DEPARTMENT OF COMPUTER SCIENCE

MASTER'S PROJECT

Visualizing Thumbnails Of Archived Web Pages

Author:

Surbhi Shankar

Advisor:

Dr. Michele C. Weigle

April 24, 2017

Acknowledgement

I express my gratitude to my project advisor **Dr. Michele C. Weigle**, Associate Professor, Department of Computer Science at Old Dominion University, for being a remarkable source of inspiration to me and for guiding me throughout the project. Love for visualization started during course Information Visualization in Spring 2016 under Dr. Weigle. Visual representations are always powerful and more effective when compared to other methods of representing data. The idea of visualization ideas for this project was proposed by Dr. Weigle and without her creative ideas, detailed guidelines and her depth of knowledge in the field of Visualization, this project would not be complete. She has been very understanding and supportive at every step and hence, I could keep track of my progress and finish the project as intended.

I take this opportunity to thank **Mr. Mathew Kelly**, PhD Candidate, Web Sciences and Digital Libraries lab, Computer Science Department, Old Dominion University, for spending time in helping me understand some of the critical parts of the project which supported my work. Mat Kelly helped me understand the parts of the implementation which is useful for my project. Also, explained the algorithm and the way the output is designed to work.

Finally, my sincere thanks the **Department Of Computer Science** for providing all the required resources and infrastructure to complete this project successfully.

Contents

1 Motivation 3
2 Introduction 4
3 Visualizations 8
4 Sources of Data and Data Cleaning Process 11
5 Insights on the System 14
5.1 Image Slider 14
5.2 Image Grid 14
5.3 Timeline 15
6 Challenges 16
7 Future Enhancements 17
8 References 18

1 Motivation

This project is a part of a bigger idea of visually representing internet archived pages. There are a large collection of archived web pages, which leads to a large number of mementos. So far, the methods used for viewing the mementos collection to easily access them may not be efficient in every situation. Opening each memento to observe the changes that the web page has had, is a tedious job and also very time consuming. To avoid this, we can use many different and innovative visualization methodologies to see the evolution of the webpage over the years. This project has three such major designs implemented. This project also uses images of the archived versions of the webpages which is a very effective method of showing different versions of archived pages and determine the distinct webpages.

Users can understand the dynamic nature of the web and the differences between each memento just by looking at the images. This whole concept was very interesting for me because these visualizations are different from the usual ones which are usually used in every application. I was excited to see how data can be represented differently and what the final product would look like. The archived pages are easily accessible based on their appearance as snapshots of the mementos that are distinct are being used.

2 Introduction

The web pages on the internet are dynamic and the content on the web changes very frequently. The content on the wen is a very important part of our lives, with technology being used everywhere. Excessive competition in all the fields have made it very competitive for websites to be dynamic and hence there may be significant changes in the contents of the webpages. These web pages are periodically stored in the form of Internet archives, which are used for research purposes by humanities scholars and social scientists.

Internet Archive thrives to save as much data of the webpages as possible to help the researchers. Most of web archive interfaces provide simple textual links to the archived versions of the webpage, known as mementos. **Mementos** are a set of archived versions of a webpage which can be explored individually. In simple words, they are archived version of a webpage at a particular time in history which shows how it looked at that time. **TimeMaps** is a list of all the mementos for a webpage. A **thumbnail** is a small image representation of archived pages taken as a snapshot. The TimeMap designs used in this project are all based on thumbnail representation of the archived pages.

Visual aspects are important for scholars and researchers of art history. As they are extending research in the field of born-digital domain, visual results for their searches are more effective and navigable than any current methods of presentations. The Timemap designs used in this project are more user-friendly and helps to access the web archives easily and is faster when compared to other methods.

The idea behind the visualizations in this project is to make it easier to access the web archived pages and also to view the evolution of the webpage. There may be so many archives among which only some are important and significantly used in research. The visualizations presented in this project represent the collections from New York Art Resources and Consortium (NYARC), and Columbia University Libraries (CUL). The NYARC

and CUL have active web archive collections hosted with Internet Archive's Archive-It service.

