Introduction to Digital Humanities: Scraping Edition
February 11, 2015
Guest speakers: Liam Andrew and Desi Gonzalez

In this session, we will introduce *data scraping*, or the process of grabbing data from unstructured documents (such as web pages) in order to extract and store it. Scraping can be done by automatic scripts at massive scales, or manually as a one-off task.

Many online tools can assist you with scraping and extracting data from the web. We will introduce **Kimono**, which can let you scrape and crawl basic websites without having to do any programming (even if you're a programmer, this can save you lots of time!)

Kimono: https://www.kimonolabs.com/

1. Register for Kimono
2. Download Kimono (Chrome extension recommended, you can also use a bookmarklet on other browsers)

Here are three sample data sets to practice using Kimono. Pick whichever one you like.

1. Historical maps: http://bit.ly/1uGheF4
    a. *Basic:* make a scraper that first 50 grabs each item's title, maker, title, date, and map type.
    b. *Advanced:* make a crawler that grabs the above and the "note" from the object page.
2. DPLA images of Grace Hopper: http://bit.ly/1zMYFPY
    a. *Basic:* make a scraper that grabs each item's title, image, and source/description.
    b. *Advanced:* make a crawler that fetches each item's title, image, partnering institutions, format, and any additional metadata that might provide insight
3. Journals from the Modernist Journals Project: http://modjourn.org/journals.html
    a. *Basic:* make a scraper that grabs each journal's title, dates, description and image
    b. *Advanced:* make a crawler that fetches each journal's title, dates, image, full description, and list of contents.

Lastly, download and review the data. Is there anything particularly interesting or strange about the results you scraped? Are there any obvious errors or formatting issues? How could these be fixed?

CMS.633 / CMS.833 Digital Humanities

Spring 2015