


### Midterm: Data Biography Reflection

The **Museum of Modern Art** was established in 1929. In July 2015, the museum released a dataset of its vast collection of almost 200,000 artworks from **the last 150 years** that is available on Github. The dataset is categorized into sections of Works and Artists. Here is what the Github [site](#) says about the exact figures: “The Artists dataset contains 15,526 records, representing all the artists who have work in MoMA's collection and have been cataloged in our database. It includes basic metadata for each artist, including name, nationality, gender, birth year, death year, Wiki QID, and Getty ULAN ID.” Also according to the site, “This research dataset contains 149,867 records, representing all of the works that have been accessioned into MoMA’s collection and cataloged in our database. It includes basic metadata for each work, including title, artist, date made, medium, dimensions, and date acquired by the Museum.”



[This project](#) was headed by Michelle Elligott, chief of the museum’s archives and Fiona Romeo, the director of digital content and strategy. There isn’t much information available about the process of completing it, but the site names John Cline and John Halderman as contributors. Romeo wrote a [Medium article](#) explaining their motives behind creating and placing the dataset in the public domain using a CC0 license, which allows anyone to access and use it freely. This gives us some insight into the intended research that would arise from it. The data release was celebrated with an exhibition titled *This is for Everyone: Design Experiments for the Common Good*. There is certainly a political bent to this innovation. The MoMa has continuously strove to

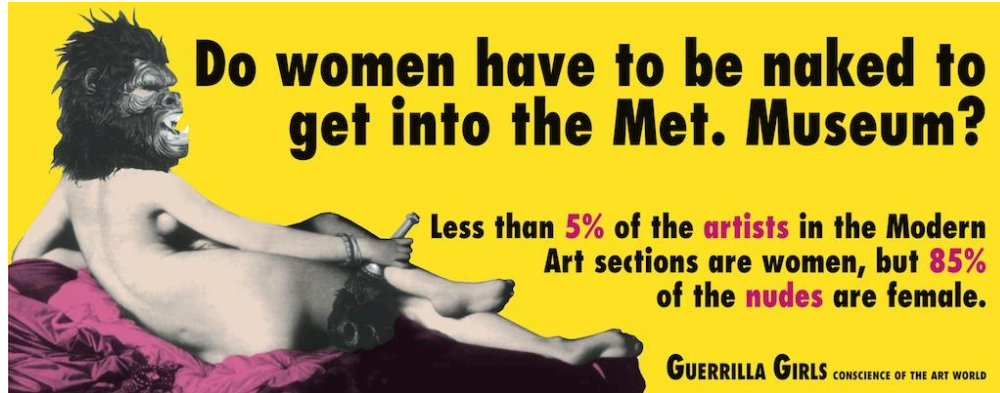
democratize its work and update its image as an elite institution. There is a Medium [article](#) being updated regularly that lists the ways in which people have used the dataset: many played with the data to assess the ‘modernity’ of the artwork and the ratios of the works’ dimensions, many used it to evaluate the representations of different groups (gender, ethnicity etc.) and artists (Picasso was one of the top represented artists). In addition, a few people created Twitter bots that post Museum items periodically, and all of these actions do serve the MoMa’s original incentive.

MoMa acknowledges many of the limitations of this dataset, some of which open a large scope for ethical concerns. Records that have incomplete information are noted as “not Curator Approved.” The datasets are not clean, as was [observed](#) by one user Celeste Grupman, who pointed out that in some cases, years are ranges (for ex. 1967-76) and some are single years. In addition, images are not part of the dataset, which affects its handling greatly considering that the artwork is visual and much of the way it is treated hinges on details regarding its appearance. In 2022, a user opened the issue that they were unable to see whether works were part of a series, and requested that this metadata be included. So the archive can be considered incomplete by many metrics. However, this also operates on a different level, where both the accessibility in and hindrances to using this dataset complicate its use. On a technical level, both datasets are available in **CSV format, which is not correctly interpreted by Excel on a Mac**, a fact that may not be known to a variety of users, posing a barrier to entry for its use. The site also says: “This data is provided “as is” for research purposes and you use this data at your own risk. **Much of the information included in this dataset is not complete and has not been curatorially approved.** MoMA offers the datasets as-is and makes no representations or warranties of any kind.” Pull

requests are not permitted, so there are limitations to the degree to which this data can be updated by the public and is ‘available’ to this domain in this sense.

There are also more base issues that deal with the digitization of this kind of work to begin with. According to the [website](#), “The collection includes an ever-expanding range of visual expression, including painting, sculpture, printmaking, drawing, photography, architecture, design, film, and media and performance art.” On a principled level, the artwork that is the most interactive—installations that render visitors a part of the work and performances that invite their participation are simply not possible to digitize or make accessible at all. This limits the impression created of the data. It could lead to many misleading conclusions about the nature of the **Moma’s** collection. The prolific performance artist Marina Abramovic’s work is a prominent example of this limitation. Much of her work in the MoMa has been hugely historic in the modern art world. Her 2009 [piece](#) *The Artist is Present* involved inviting visitors to sit across from her, and the dataset only includes audio clips of her interviews and explanations, which turn the work into a different form entirely.

When I first encountered this data set, I thought of the Guerilla Girls, a vigilante activist group of anonymous women who have been striving to improve the representation of women artists in galleries since 1985. One particular reason I made this mental connection was because for a long time, they have been using an abbreviated version of datasets that are adjacent to this one, as can be seen from their works:



A dataset such as this one would only provide more opportunities to point out the exact ways representation has been skewed in favor of men. One student Yuna Shin actually [did](#) so, scrubbing the data for categories such as ethnicity, gender, and top female artists. What is most curious to me in all this democratization, however, is the way in which in this case the benefits are mutual: an annotated history of their work is now available as a digital [timeline](#) on the MoMa website.