The goals that this project aims to fulfill are as follows:

- Through visualizations, showing how a single web page has changed over time, without the user having to open every memento separately.
- Reducing the search time required to look for a memento, as images are displayed for quick access.
- Highlighting the temporal and dynamic nature of a website through these visualizations.

Currently, the web archives are presented in ways which may be difficult to search for a particular page that is required. It is difficult to search through every archived page to find what is needed. This process is more time consuming and also the interface may not be user-friendly. Figure 1 and Figure 2 below shows how web archives are made accessible for research purposes, which may be inefficient sometimes. The calendar visualization also has some overlaps in the view which makes it hard to understand the information that it intends to display.

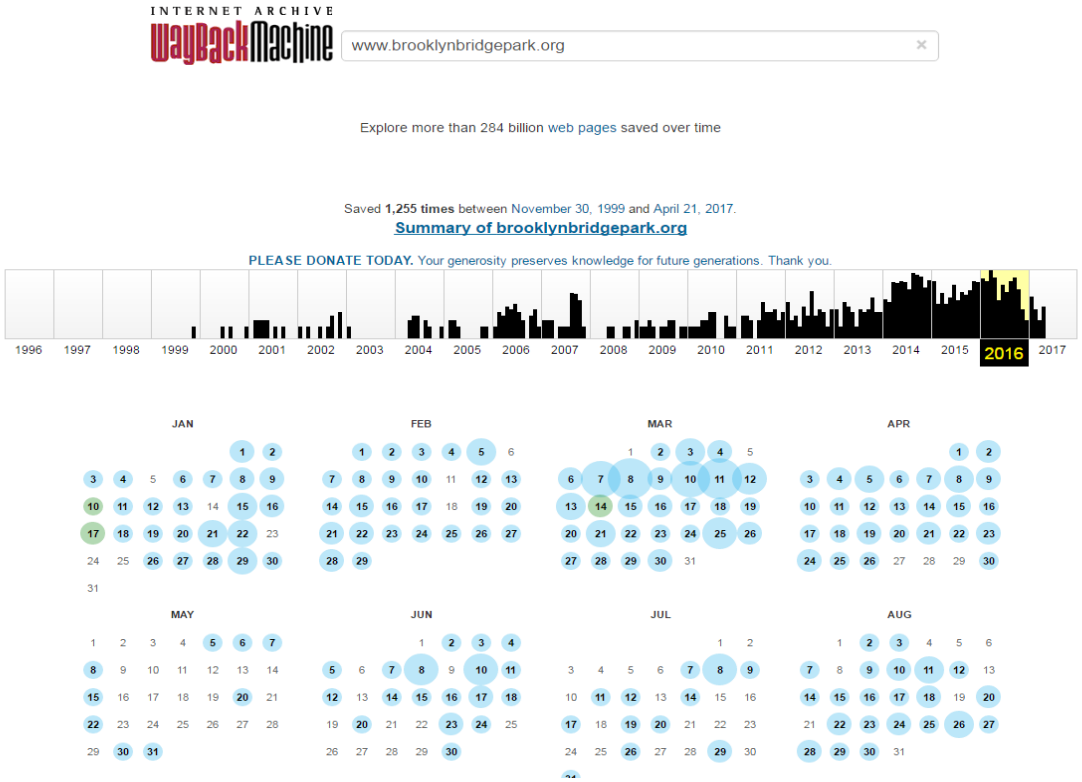

Figure 1: Visualizations on Wayback machine website using calendar view and bar graph.

Figure 2: One of CUL's Archived webpage collections, which needs a user to open every link to see how the website evolved.

