Rachel Sinquefield-Kangas &
Introduction to Digital Humanities
Final Project 22-12-2016

1. Research questions and corpus

Step one: Ok, so I wanted to see if I could challenge myself a bit, maybe in force myself to take a stab at coding, something that I have never done before. In order to do this, I needed to find some different data from my initial choice, so I used one of the data archives suggested in the PowerPoint *Collaborative⇒ open, shared and reproducible data*. Perusing through the list of suggested data archives, I noticed the link to the Finnish National Gallery of Art. From the link I downloaded both the Json files and the txt. Files. After opening the json file and skimming the text (some Finnish and some English) I came up with a research question.

⇒ How many painting in the Finnish National Gallery are self-portraits? …we'll see? Other potential questions; paintings acquired in 1977? How many are not on canvas? How many of the paintings are of women? * really cool stuff- where has one painting traveled to exhibitions? What are the different names of the collections and how many collections are there (private vs. publicly owned)?

From which only one was roughly at about my level of technological abilities:

Which artist's self-portraits belong to each of the National Galleries collections (Ateneum, Kaisma, or Sinebrychoff)?

2. Processing the data

Step one: Eetu unzipped the large file and from there we began putting the data in Jupyter to see it were possible to search the data for only the bits of data that involved the term self-portrait. After some discussion and trial and error this code was used to successfully extract only information from the data that was associated with self-portraits.

```
f = open('self-portraits.json','w')
selfportraits = []
for artwork in json_data['descriptionSet']:
    isSelfPortrait = False
    if 'description' in artwork:
        for description in artwork['description']:
            if 'fi' in description and 'makuv' in description['fi']:
                isSelfPortrait = True
    if 'title' in artwork:
        for title in artwork['title']:
            if 'fi' in title and 'makuv' in title['fi']:
                isSelfPortrait = True
    if isSelfPortrait:
```

```
        selfportraits.append(artwork)
json.dump(selfportraits,f)
f.close()
```

From there we used the next bit of code to further extract both the date of acquisition and the date of potential creation for each self-portrait.

```
yearCountDict = dict()
for artwork in json_data['descriptionSet']:
    if 'date' in artwork and 'acquisition' in artwork['date'][-1]:
        dateOfCreation = artwork['date'][-1]['acquisition']
        isSelfPortrait = False
        if 'description' in artwork:
            for description in artwork['description']:
                if 'fi' in description and 'makuv' in description['fi']:
                    isSelfPortrait = True
        if 'title' in artwork:
            for title in artwork['title']:
                if 'fi' in title and 'makuv' in title['fi']:
                    isSelfPortrait = True
        if isSelfPortrait:
            if not dateOfCreation in yearCountDict:
                yearCountDict[dateOfCreation] = 1
            else:
                yearCountDict[dateOfCreation] += 1
yearCountDict
```

After this step the data was still looking a bit messy as it was a rather large amount of data. Eetu again worked some of his magic and was able to provide me with an Excel spread sheet containing the information in the data file. The spread sheet contained information from the data that included; Creator, descriptor, year acquired, year created, a lot of columns of containing description terms, size of artwork, artwork media, name of the personal collection, as well as to which of the national galleries the artwork belongs.

3. Visualization of the data

My plan was to put it the data into Palladio and try to visualize the data. First, I copy and pasted the entire spread sheet which resulted in a mess of mass chaos. So, I had to revisit the spreadsheet and purse some of the information. I removed all of the columns until I was left with only the date of acquisition, date of creation, artists name, museum it belongs to, and collection. However, these results did not fare well in Pallidio well either. I was unable to really decipher what I was looking at on the visualization.

Revisiting the Excel spread-sheet I removed even more of the data until I was left with the name of the the museum that the painting belongs to, and the name of the artist that painted the self-portrait. For some reason the dates enter into the original data set were done in a variety of different ways. So, before eliminating the date entirely I tried to make the date entries all the same format but even still, when I uploaded it into Palladio the visualization was still unclear.

Finally, I decided to remove the dates and look at just try to visualize the national museum that housed the self-portrait in conjunction with the name of the artist. What resulted was were three clearly discernable nodes with the lables; Ateneum, Sinebrychoff, and Kaisma. Out from each of the nodes is a link to the all the self-portrait artist housed in that museum. I downloaded the Palladia visualization in an svg. File but it was hard to get all three clusters on the same page without zooming out quite far. Or, at least far enough to where the graph is hard to read.

4. Analysis

While the content of the data was pretty straight forward, it seems as though the initial input of the data did not follow a consistent format. This made it difficult to visualize the data using the tools without having to spend a lot of time clearing up the data. Also, there was a lot of redundant information in the data as it was entered in Finnish, Swedish, and English.

When processing the data, particularly the dates, it was tricky because some of the artworks were not dated or were entered as a range of dates. Seeing as how there were 571 rows of information (and this was just for self-portraits) having to go through all that manually normalize the data was quite time consuming. In order to do this in a more efficient way, I'm sure that it would require some coding skills.

With the Palladia tool I was able to determine the number of self-portraits from the national collection as they are housed is in each museum:
Ateneum = 363
Kiasma = 165
Sinebrychoff = 42.
However, this number is off by one..? It also provided a list of the artists names in accordance to these numbers.

This information has little connection to my own person research. My reasons for enrolling the course was really based on my own personal curiosity. The course has been very interesting and I can see a lot of potential for being able to master the art of both analyzing and visualizing 'big data'. It also brings into awareness some of the ethical and validation concerns to keep in mind when reading the results from other publications.