Methodology report

Generally, this task has shown to me how important it is to critically reflect on **how to organize** things within the field of Curating Data. Also this assignment coined my notion on taxonomy in general. Taxonomies are mostly concerned with the description and identification of a given subject. In the last assignment we saw and analyzed the process of datafication. For now, in this assignment the focus is more laid on the "data structuring" (Mikkel & Murray, 2018), when it comes to reshaping the given data. In that context it is important to highlight the influence of social activity. Certain datasets are reshaped in order to follow certain aims, or to persuade a given target audience, or to display a given phenomenon in a given context. Due to the fact that this is based on human decisions which have a given picture in mind of how to realize a given structure, taxonomies have a certain bias. Also in our assignment when constructing our own taxonomy I realized that we for example have set the authors as a main information branch of the taxonomy, whereas other information is more linked to the author. We for example added a few links to the author like occupation or citizenship. This is because we decided that the idea of our taxonomy is to structure the data of the texts in a way that the contribution of the different authors is the main knowledge provided. Connected to that, also our visualization is biased in a way that we first displayed a map where the citizenships are displayed, second we have a dendrogram showing only occupation and citizenship, as well as the title of the text which was produced and third a visualization showing the countries written, however the text of a given country is bigger if more texts were produced with that background. Especially when it came to the visualization of the dendrogram I chose to make the nodes having a different color for each author respectively, or to make the title of the different texts as a starting point to demonstrate which authors with different occupation and citizenship have contributed to the specific texts. The procedure with Open refine and wikidata we chose was in a way showing off the difficulties one has to face when doing a dataset, namely coping with missing information, or making decisions on how to depict data in a dataset, which columns should be reconciled, what additional information would be important for our purpose? The task of data structuring was facing our group with several challenges.

Taxonomies are created because they give a way of working with collections in a structured way and that can be used in machine learning to automatize aspects of our lives (Nadim 2021). Moreover, taxonomies are a way of seeing the world. It is the way we group together something, and how we define what is what. Because taxonomies often are reductions of the world, different taxonomies can serve different purposes, political, cultural or scientific. That knowledge is specifically important when thinking about working with given taxonomies. One has to consider which purpose is underlying and to what extent the reduction of the world was done. Is the given taxonomy fitting to

my application? Or in other words: What is the essence, in which knowledge am I the most interested in?

Personally, it was the first time for me to work with wikidata. I was wondering that there is no common convention on how to do such an entry with the data especially when it came to different labeling. Sometimes there were even 2 different links for the same person. One has to be aware of the somewhat inconsistent usage of wikidata which can prove to be difficult. Nevertheless, it is definitely a powerful tool to show linkage of data.

Nadim, Tahani. 2021. "The Datafication of Nature: Data Formations and New Scales in Natural History." *Journal of the Royal Anthropological Institute* 27 (S1): 62–75. https://doi.org/10.1111/1467-9655.13480.

Flyverbom, Mikkel, and John Murray. 2018. "Data Structuring—Organizing and Curating Digital Traces into Action." Big Data & Society 5 (2): 2053951718799114. https://doi.org/10.1177/2053951718799114.

Lisa Golla

This is for assignment 3 - documenting the workflow

Aim: In this assignment you work in your study group to 1) map all of the tasks performed when generating your taxonomic model for the previous assignment, and 2) to visualize these tasks through a diagram or a flowchart.

The order of what we did

1. The first thing we did was attending the two workshops hosted by Lozana in week 41.
    a. Here we learned how to use/do:
        i. WikiData
        ii. OpenRefine
        iii. Visualize this data
2. We found all the WikiData links for each text and author in the syllabus.
    a. The links we were able to find were put in a list we had in an online document.
    b. The texts that did not have a link, we decided, were going to be added to the query manually - if possible.
3. We made a query with all the data.

    a. In the beginning we used many example codes to see if they could help us understand how we could create our own query with the different texts from the syllabus.
    b. We encountered many problems with the code.
        i. Mattias searched for the texts Q value for the query to see if that worked.
        ii. Alexander used the authors Q value to see if we could make a query using that and then filtering out the texts that aren't in our syllabus.
4. We put the data into OpenRefine.
    a. We reconciled the data for each section and added some of the values that weren't available on WikiData.
        i. We thought it was an important note to make, that sometimes you do not have access to all the data and need to find ways to deal with it.
5. In the end we had our taxonomy.
6. The visualization
    a. We made a map with the coordinates for each author's place of birth to see if there's any relations between them and where they're from.
    b. Lisa made a dendrogram which displays the different occupations and origin from the authors. That way the diversity of the contribution is shown.
7. For the visualization in general to work we needed to find ways to fill the missing knowledge in our data. For that we did some research about what the country of citizenship was and what the occupation was (if it was missing) and filled it manually.

