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10/14</td>
<td>SU</td>
<td></td>
<td>release hw7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>project proposal due</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>10/16</td>
<td>TU</td>
<td>Linked Data Technologies</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10/18</td>
<td>TH</td>
<td></td>
<td>hw7 due &amp; release hw8</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>10/23</td>
<td>TU</td>
<td>Publishing Structured Web Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10/25</td>
<td>TH</td>
<td></td>
<td>hw8 due &amp; release hw9</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>10/30</td>
<td>TU</td>
<td>Making Use of Structured Web Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11/1</td>
<td>TH</td>
<td></td>
<td>hw9 due</td>
<td></td>
</tr>
<tr>
<td></td>
<td>11/4</td>
<td>SU</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
INFO/CS 4302 Syllabus

<table>
<thead>
<tr>
<th></th>
<th>Date</th>
<th>Day</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>11/6</td>
<td>TU</td>
<td>Student Projects</td>
</tr>
<tr>
<td></td>
<td>11/8</td>
<td>TH</td>
<td></td>
</tr>
<tr>
<td></td>
<td>11/11</td>
<td>SU</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>11/13</td>
<td>TU</td>
<td>Cross Cutting Issues</td>
</tr>
<tr>
<td></td>
<td>11/15</td>
<td>TH</td>
<td>Cross Cutting Issues</td>
</tr>
<tr>
<td></td>
<td>11/18</td>
<td>SU</td>
<td>release hw10</td>
</tr>
<tr>
<td>14</td>
<td>11/20</td>
<td>TU</td>
<td>Cross Cutting Issues</td>
</tr>
<tr>
<td></td>
<td>11/22</td>
<td>TH</td>
<td>(no class) Thanksgiving Recess</td>
</tr>
<tr>
<td></td>
<td>11/25</td>
<td>SU</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>11/27</td>
<td>TU</td>
<td>Cross Cutting Issues</td>
</tr>
<tr>
<td></td>
<td>11/29</td>
<td>TH</td>
<td>Recap</td>
</tr>
</tbody>
</table>

Homeworks: due Sun night @ 11:59PM (CMS)
(exception: hw 6 due already Fri night @ 11:59 PM)

Requisite: CS 2110 or similar (object oriented programming, Java or python)
Group Projects

Groups of 3 students.
Design a web information systems of your choice.
Detailed requirements on course website.

Milestones:
Kick-off in class: early September
Project proposal due: October 14th
Intermediary project presentation: November 6th + 8th
Final project presentation: December 13th from 5-9 pm
(make up: December 5th from 7-9:30pm, you need to let us know by August 31st)
Final project report due: December 13th @ 11:59pm
Participation & Support

Piazza:
• your questions on course content
• announcement of useful resources
• formation of teams, search for project partners
• your answers to challenges

Office Hours:
• offered every weekday @ 301 College Avenue
Email policy

• Send all questions about course content via piazza (not in personal email to instructors)
  o Quick turn-around
  o Others learn too, reducing overhead
Grading

45%  
Homework Assignments

10%  
Project proposal

10%  
Mid Term Project Presentation

25%  
Final Project Presentation & Report

10%  
Participation
More course information:

• Academic integrity
  o Group assignments are meant to be worked on in groups. They are not meant to be done by one person without review and passed off as the group's work.
  o Individual assignments are meant to be worked on alone.
  o In both cases, looking things up and getting ideas from other sources is okay, if you cite it. Plagiarism (copying of others' work and attempting to pass it off as your own) is not.

• Lecture slides (posted after lecture)

• Instructions for submitting homework & code

http://www.infosci.cornell.edu/Courses/info4302/2011fa
Next week

Week 2: **History and Architecture of the Internet**

This Sunday (8/26): release of homework 1

Questions on readings:

1. V. Bush. *As We May Think*; Atlantic Monthly; July 1945.
2. T. Berners-Lee et al. *Creating a Science of the Web*
3. T. Heath, C. Bizer
   *Linked Data: Evolving the Web into a Global Data Space, Chapters 1-3*