Huo(s)-Data-Biography
Andrew Huo
HUM 346
02/19/2024

Photography at the Carnegie Museum of Art Data Biography

For their 120th anniversary the Carnegie Museum of Art in Pittsburg, PA, decided to publicize a digital record of all its accessioned artwork. This included objects from different departments such as fine arts, decorative arts, contemporary arts, and so forth, but this data biography will focus specifically on photographs. A dataset of 87,291 photographs was posted on GitHub and grouped into two sections: (1) CMOA Collection Data (28,259) and (2) the Charles "Teenie" Harris Archive (59,032), a photographer for the *Pittsburgh Courier* focusing on the every day of Black life in 20th-Century America. Each photograph data was released in a CSV format and JSON format and consisted of 31 pieces of information, including title, dimension, medium, artist information, image URLs, and much more, except for the digitized photograph itself. The decision to release the data on GitHub itself invited version control and continuous updating for better accuracy and accessibility. The project was overseen by David Newbury (Carnegie Museum of Art Lead Developer on Art Tracks) and Daniel Fowler (Open Knowledge International).

This collections data publication was an offshoot of another CMOA project, Art Tracks, which had the sole purpose of vizualizing provenance data. Its goal was to shift traditional handwritten provenance records into searchable structured data using the building tools of open-source software that can be accessed from multiple institutions. For museum art objects, tracking provenance is incredibly important, and thus, it was necessary that qualified professionals took charge of the process. Travis Snyder (Collections Database Administrator)

created XML reports taken from the collections management system, which were uploaded to an internal FTP site. Then, David Newbury transformed the XMLs, making simpler labels and connecting data across tables.

Later, for the larger collections data project, Carnegie Museum brought in other professionals to contribute to the release of data: David Brennon (minor changes to metadata), Bob Gradeck (advice and assistance with datapackage.json), Matthey Lincoln (documentation review and suggestions), and Zac Yu (Documentation fixes). Six months were given to allow departments to correct and certify information as well as allow information to be held back by anyone who was not ready for publication. The data was placed in a single command line using Rake (a Ruby library that automated repetitive tasks). The task downloaded the XML, reprocessed the information, and then created both CSV and JSON formats, and finally uploaded the data on GitHub. Lastly, someone would run the exports to fix any issues before publishing the data. The platform of GitHub made it easier for researchers to use large collections of data without any further necessary tools. CSV was the most accessible format for quantitative analysis, and JSON exports were the best formats for developers to load and process the data.

The number one purpose of this project was for the Carnegie Museum of Art to allow Open Access to all of its accessioned artwork, believing that "exploration, education, experimentation, and fun" (Fowler, 2017) were necessary components of museum practice. On the Art Tracks website, the overview mentions the potential complex questions that this original project could help answer: (1) "Which works in our collection were in the same city in the same year?" and (2) "Which artworks in our collection were owned by an artist whose work is also in our collection?" Lastly, the overview ends with a statement that with the newly digitized data, "we can discover gaps in knowledge, shape collections policy, and better understand the

ecosystem of the collection and the institution." Furthermore, CMOA based the idea of digitization on what other large museums and academic institutions did such as MoMA, and they collaborated with universities such as the University of Pittsburgh's Information Science program, the Carnegie Mellion Digital Humanities program, and the Frank Raytche Studio for Creative Inquiry. The faculty of these university programs helped CMOA by giving access to their own collections for research and teaching.

Regarding limitations, the most obvious answer is provided on the "READ ME" file on GitHub, which lists information about the data structure and usage guidelines. Under "Data Integrity," it reads: "Please be aware that the dataset contains incomplete data and/or errors. CMOA staff does not guarantee or provide curatorial approval for these records." At the end of the day, the data collection and formatting was handled by people so it is liable to human error. However, the capabilities of GitHub (version-controlled data and constant updating) tries to limit this as much as possible. Also, during the six months collection process, anyone had the ability to first, correct any information that they felt was unnecessary, and second, retract information they thought was not ready. All they had to do was write one sentence on why the data should not be published. Thus, there could be missing information or different departments could have had different criteria for publishable or unpublishable data. Lastly, all information was taken from paper records which could have had errors before digitization. Artwork provenance, for example, has been known to have gaps or errors with historical information in the past such as Christie's unknowingly handling two Nazi-stolen pieces of art. Overall, it is impossible to consistently verify all the data in the collection.

The Carnegie Museum of Art Data Collection Project had the vague idea of allowing the public to use information for "discovery, inspiration, and innovation, [. . .] to creatively

re-imagine and re-engineer our collection in the digital space" (Fowler, 2017). It is unclear when the data was collected considering the paper records would have come from different points since the museum's first gallery opening on November 5th, 1895. However, the digitization process would have happened between 2015 (the museum's 120th anniversary) and August 8th, 2017 (the publishing of the 'Collections as Data Facets document collections). With the information readily available on GitHub, anyone can answer similar questions to those states on the Art Tracks website overview and much more.

Bibliography

Cmoa. "CMOA/Collection: The Collection Data of the Carnegie Museum of Art in Pittsburgh, Pennsylvania." *GitHub*, github.com/cmoa/collection. Accessed 21 Feb. 2024.

Fowler, Dan. *Frictionless Data*, 8 Aug. 2017, frictionlessdata.io/blog/2017/08/09/collections-as-data/.

"Art Tracks: Art Tracks." *Art Tracks | Art Tracks*, www.museumprovenance.org/. Accessed 21 Feb. 2024.