This project is based on the archived pages from Archive-It collections which belongs to NYARC and CUL. This project can be used with an implementation of Ahmed AlSum's 2014 ECIR paper titled "Thumbnail Summarization Techniques for Web Archives" [1], which Mr. Mat Kelly and team worked on. This is an implementation for Web Archiving Incentive Program for Columbia University Libraries' grant, "Visualizing Digital Collections of Web Archives". The algorithm implemented crawls through the web pages and creates hash values for each of the archived versions based on how distinct they are from each other. The output would be a list of archived versions which are distinct from each other. All the URLs are dumped into a text file which can be used as a data source for this project and JSON file can be created from it.

Creating a visualization design, which has everything on a single screen is very helpful in terms of specific kinds of research and also as a tool for presenting the deployment of web archives in lectures and publications. This helps increase the use of web archives by the scholars, researchers and art historians, who have begun using the web archives extensively to aid their work. NYARC supports and sees the potential advantage of having a visual and user-friendly interface of comprehensive snapshots of an artist's preferred work highlights over time. This is a better way to study and research about artistry and history which is digitally captured through some of the websites. New York City Galleries [6] is a collection under NYARC that has been used in this project.

CUL supports research on building web archives in thematic areas corresponding to existing collection strengths. The Human Rights web archives [4] is a huge collection that belongs to CUL. The evolution of public areas and environment in New York City are documented by the Avery Library Historic Preservation and Urban Planning web archive [5]. They publish all the essential and important information captured by non-governmental organizations, advocacy campaigns, parks conservancies, neighborhood associations and other such organizations and projects that is usually found online only.

3 Visualizations

This is a system which contains three different parts in it – three different kinds of TimeMaps representing the archived web pages. Having a visual representation or thumbnails of each of the mementos is very useful and effective in quickly determining the evolution of the webpage over a certain period and also helps a user to focus on the webpage. I used JSON file for the data and used technologies like HTML, JavaScript, JQuery and Ajax to create all the visualizations. All the visualizations are based on images of the web archived pages of websites from the archived collections of NYARC and CUL as discussed earlier. There are no graphs or trends used in this system and the visualizations show the best possible ways to be able to access the web archives.

The idea behind "Image Slider Visualization" is to imitate the photo roller property used in a photo album in iPhoto. Simply by moving the cursor over the images, which are the snapshots of the archived pages of a particular collection, we can see the images sliding and click on the page that we are interested in exploring more. There are additional properties added to this visualization, Play/Pause buttons, for example.

The "Image Grid View" is a much simpler representation that shows a responsive grid view to show all the snapshots of archived pages at one glance. This helps in viewing and comparing all the snapshots and choosing the ones that are subject of interest. The "Timeline View" is a very effective visualization of web archives as it represents the snapshots of the archived web pages on a time line based on the date that it was archived. Users can easily search and select the page based on year that the page was archived and also can see how distinct the pages are.