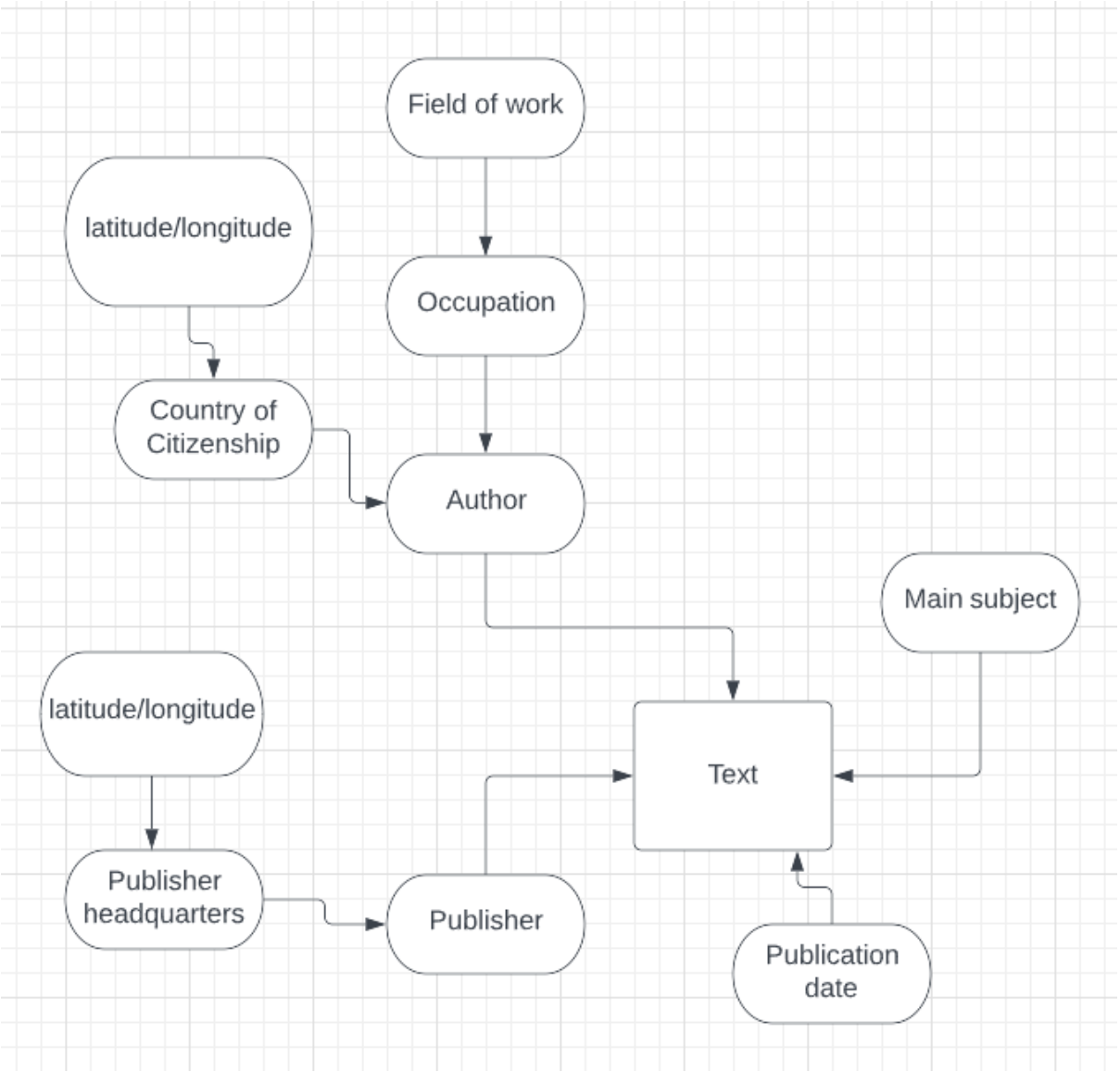
The taxonomy should include:

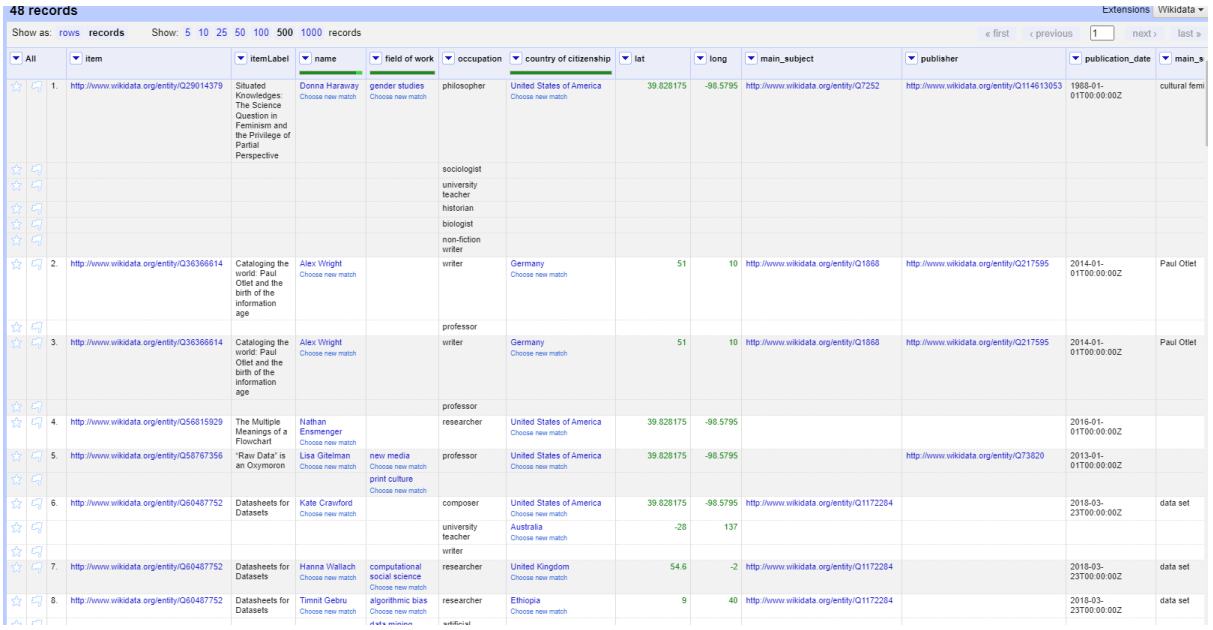
- data structure - consisting of classes (and sub-classes) and relations, including definitions for each class and the types of relations possible between classes.

- Using the text as a source to link to the authors; Authors background studied in order to understand the contribution it has on the text (multiple authors bring different factors into the written text)
    - Location, occupation; field of work (as a subclass), etc. to get the idea of how different backgrounds make up the field of curating data

Definitions for each class are as follows; text relates to the text as an object with relational values such as publisher, author, publication date and publisher. Under each value with a geographical location, a subclass of latitude and longitude are added. The class *author* has sub-classes occupation and field of work nested within. The database was created from the relations the texts have to their authors (if they are linked), and which was then used to expand the database with information such as field of work, publisher etc. We wanted to build our database in this fashion because we wanted to have geographical data as well as semantic, qualitative data about the texts and their authors, as well as a relation between these. We also decided to build the database in this way to have an understanding of each author's contributions and background, as the authors in our syllabus have a diverse range of backgrounds. We also chose to reduce our data in order to visualize it, as we had duplicates of certain texts over 14 times. This was because if a text has more authors, the text would appear once for each author. If the author had several occupations, the text would also appear for each

occupation. We solved this issue by reducing occupations to one value, and deleting duplicate rows in our dataset.





