Information Integration on the Web: Homework 1

Due on January 23, 2015

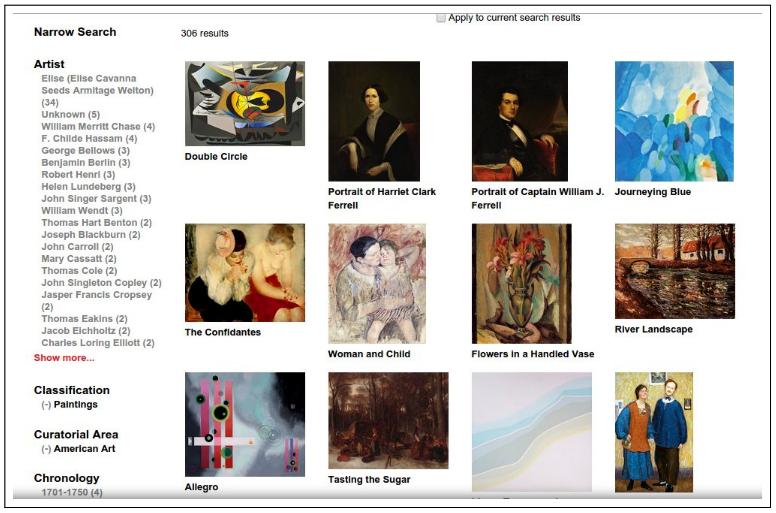
Professor Ambite & Knoblock

Tushar Tiwari

Statement of Integrity

I, Tushar Tiwari, 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.

Screenshots



☐ Apply to current search results

Narrow Search

306 results

Artist
- Elise (Elise Cavanna Seeds Armitage Welton) (34)
- Unknown (5)
- William Merritt Chase (4)
- F. Childe Hassam (4)
- George Bellows (3)
- Benjamin Berlin (3)
- Robert Henri (3)
- Helen Lundeberg (3)
- John Singer Sargent (3)
- William Wendt (3)
- Thomas Hart Benton (2)
- Joseph Blackburn (2)
- John Carroll (2)
- Mary Cassatt (2)
- Thomas Cole (2)
- John Singleton Copley (2)
- Jasper Francis Cropsey (2)
- Thomas Eakins (2)
- Jacob Eichholtz (2)
- Charles Loring Elliott (2)

Show more...

Classification
(-) Paintings

Curatorial Area
(-) American Art

Chronology
1701-1750 (4)

Double Circle

Portrait of Harriet Clark Ferrell

Portrait of Captain William J. Ferrell

Journeying Blue

The Confidantes

Woman and Child

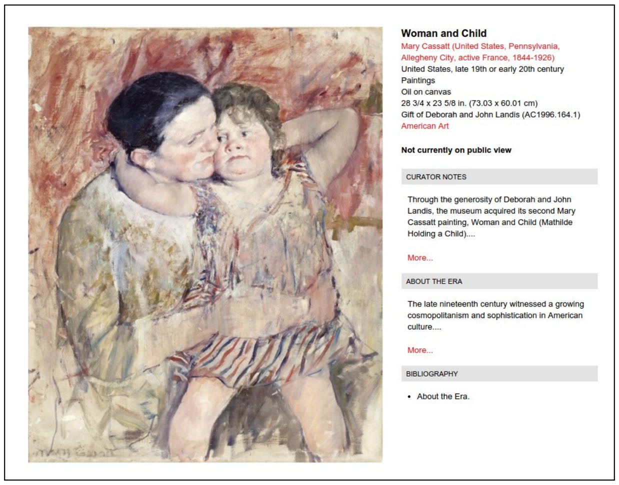
Flowers in a Handled Vase

River Landscape

Allegro

Tasting the Sugar

Figure 1: The search page with the option American Paintings already chosen.
This provides us with a custom url that can be exploited to navigate and then extract all paintings.
`http://collections.lacma.org/search/site/?page=0&f[0]=bm_field_has_image%3Atrue&f[1]=im_field_curatorial_area%3A32&f[2]=im_field_classification%3A22`



Woman and Child
Mary Cassatt (United States, Pennsylvania, Allegheny City, active France, 1844-1926)
United States, late 19th or early 20th century
Paintings
Oil on canvas
28 3/4 x 23 5/8 in. (73.03 x 60.01 cm)
Gift of Deborah and John Landis (AC1996.164.1)
American Art

Not currently on public view

CURATOR NOTES

Through the generosity of Deborah and John Landis, the museum acquired its second Mary Cassatt painting, Woman and Child (Mathilde Holding a Child)....

More...

