CSCI 548 Homework 1: Construct Your Scrapper

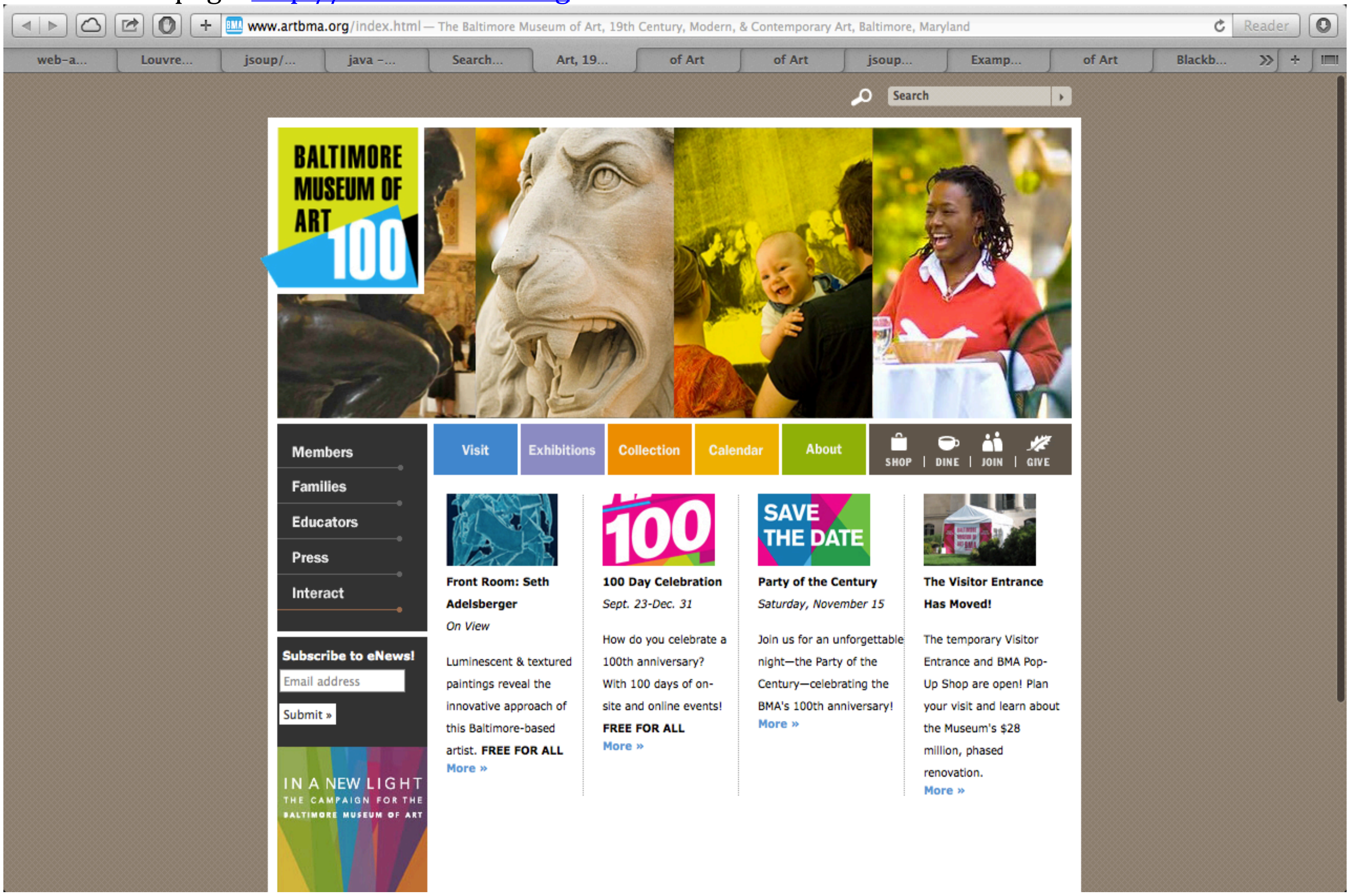
Name: Priya Ankush Kotwal
Email: Kotwal@usc.edu

