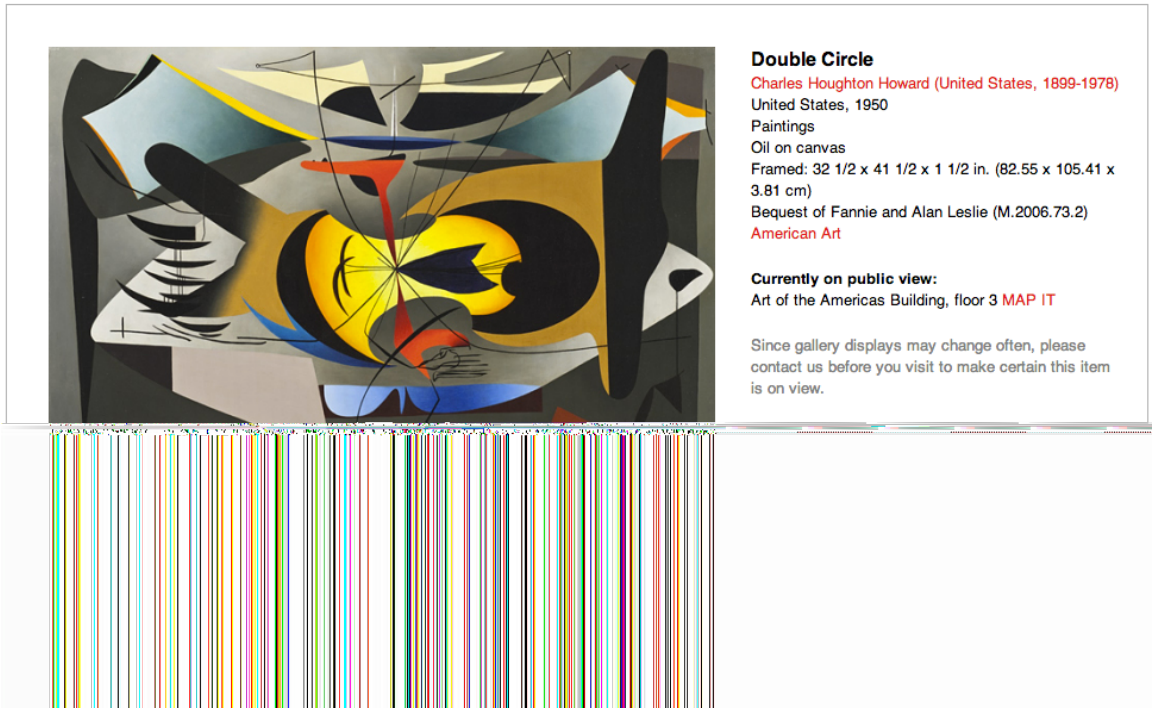
CS548 Homework 1

Due Date: 2014/09/02 @11:59pm on Blackboard.

This homework must be done individually. You can ask others for help with the tools, but the submitted homework needs to be your own work.

The objective in this homework is for you to write a wrapper/scrapper to get data from a museum Web site. The software that you write must collect data about paintings and their authors. For example, below is an image of a page from the LACMA museum[1]:



A scrapper for this web site should extract the URL of the image, and all the data about the artwork and artist (title, author name and biographical data, location, category, material, extent, etc.) Different museums provide different data, so the fields you extract may differ from the ones in this example. We want to see that you extract all the data you can reasonably extract.

You will be using the dataset you produce in this homework in future homeworks, so it is a good idea to do this well so that you have good data for those homeworks.

[1] http://collections.lacma.org/node/184481

Tasks to Do

Select a Web Site

Every student must select a different museum site. You can choose any museum you want, but you must extract data about American Art or European art after 1700.

Record your Web site in the following Google Spreadsheet after making sure that nobody else selected the same Web site (you cannot use LACMA).

Select a Scrapping Tool

Many scrapping/wrapper tools are available on the Internet. You can use any tool you like, you can use one of several available libraries and interactive tools. Search in Google to find your favorite tool. For example, stackoverflow has a list to get you started: http://stackoverflow.com/questions/2861/options-for-html-scraping

Construct Your Scrapper

Your scrapper must navigate the Web site you select automatically, and produce a dataset of at least 300 artworks. You are not allowed to manually provide your scrapper a list of 300 URLs.

The output of your scrapper should be one or several JSON or XML files with all the data in the pages. You need to extract data about the artworks and their authors.

What to Hand In

1. 2-page description of your work including:
    a. At the top of the page include the following statement and sign it. "I, _____, declare that the submitted work is original and adheres to all University policies and acknowledge the consequences that may result from a violation of those rules"
    b. Web site with screenshots of the types of pages
    c. One or two sample records that you extract showing the fields that you extract
    d. A one paragraph description of the most difficult technical challenge you had to overcome
    e. A one paragraph description listing the tool you chose and a justification for choosing it
    f. A one paragraph description listing the fields in the page you didn't extract and a justification for not extracting them
2. Zip file of your datasets
3. Zip file of your software