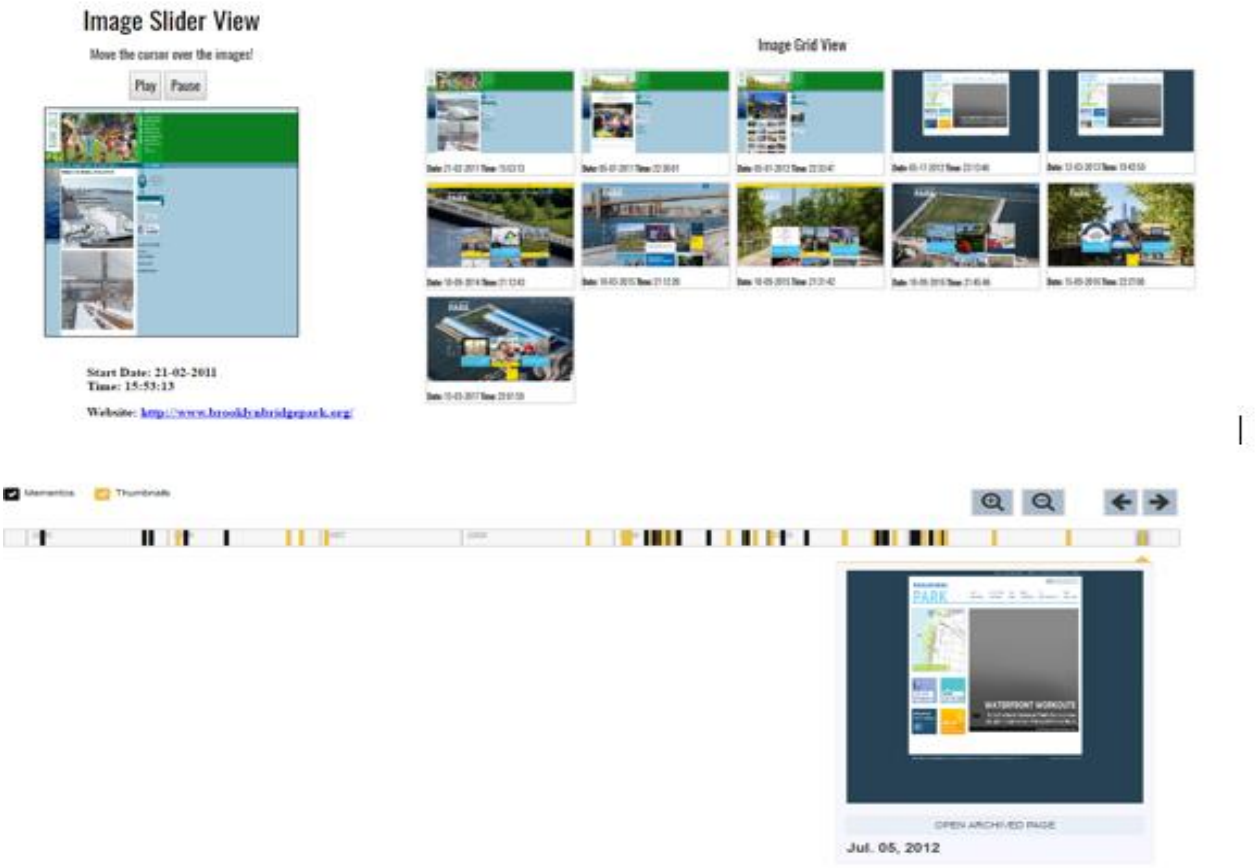

Figure 3: Overview of the system

To create timeline visualizations, I used the libraries that are used in Timeline Setter website - https://www.propublica.org/special/tbi-psycho-platoon-timeline [2]. There is an attribute in the JSON file named as "Timestamp", using which the lines representing the archived time of the website have been marked on the timeline. Timeline attribute is nothing but the dates that are converted to milliseconds format. I used this website to convert Human readable date to timestamp - http://timestampconvert.com/ [12].

A glimpse of the TimeMaps are shown above in Figure 3. All the three visualization systems were put together using "iframe" tag in order to be able to compare the visual effectiveness of each of

the representations. Insights for all of the visualizations are given in the section. Every archived collection is separately placed, the links for which are listed below.

Atlantic Yards - http://www.cs.odu.edu/~sshankar/MS_Project/project_AtlanticYards.html

Brooklyn Bridge Park - http://www.cs.odu.edu/~sshankar/MS_Project/project_BrooklynBridgePark.html

Cage Prisoners - http://www.cs.odu.edu/~sshankar/MS_Project/project_CagePrisoners.html

Cofadeh - http://www.cs.odu.edu/~sshankar/MS_Project/project_Cofadeh.html

Gulf Labor - http://www.cs.odu.edu/~sshankar/MS_Project/project_GulfLabor.html

Peter Blum Gallery - http://www.cs.odu.edu/~sshankar/MS_Project/project_PeterBlumGallery.html

All of the my work can be viewed in the following Gitlab link - https://git-community.cs.odu.edu/sshankar/MS_Project.git

4 Sources of Data and Data Cleaning Process

Collecting data for this took a lot of time because NYARC and CUL have a vast collection of archived pages. Snapshots of every page has been stored and used in the TimeMaps. All the mementos are chosen from collections of the New York Art Resources and Consortium (NYARC) and Columbia University Libraries (CUL). I selected some of the archived collections from NYARC and CUL which are distinct from each other. Links for the same are displayed below.