(Dataset produced comprises of 349 artworks.)

a. "I, Priya Ankush Kotwal, 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. Website used: http://www.artbma.org Baltimore Museum of Art
Screenshots:

1. Homepage: http://www.artbma.org



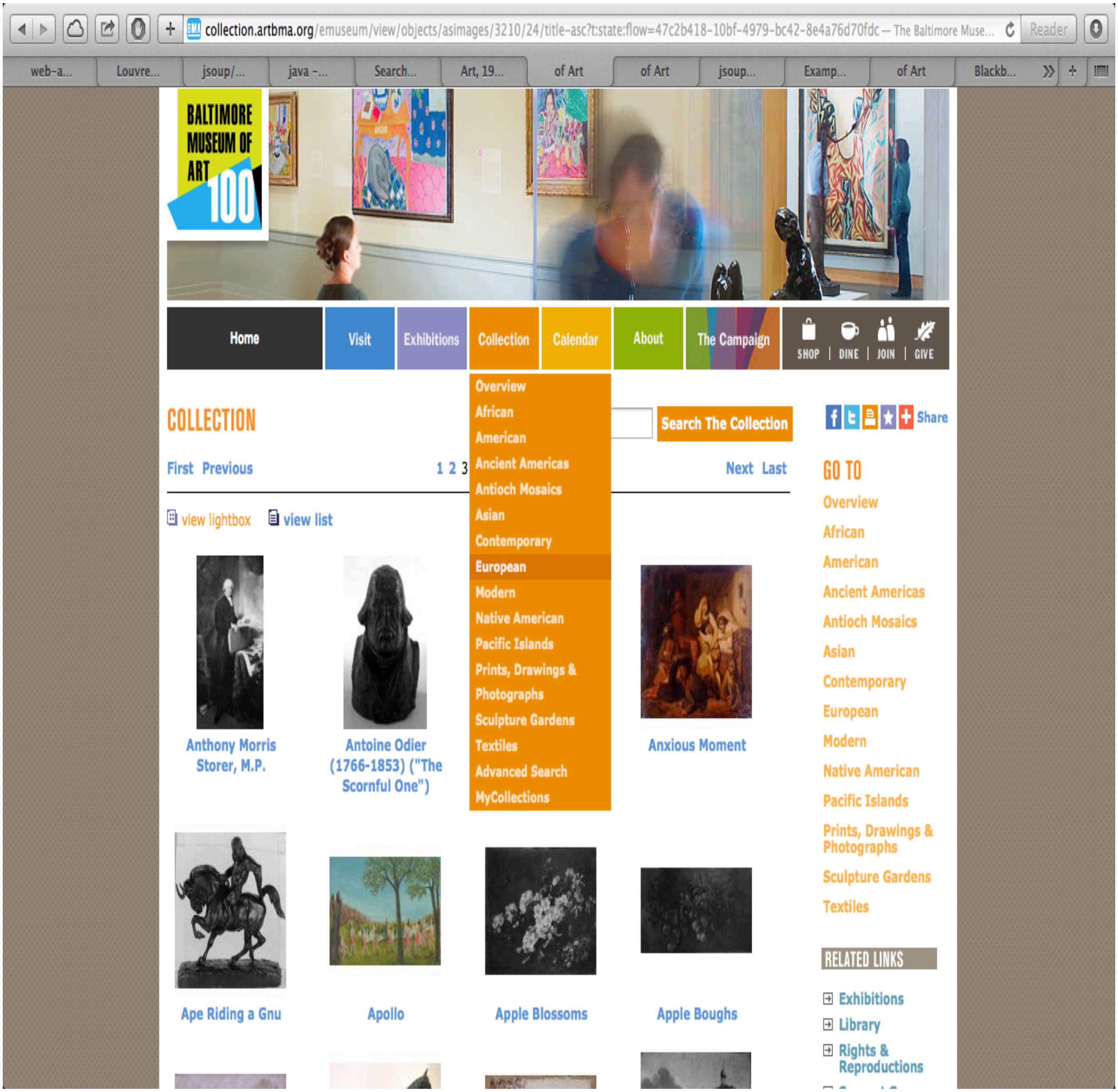
Front Room: Seth Adelsberger
On View
Luminescent & textured paintings reveal the innovative approach of this Baltimore-based artist. **FREE FOR ALL**
More »