48 records

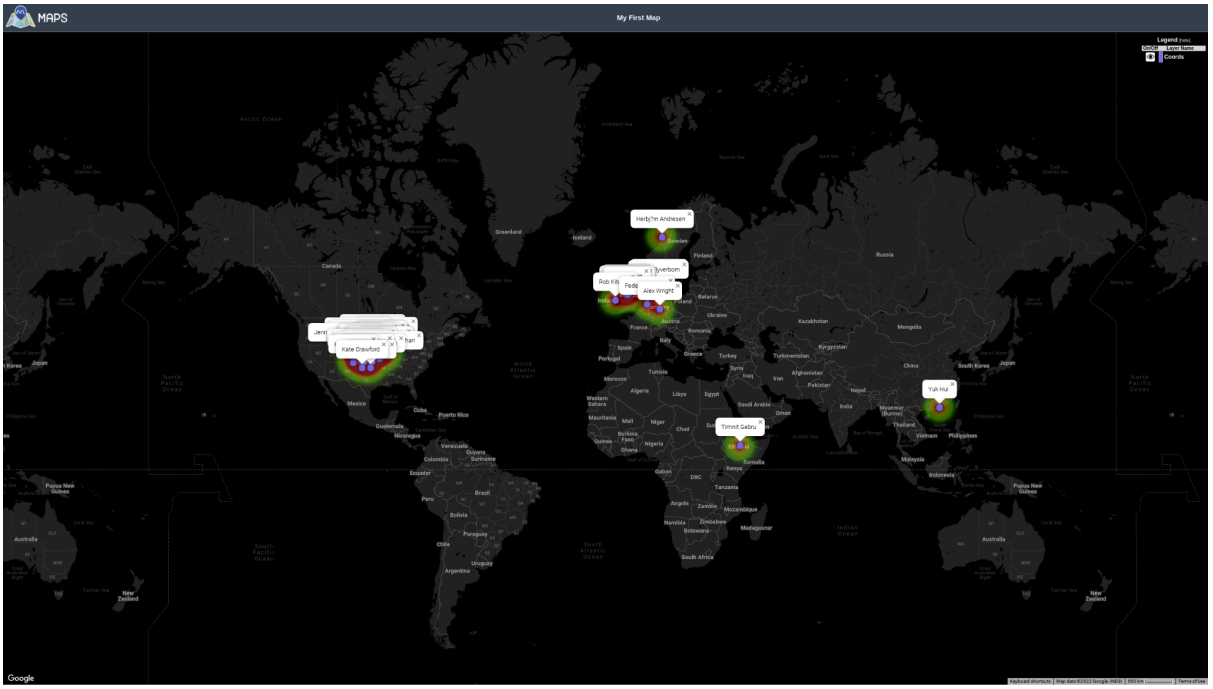
Extensions Wikidata

Show as: rows records Show: 5 10 25 50 100 500 1000 records « first ‹ previous 1 next › last »