- https://wayback.archive-it.org/1068/*/http://www.cageprisoners.com/
- https://wayback.archive-it.org/1757/*/http://www.brooklynbridgepark.org/
- https://wayback.archive-it.org/1068/*/http://www.cofadeh.hn/
- https://wayback.archive-it.org/1068/*/http://gulflabor.org/
- https://wayback.archive-it.org/1757/*/http://www.atlanticyards.com/
- https://wayback.archive-it.org/4847/*/http:/peterblumgallery.com/
- http://wayback.archive-it.org/1068/*/http://www.amnesty.org/

The archived pages are hosted with the Internet Archive's Archive-It service. NYARC's collections are relatively smaller and more recent ones. On the contrary, CUL collection is vast and maintains a long-running human rights collection when compared to other collections.

The URIs collected from these Archived collections and snapshots are stored. The path for the images and also the URIs are stored in a JSON file. JSON file is created using a python script. Input for the python code will be a text file containing all the URIs listed. Figure 4 is a snapshot of a sample JSON file which has been used for Image Grid and Image Slider visualizations.

```
[
  - {
        website: "http://www.amnesty.org/ ",
        photo: "20110902210647.png",
        content: "http://wayback.archive-it.org/1068/20110902210647/http://www.amnesty.org/ ",
        start: "02-09-2011",
        time: "21:06:47",
        id: "1"
    },
  - {
        website: "http://www.amnesty.org/ ",
        photo: "20120302210436.png",
        content: "http://wayback.archive-it.org/1068/20120302210436/http://www.amnesty.org/ ",
        start: "02-03-2012",
        time: "21:04:36",
        id: "2"
    },
  - {
        website: "http://www.amnesty.org/ ",
        photo: "20130307001447.png",
        content: "http://wayback.archive-it.org/1068/20130307001447/http://www.amnesty.org/ ",
        start: "07-03-2013",
        time: "00:14:47",
        id: "3"
    },
  - {
        website: "http://www.amnesty.org/ ",
        photo: "20141001174452.png",
        content: "http://wayback.archive-it.org/1068/20141001174452/http://www.amnesty.org/ ",
        start: "01-10-2014",
        time: "17:44:52",
        id: "4"
    },
  - {
        website: "https://www.amnesty.org/en/",
        photo: "20151001222753.png",
        content: "http://wayback.archive-it.org/1068/20151001222753/https://www.amnesty.org/en/",
        start: "01-10-2015",
        time: "22:27:53",
        id: "5"
    }
]
```


Figure 4: JSON file format for Image Slider and Image Grid views

Figure 5 shows the samples of snapshots of the archived pages which are used for visualizations. These are stored in ".PNG" format and the path for these images are included in the JSON file. Also, archived date and time is stored in the JSON file.