100 Day Celebration
Sept. 23-Dec. 31
How do you celebrate a 100th anniversary? With 100 days of on-site and online events!
FREE FOR ALL
More »

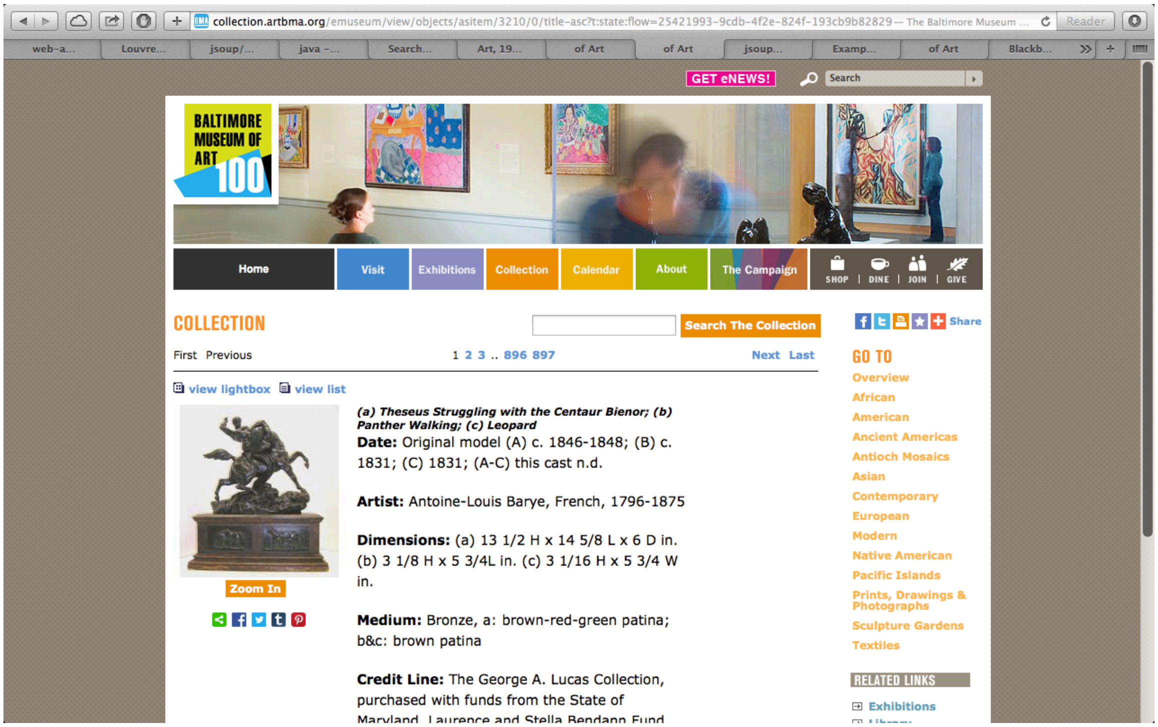
Party of the Century
Saturday, November 15
Join us for an unforgettable night—the Party of the Century—celebrating the BMA's 100th anniversary!
More »

The Visitor Entrance Has Moved!
The temporary Visitor Entrance and BMA Pop-Up Shop are open! Plan your visit and learn about the Museum's $28 million, phased renovation.
More »

2. Collection: European: http://www.artbma.org/collection/european.html
Browse collection(List View):
http://collection.artbma.org/emuseum/view/objects/asimages/3210?t:state:flow=06edea01-17f8-49cd-a8e7-c0a92f273778



3. Detail view of the artwork:
   http://collection.artbma.org/emuseum/view/objects/asitem/3210/0/title-asc?t:state:flow=9a309cee-9326-445f-8655-d1ebbd1cdf6a



(a) Theseus Struggling with the Centaur Bienor; (b) Panther Walking; (c) Leopard
Date: Original model (A) c. 1846-1848; (B) c. 1831; (C) 1831; (A-C) this cast n.d.