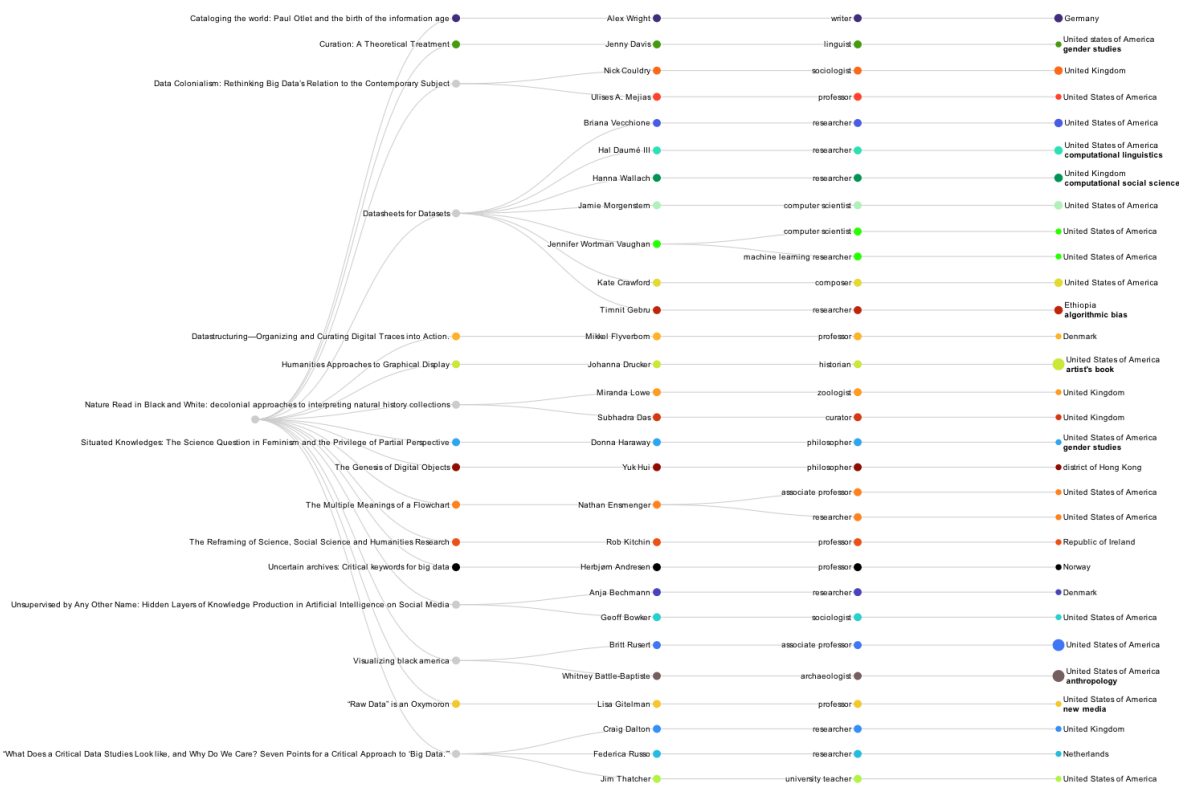
All		item	itemLabel	name	field of work	occupation	country of citizenship	lat	long	main_subject	publisher	publication_date	main_s
	1.	http://www.wikidata.org/entity/Q29014379	Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective	Donna Haraway Choose new match	gender studies Choose new match	philosopher	United States of America Choose new match	39.828175	-98.5795	http://www.wikidata.org/entity/Q7252	http://www.wikidata.org/entity/Q114613053	1988-01-01T00:00:00Z	cultural femi
						sociologist							
						university teacher							
						historian							
						biologist							
						non-fiction writer							
	2.	http://www.wikidata.org/entity/Q36366614	Cataloging the world: Paul Otlet and the birth of the information age	Alex Wright Choose new match		writer	Germany Choose new match	51	10	http://www.wikidata.org/entity/Q1868	http://www.wikidata.org/entity/Q217595	2014-01-01T00:00:00Z	Paul Otlet
						professor							
	3.	http://www.wikidata.org/entity/Q36366614	Cataloging the world: Paul Otlet and the birth of the information age	Alex Wright Choose new match		writer	Germany Choose new match	51	10	http://www.wikidata.org/entity/Q1868	http://www.wikidata.org/entity/Q217595	2014-01-01T00:00:00Z	Paul Otlet
						professor							
	4.	http://www.wikidata.org/entity/Q56815929	The Multiple Meanings of a Flowchart	Nathan Ensmenger Choose new match		researcher	United States of America Choose new match	39.828175	-98.5795			2016-01-01T00:00:00Z	
	5.	http://www.wikidata.org/entity/Q58767356	'Raw Data' is an Oxymoron	Lisa Gitelman Choose new match	new media Choose new match	professor	United States of America Choose new match	39.828175	-98.5795		http://www.wikidata.org/entity/Q73820	2013-01-01T00:00:00Z	
					print culture Choose new match								
	6.	http://www.wikidata.org/entity/Q60487752	Datasheets for Datasets	Kate Crawford Choose new match		composer	United States of America Choose new match	39.828175	-98.5795	http://www.wikidata.org/entity/Q1172284		2018-03-23T00:00:00Z	data set
						university teacher	Australia Choose new match	-28	137				
						writer							
	7.	http://www.wikidata.org/entity/Q60487752	Datasheets for Datasets	Hanna Wallach Choose new match	computational social science Choose new match	researcher	United Kingdom Choose new match	54.6	-2	http://www.wikidata.org/entity/Q1172284		2018-03-23T00:00:00Z	data set
	8.	http://www.wikidata.org/entity/Q60487752	Datasheets for Datasets	Timnit Gebru Choose new match	algorithmic bias Choose new match	researcher	Ethiopia Choose new match	9	40	http://www.wikidata.org/entity/Q1172284		2018-03-23T00:00:00Z	data set

- visual representation of taxonomic model (a diagram or a chart)

- max 2-3 pages (4800-7200 characters) methodology (incl. information on how each student was involved and their responsibilities in this task, and steps along the way)



https://maps.co/map/635903a1787a9065940372bzh7b4605



done with raw graphs


```
coder:/work$ /usr/bin/python /work/curatingData/data.py
{'usa': 34, 'england': 10, 'germany': 2, 'ethiopia': 2, 'denmark': 2, 'australia': 1, 'netherlands': 1, 'norway': 1, 'hongkong': 1, 'ireland': 1}
coder:/work$
```