ABOUT THE ERA

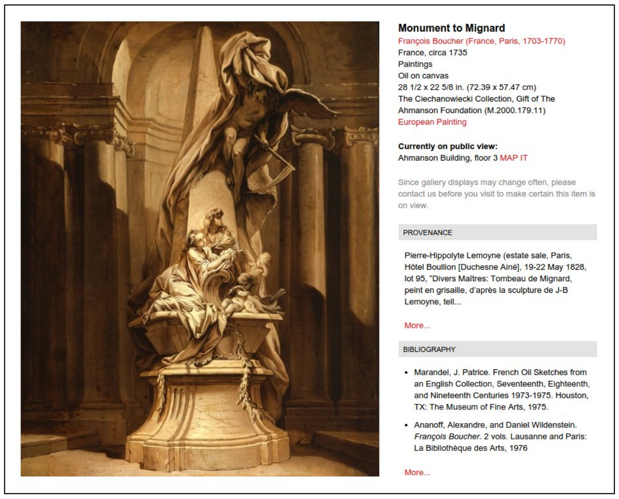
The late nineteenth century witnessed a growing cosmopolitanism and sophistication in American culture....

More...

BIBLIOGRAPHY

- About the Era.

(a) An american painting



Monument to Mignard
François Boucher (France, Paris, 1703-1770)
France, circa 1735
Paintings
Oil on canvas
28 1/2 x 22 5/8 in. (72.39 x 57.47 cm)
The Ciechanowiecki Collection, Gift of The Ahmanson Foundation (M.2000.179.11)
European Painting

Currently on public view:
Ahmanson Building, floor 3 MAP IT

Since gallery displays may change often, please contact us before you visit to make certain this item is on view.

PROVENANCE

Pierre-Hippolyte Lemoyne (estate sale, Paris, Hôtel Boullion [Duchesne Ainé], 19-22 May 1828, lot 95, "Divers Maîtres: Tombeau de Mignard, peint en grisaille, d'après la sculpture de J-B Lemoyne, tell...

More...

BIBLIOGRAPHY

- Marandel, J. Patrice. French Oil Sketches from an English Collection, Seventeenth, Eighteenth, and Nineteenth Centuries 1973-1975. Houston, TX: The Museum of Fine Arts, 1975.
- Ananoff, Alexandre, and Daniel Wildenstein. *François Boucher.* 2 vols. Lausanne and Paris: La Bibliothèque des Arts, 1976

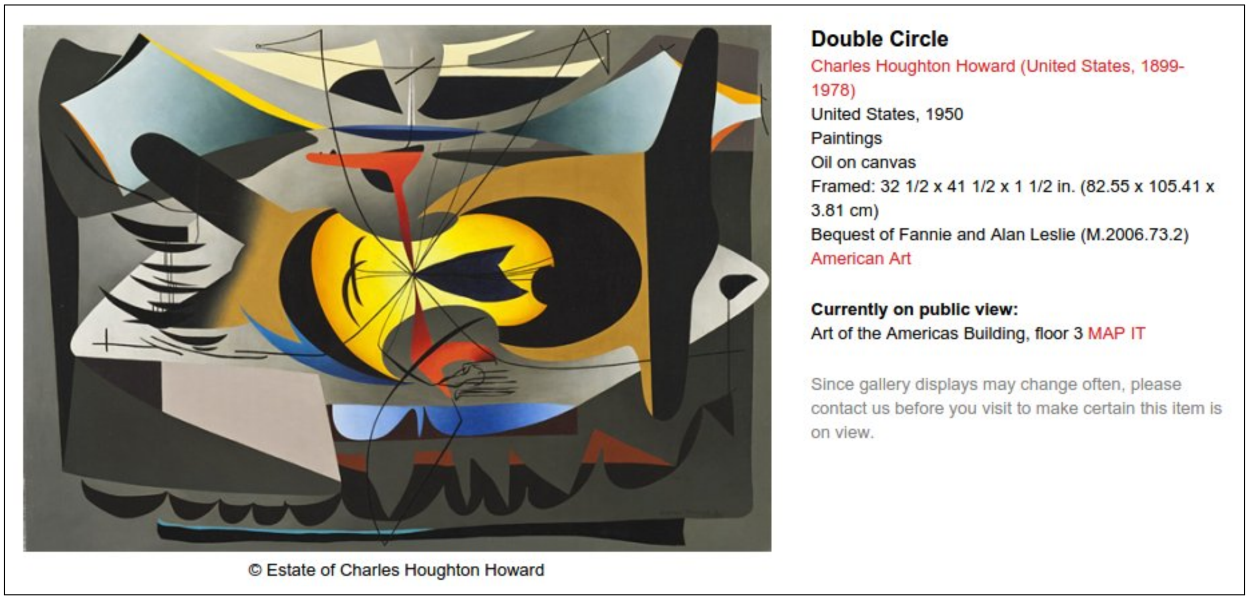
More...

(b) A european painting

Figure 2: Examples of the paintings on the website

Sample Record

Let's look at a sample painting and the output that it produces. For purposes of demonstration the link to this sample painting is hard-coded into the program so that the output produced is only for the painting in consideration.