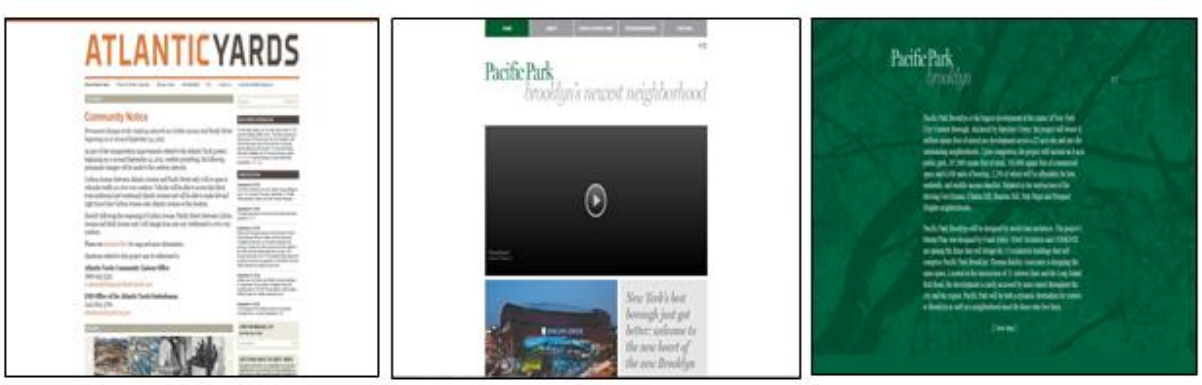

Figure 5: Sample Thumbnails used for visualizations.

The JSON file that is used for Timeline Visualization is slightly different due to the way the data is used in JavaScript files. Figure 6 shows the structure of the file.


```
[
  - {
        timestamp: 1298321593,
        event_series: "Mementos",
        event_html: "<img src='http://www.cs.odu.edu/~sshankar/MS_Project/timeline/photos/BrooklynBridgePark/20110221155313.PNG' height='300px' width='300px' />",
        event_date: "Feb. 21, 2011",
        event_display_date: "",
        event_description: "",
        event_link: "https://wayback.archive-it.org/1757/20110221155313/http://www.brooklynbridgepark.org/"
    },
  - {
        timestamp: 1298322085,
        event_series: "Thumbnails",
        event_html: "<img src='http://www.cs.odu.edu/~sshankar/MS_Project/timeline/photos/BrooklynBridgePark/20110221160125.PNG' height='300px' width='300px' />",
        event_date: "Feb. 21, 2011",
        event_display_date: "",
        event_description: "",
        event_link: "https://wayback.archive-it.org/1757/20110221160125/https://www.brooklynbridgepark.org/"
    },
```


Figure 6: JSON format for Timeline Visualization.

5 Insights on the System

I examined each of the visualization while before and after developing them to see if the design is effective and if the design is achieving the basic goal of this system. The common goal of all these visualization is to provide a user-friendly interface to researchers and scholars who use web archives for their research. I have a brief description and observations of each of the Visualizations listed down here.

5.1 Image Slider

a. Snapshots of archived pages which are distinct is displayed on a single thumbnail view, along with the information related to the page below the image itself.
b. The thumbnail images change when the cursor is moved across the image, which makes it easier for a user to pick their choice of archived page. This style is similar to iPhoto image previews.
c. Play/Pause button is included to make the image slider work like a video. Also, when clicked on the image, it navigates to the corresponding archived page.
d. Image slider helps swipe through time and view all the thumbnails of various archived pages which show the evolution of the webpage clearly, in order to compare between the snapshots how distinct they are from each other.

5.2 Image Grid

a. Thumbnails are placed as a 5 X 5 or 4 X 4 grid (depending on the number of images to be included) with images placed next to each other giving room for comparison between the thumbnails showing the snapshots of the archived pages. The idea is to show the entire thumbnail summary in one grid.
b. The interface is responsive so that more snapshots can be added at any point of time to the JSON file from where the data is pulled out from.

c. Under each image, date and time of archival is displayed. On clicking on any of the image, it navigates to the archived page itself.
d. This visualization helps users to see all the thumbnails together and easily view the distinction between them.

5.3 Timeline

a. The interactive timeline view includes all the thumbnails on a timeline which indicates the time when the page was archived.
b. The lines on the timeline represent the thumbnails and mementos and on clicking on the line we can see a twitter card style thumbnail displayed below the timeline.
c. Two different colors are used to represent the “Thumbnails” and “Non-thumbnail mementos”, yellow and black respectively.
d. This difference between this and other visualizations is that depending on the size of the TimeMap, the timeline includes mementos that are not selected as a part of the summary as well.

