Assignment 1
Collecting: Building Datasets

Mie Buchhave Ryborg (202004812@post.au.dk)

Introduction

This assignment focuses on the collection and digitization of analogue objects. The aim is to build a dataset of at least 20 objects and reflect and evaluate on methods of digitizing and curation in terms of how analogue objects are labelled and become digital. Following three questions will be answered and critically reflected upon:

1. What will be collected?
2. How will the collection be digitized?
3. What variables will be collected?

Repository:

The collection can also be accessed in the following repository on GitHub:
https://github.com/miesogaard/mugcollection

1. What will be collected?

In this assignment, the object to be digitized will be my personal collection of coffee mugs, counting 21 unique mugs. The fact that the mugs are used in everyday life will make it valuable to have a more organized overview over the collection, as some of them have to be packed away due to limited amount of shelf space.

2. How will the collection be digitized?

With the emergence of modern computational machines, all kinds of analogue, tangible objects can be converted into and represented in digital formats (Hui, 2016). When digitizing a collection, one needs to make sure that the system one uses actually supports the representation that one aims for the digital collection. For this collection, it has been chosen to work with a flat, tabular format with each column representing a variable and each row corresponding to a given value of the column's variable (*relational systems,* Dourish, 2014). Alternative data representations are undoubtedly important to assess when creating a digital collection, as different systems have different levels of nesting and structures. The relational system has a flexibility and extensibility that makes sense for this collection in order to organize, update, extend, and retrieve information.

The mugs are also going to be digitized by creating a bank of pictures of the mugs. Every mug has been photographed from two angles; the front and top. Front has been defined as the side of the mug that reveals the most information about the mug as possible in one single picture. If a mug did not have any significant features, a random angle from the side has been photographed. Pictures of more angles can be considered to add in the future, as some perspectives on the mugs are missing (especially with regards to mugs with extensive features). Pictures of the mugs were taken with an iPhone 8 in a quadrat format, and files are stored as PNG-files.

For the purpose of this assignment, no paid CMS or DAM system is used for storing the collection. Instead, a repository on GitHub has been created, containing the csv-file of the data, a folder with the pictures of the mugs, and a README for documentation of the project and variables. This way, it will be possible to update the collection and maintain version control.

3. What variables will be collected?

Today, there is an abundance of data surrounding the digital environment (Hui, 2016), and digital curation has become one of the central component in order to discriminate the giant pool of data to identify relevant materials for display (Davis, 2017). In this collection, decisions has been made in

terms of what type of information about the objects is registered and displayed in the dataset. This dataset captures ID, name ID, height, diameter of the top, handle, volume, color (primary and secondary), motif, the country in which they were bought, thermal capability, and links to the pictures of the mugs. Some values and attributes were neglected due being non-retrievable across all mugs or not considered relevant enough for this particular collection.

Besides considering *what* information is displayed, it is also important to consider *how* information is displayed (Haraway, 1988). For instance, one could choose to represent color by HEX or RGB codes. However, a color code has the disadvantage that it makes it hard to filter and organize the collection based on a more broad categorization of color, and reading a color code is often not very meaningful to humans. Since readability and filtering has been weighted as essential mechanisms in this digital collection, it was chosen to represent and classify colors in terms of 12 colors (Boynton, 1997). Consequently, the colors have been registered based on subjective assessments of the mugs in real life that does not capture specific nuances of colors.

The *name ID*-variable attempts to group the mugs through associations that cannot be obtained by any other variables in the dataset (e.g., belong to the same brand series). However, some of the names are based on subjective assessments and groupings of the mugs in terms of whether they were bought to complement each other, or have a certain personal association. Additionally, some of the name ID's consists of abbreviations to prevent long variable names. Thus, it will require understanding context to a certain degree be able to interpret the data from the creator's perspective.

It was also chosen to represent all true/false variables as Boolean logic (0 and 1). In other words, all 1's in the *"handle"* variable corresponds to "True" (*"there is a handle on this mug")* and 0's corresponds to "False" (*"there is no handle on this mug"*). Again, this is a digital representation of the objects where it is important to understand the structure of Boolean logic and what is coded as 0 and 1, otherwise it might be easy to confuse the variables.

These examples of considerations made in the creation of this collection show how digital representations of objects are limited to human perception; a product of particular worldviews into ideas, artifacts and experiences which guides decisions for representation (Acker, 2021). Thus, as any specific collection with a certain goal behind it, the dataset becomes a selection and one specific digital representation of the entirety (Davis, 2017). This is important to acknowledge in any act of collecting and curating, and thus, also when working with this collection.

References

Acker, Amelia. 2021. "Metadata." In *Uncertain Archives: Critical Keywords for Big Data*, edited by Nanna Bonde Thylstrup, Daniela Agostinho, Annie Ring, Catherine D'Ignazio, and Kristin Veel, 321–29. Cambridge, Massachusetts: The MIT Press.

Boynton, R. M. (1997). Insights gained from naming the OSA colors. In C. L. Hardin & L. Maffi (Eds.), *Color categories in thought and language* (pp. 135–150). Cambridge University Press. https://doi.org/10.1017/CBO9780511519819.006

Davis, Jenny L. (2017). "Curation: A Theoretical Treatment." *Information, Communication & Society* 20 (5): 770–83. https://doi.org/10.1080/1369118X.2016.1203972.

Dourish, Paul. (2014). **"**No SQL: The Shifting Materialities of Database Technology." *Computational Culture,* no. 4 (November). http://computationalculture.net/no-sql-the-shifting-materialities-of-database-technology/

Haraway, Donna J. (1988). "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective." Feminist Studies 14 (3): 575–99. https://doi.org/10.2307/3178066.

Hui, Yuk. (2016). "The Genesis of Digital Objects." In On the Existence of Digital Objects, 47–74. Electronic Mediations 48. Minneapolis: University of Minnesota Press.