Artist: Antoine-Louis Barye, French, 1796-1875

Dimensions: (a) 13 1/2 H x 14 5/8 L x 6 D in. (b) 3 1/8 H x 5 3/4L in. (c) 3 1/16 H x 5 3/4 W in.

Medium: Bronze, a: brown-red-green patina; b&c: brown patina

Credit Line: The George A. Lucas Collection, purchased with funds from the State of Maryland, Laurence and Stella Bendann Fund

c. Sample Records:
   Extracted fields:

   1. Artist
   2. Medium
   3. Dimensions
   4. Date
   5. Object Number
   6. Credit Line
   7. Title
   8. Inserted the field Painting # just for reference in order to keep track of the count.

Sample records: (from artworkDataset.json)

```
{
  "Artwork": [
    {
      "Artist: ": "Antoine-Louis Barye, French, 1796-1875",
      "Medium: ": "Bronze, a: brown-red-green patina; b&amp;c: brown patina",
      "Dimensions: ": "(a) 13 1/2 H x 14 5/8 L x 6 D in. (b) 3 1/8 H x 5 3/4L in. (c) 3 1/16 H x 5 3/4 W
in.",
      "Date": "Original model (A) c. 1846-1848; (B) c. 1831; (C) 1831; (A-C) this cast n.d.",
      "Painting": 1,
      "Object Number: ": "1996.46.41a-c ",
      "Credit Line: ": "The George A. Lucas Collection, purchased with funds from the State of Maryland, Laurence and Stella Bendann Fund, and contributions from individuals, foundations, and corporations throughout the Baltimore community",
      "Title": "(a) Theseus Struggling with the Centaur Bienor; (b) Panther Walking; (c) Leopard"
    },
    {
      "Artist: ": "Th&eacute;ophile-Victor-Emile Lemmens, French, 1821-1867",
      "Medium: ": "Oil on wood panel",
      "Dimensions: ": "6-3/4 x 9-11/16 in. (17.2 x 24.6 cm.); Frame: 15-3/4 x 18-3/4 x 2-3/4 in. (40 x 47.6 x 7 cm.)",
      "Date": "1866",
      "Painting": 2,
      "Object Number: ": "1996.45.168 ",
      "Credit Line: ": "The George A. Lucas Collection, purchased with funds from the State of Maryland, Laurence and Stella Bendann Fund, and contributions from individuals, foundations, and corporations throughout the Baltimore community",
      "Title": "A Lesson for Wrongdoers"
    }
  ]
}
```

d. Description of the most difficult technical challenge in the project

   1. Learning and using a new tool that is jsoup for the scraper and extracting the data from the website. The website contained some tags which were inconsistent for example Artist and Date tags. Some artwork contained the manufacturer tag instead of the artist tag since the artwork was a sculpture and not a painting so had to exclude that artwork. So parsing the data was a challenging task.

e. Description listing the tool selection

   Project Name: ArtMuseumScraper inside the HW1workspace.zip file. Extract the project and import it in your Eclipse IDE. Json and jsoup jars included in the project. Run the artScraper.java file.
   I have limited the count to 349 since there are 897 paintings which would have consumed some time.
   Jsoup is a Java library for working with real-world HTML. It provides a very convenient API for extracting and manipulating data, using the best of DOM, CSS, and jquery-like methods.
   I chose this tool because I am well acquainted with JAVA and have done many projects using this language. I was able to implement the scraper by adding two external jars : the jsoup jar and the json jar. The API's provided by jsoup and json are easy to implement and execute in Java , thus I chose this tool.

f. Description listing the fields in the page that were not extracted and a justification for not extracting them

   The website provided all the fields listed in the assignment and I extracted them complying to the requirements. I did not extract the artwork which contained the field manufacturer since that type of artwork was not a painting and not required in the assignment.