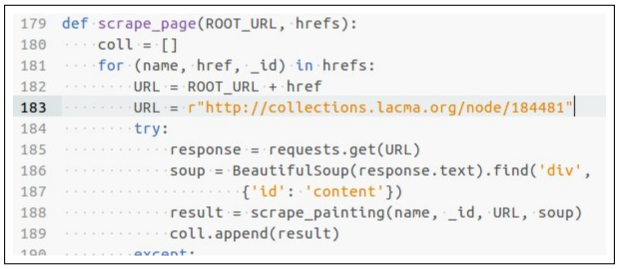
© Estate of Charles Houghton Howard

Double Circle
Charles Houghton Howard (United States, 1899-1978)
United States, 1950
Paintings
Oil on canvas
Framed: 32 1/2 x 41 1/2 x 1 1/2 in. (82.55 x 105.41 x 3.81 cm)
Bequest of Fannie and Alan Leslie (M.2006.73.2)
American Art

Currently on public view:
Art of the Americas Building, floor 3 MAP IT

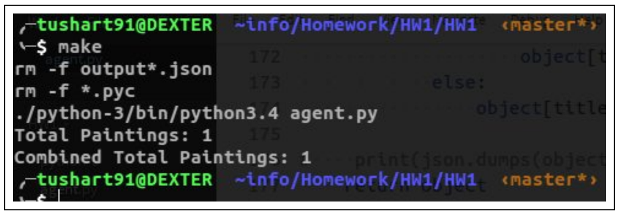
Since gallery displays may change often, please contact us before you visit to make certain this item is on view.

Figure 3: The chosen painting found at `http://collections.lacma.org/node/184481`.



```
def scrape_page(ROOT_URL, hrefs):
    coll = []
    for (name, href, _id) in hrefs:
        URL = ROOT_URL + href
        URL = r"http://collections.lacma.org/node/184481"
        try:
            response = requests.get(URL)
            soup = BeautifulSoup(response.text).find('div',
                    {'id': 'content'})
            result = scrape_painting(name, _id, URL, soup)
            coll.append(result)
```

(a) Hardcoding the url into the code



```
┌─tushart91@DEXTER  ~info/Homework/HW1/HW1  ‹master*›
└─$ make
rm -f output*.json
rm -f *.pyc
./python-3/bin/python3.4 agent.py
Total Paintings: 1
Combined Total Paintings: 1
┌─tushart91@DEXTER  ~info/Homework/HW1/HW1  ‹master*›
└─$
```

(b) Execution of make command

Figure 4: Sample Run

Note:

- The url for the painting is hardcoded only for purposes of demonstration. The usual run will automatically extract image urls from the search page.
- Not all paintings will have all data fields assosciated with them since the data is unavailable for those paintings. These missing fields have been ignored as they serve no purpose except for increasing the file size which is not desirable.

```
[
    {
        "_id": 184481,
        "acquired_from": "Bequest of Fannie and Alan Leslie",
        "acquired_id": "M.2006.73.2",
        "artist": {
            "birthplace": {
                "country": "United States"
            },
            "dates": {
                "born": "1899",
                "died": "1978"
            },
            "name": {
                "first": "Charles",
                "last": "Howard",
                "middle": "Houghton"
            }
        },
        "category": "american",
        "copyright": "Estate of Charles Houghton Howard",
        "created_country": "United States",
        "created_year": "1950",
        "image_link": "http://collections.lacma.org/sites/default/files/
        remote_images/piction/ma-31205657-WEB.jpg",
        "link": "http://collections.lacma.org/node/184481",
        "name": "Double Circle",
        "on_view": "Art of the Americas Building, floor 3",
        "painting_type": "Oil on canvas"
    }
]
```

Figure 5: Output JSON of sample run

Challenges

- Unicode characters are a problem in Python 2 since the default string supports only ascii characters. Due to this certain string operations become very complicated. The problem was oversome by porting the code to Python 3 whose default string supports unicode
- Extracting the artist's details required complex regular expressions since all information pertaining to the artist, like name, date of birth, place of birth, was provided in a single string in the html.

Tools Used

Python with BeautifulSoup

Python is one of the most easiest to use yet powerful languages out there which works cross-platform. Installing and distributing dependencies is also very easy. Beautiful Soup is a Python library designed for quick turnaround projects like screen-scraping. It parses anything you give it, and does the tree traversal stuff for you. You can tell it "Find all the links", or "Find all the links of class externalLink", or "Find all the links whose urls match "foo.com", or "Find the table heading that's got bold text, then give me that text.". Some of the features that make it powerful are:

- Provides a few simple methods and Pythonic idioms for navigating, searching, and modifying a parse tree: a toolkit for dissecting a document and extracting what you need. It doesn't take much code to write an application
- Automatically converts incoming documents to Unicode and outgoing documents to UTF-8. You don't have to think about encodings, unless the document doesn't specify an encoding and Beautiful Soup can't detect one. Then you just have to specify the original encoding.
- Sits on top of popular Python parsers like lxml and html5lib, allowing you to try out different parsing strategies or trade speed for flexibility.

Fields not Extracted

There were no fields that were left untouched. Every field was extracted. But, not all paintings have data for every field. All fields were checked for data. If valid information was present only then it was extracted.