6 Challenges

This project was a great learning experience for me. I had an opportunity to work different kinds of visualization techniques and also learnt about web archiving concepts. However, the project was not very easy for me and there were a few challenges that I faced in the path. Some of the challenges I faced while I working on the project are mentioned below.

i. Collecting data from each of the archived collections was a time consuming task. I had to open each page and determine if they are distinct and then select the mementos to be represented in the TimeMap. Initially, it was easy because I had to do a mock up for just one of the collections, but later this task got tougher because there are six different collections that I have represented in the TimeMap. Moreover, this data grows whenever it is archived and hence the data file has to be updated each time.

ii. Algorithm was difficult to understand in the beginning. There were some of the Node.js and Phantom.js libraries and packages which were not compatible in MAC OS and also few packages which were outdated hence had to be replaced. This was also a difficult task as it was the first time for me to run a Node.js code.

iii. Timeline visualization has a JSON file with a slightly different format. I used a separate source for date to timestamp conversion. Also, I had to create separate JSON files for each of the collection again, which took a lot of time.

7 Future Enhancements

Here are some of improvements and further implementation ideas, although the project is substantial.

- This project can be coupled with the Thumbnail Summarization implementation by just obtaining the dump of all the distinct URLs in a text file. This can be put into a JSON file in the format that is required and represented in the TimeMaps implemented in this project.
- For Image Grid view, we can also include a two dimensional responsive interface and have resizable thumbnails. This way, the thumbnails arrange itself depending on the number of records in the data file.
- Timeline implementation can be improvised by making it more interactive. Including filters to select years or months can be one such example. Also, we can include months and dates as we zoom into the timeline. We can also include zoom-in and zoom-out property with scrolling of the mouse.
- The data files have to be updated every time a new archival is added to the collection. Instead a script can be written to update the data file periodically such that the data is just appended to the existing data files.

8 References

[1] AlSum A., Nelson M.L. (2014) Thumbnail Summarization Techniques for Web Archives. In: de Rijke M. et al. (eds) Advances in Information Retrieval. ECIR 2014. Lecture Notes in Computer Science, vol 8416. Springer, Cham

[2] "Timeline Setter 0.3.2" ProPublica Journalism in the Public Interest <http://propublica.github.io/timeline-setter/>

[3] "Brooklyn Bridge Park website" Nov 30. 1999 .Web Apr 21. 2017 <http://www.brooklynbridgepark.org/> .Internet Archive. <https://web.archive.org/web/*/http://www.brooklynbridgepark.org/>

[4] "Human Rights web archives", Columbia University Libraries, May 2008, < https://archive-it.org/collections/1068 >

[5] "Avery Library Historic Preservation and Urban Planning", Columbia University Libraries, Jan 2010, < https://archive-it.org/collections/1757 >

[6] "New York City Galleries", New York Art Resources Consortium (NYARC), September 2014, < https://archive-it.org/collections/1757 >

[7] "Atlantic yards" Sept. 26. 2012 .Web Mar. 15. 2017 < http://pacificparkbrooklyn.com/> .Internet Archive. < https://wayback.archive-it.org/1757/*/http://www.atlanticyards.com/>

[8] "Cage Prisoners" Aug 09. 2011 .Web Jan 01. 2015 <http://www.cageprisoners.com/> Internet. Archive. < https://wayback.archive-it.org/1068/*/http://www.cageprisoners.com/>

[9] "Cofadeh" Jun 11. 2013 .Web Oct 01. 2015 < http://www.cofadeh.hn/ > Internet. Archive. < https://wayback.archive-it.org/1068/*/http://www.cofadeh.hn/ >

[10] "Gulf Labor" Apr 07. 2014 Web. Apr 14. 2017 < http://gulflabor.org/ > Internet. Archive. < https://wayback.archive-it.org/1068/*/http://gulflabor.org/>

[11] "PeterBlum Galleries" Nov 13. 2014 .Web Mar 28. 2017 < http:/peterblumgallery.com/> Internet. Archive. < https://wayback.archive-it.org/4847/*/http:/peterblumgallery.com/>

[12] "TimestampConverter.com" <http://timestampconvert.com/>