alj

Disparate Datasets and Data Squishiness: Growing Data Awareness at the National Gallery of Art

Rachel McPherson

This article was based on a presentation that was given at the IFLA-Art Libraries Section Satellite Conference. In an effort to push toward a more multidisciplinary and holistic approach to the data generated throughout the National Gallery of Art in Washington, DC, staff have been investigating new ways in which to amass, organize, and use that data. Segueing from a GIS mapping initiative, one such project sought to amalgamate disparate kinds of data centering around pigment analysis. This led to a multidepartment team trialing software solutions, raising questions and identifying challenges about how the institution creates and standardizes its data, and what changes are needed as we look to the future.

The theme of the 2023 International Federation of Library Associations and Institutions (IFLA)-Art Libraries Section Satellite Conference was *Big Ideas, Challenging Questions*. Two days focusing on how multidisciplinary the cultural heritage industry has become, the challenges that have emerged from the multitude of those interconnections, and what roles art libraries can contribute in this sea change. As our institutions come to terms with the sheer quantity of data that is being created about the collections that they keep in trust, they are looking to find ways in which to analyze, contextualize, and visualize that data.

The National Gallery of Art in Washington, DC has been no exception to this change and has begun initiatives to determine where the institution could grow in terms of utilizing the rich resources pertaining to its collections. This included events such as a datathon, an event where data scientists and art historians were invited to collaborate together, and the creation of the Data Curation and Digital Tools Working Group. This cross-departmental group, which is spearheaded by colleagues from the Center for Advanced Study in the Visual Arts (the Center), IT, and the Library, utilizes staff expertise to examine the possibilities.

Piloting New Initiatives

In 2022 an internal pilot project was formulated called *Mapping Our Museum: GIS@NationalGallery*. The initial objective was to focus on mapping to create a proof of concept in the benefits of harnessing and utilizing collection data using digital tools in new ways. During this pilot, there were six preliminary teams, each concentrating on a project within the museum that could benefit from utilizing maps. For example: mapping upcoming exhibitions and art loans, the diversity of Center alumni, and the locations where Cézanne painted based on an art historian's photo archive.

I was a part of the team Giallorino, which sought to map the geographical use of mineral pigments in artwork. Conceived by the National Gallery of Art's Senior Conservation Scientist Barbara Berrie, the team also included Eve Straussman-Pflanzer and Corinne Annalisa Scala from the Department of Italian and Spanish Paintings in Curatorial, Rheagan Martin, a Fellow at the Center, and myself, the Digital Projects Coordinator from the Library. Barbara had long

 doi:10.1017/alj.2024.11

fostered a desire to see this concept come to life, as she has a close colleague in Rome who has written extensively on yellow pigments used in the Italian Baroque and Renaissance eras.[1] To this end, she had begun a spreadsheet that she could use to organize the data about these yellow pigments around those eras. The data would include the names of the artists that used them, what artworks they were used in (according to conservation testing), the dates and places in which these artworks were created, and ultimately the different types of yellow pigments present or not present, including Lead Tin Yellows (generic, Type 1 and 2, Antimony), and Naples Yellows[2].

Ultimately our research question was: could we pick out the trends of the usage of particular pigments in specific cities? Our goal was not to attribute dates or even places to paintings, or authenticate them in any way, because with art there are differing degrees of certainty that can be obtained. We did not want to speculate that a certain painting was likely to have been painted in Naples because of its use of Naples yellow. However, the presence of that particular pigment could add to the evidence used to observe an artist's movements throughout their lifetime, because they were perhaps using different pigments in different cities as they were available.

1. Seccaroni, Claudio. *Giallorino : storia dei pigmenti gialli di natura sintetica* (Ann Arbor: University of Michigan, 2006).

2. Giallorino is an umbrella term used to encompass those yellow pigments and where the team derived its name from.

Giallorino's Formulation

From the standpoint of a librarian, who is used to dealing with massive quantities of data, it was apparent from the beginning that this project had the potential to be a massive undertaking on a truly infinite scale. With this in mind, the group decided to first focus on yellow pigments within the works of one or two artists. Artemisia Gentileschi was chosen, not only because several teammates were from the Italian Paintings department, but for the fact that there exists a good but not overwhelming number of her paintings. In addition, her career encompassed several major cities throughout Italy, and from there, other artists could be added, such as her father Orazio.

For this initial foray a map was created to show the cities in which the different artworks were created across Artemisia's career. This simple visualization was able to show the use of the different types of pigments and how that timeline matched up with where she had lived throughout her life[3]. Wanting to take this project further, the team set out to amass more data from more artists, since it would be easier to see trends as more data was compiled. This included not only compiling lists of all the known artworks of these artists, but researching any information regarding where those paintings were created, as well as reaching out to the institutions that currently housed them to see if any pigment analysis could be shared.

3. https://timemapper.okfnlabs.org/anon/vqsqd1-use-of-yellow-pigments-in-artemisia-gentileschi-paintings

Spreadsheets are a great way to compile data and are commonly used in digital projects. However, the team quickly learned that there would be a battle with, not only trying to find a way to keep very disparate types of data together and manageable, but also in making the data searchable while also accounting for what the team has now lovingly termed "data squishiness." This term has been coined for instances when the data is not set in stone and firm. For example, dates can float over a span of years, spellings of artist names and the language of artwork titles can vary widely, and even what may seem concrete, such as the scientific analysis of the mineral pigments, can actually be open to interpretation. Oftentimes the context and explanation of the data can be just as important as the data itself, in how and where that data was obtained, determined, or verified.

In the original spreadsheet created by Barbara, she attempted to capture this squishiness by designating a system of 0s and 1s, whether the pigment was definitively identified as being present in the painting, whether it was determined not to be, or perhaps it wasn't truly known. But ultimately, in a spreadsheet it was hard to capture the nuances fully. In order to combat this, and hoping to move beyond spreadsheets, a FileMaker database was created to further the proof of concept. The team hoped to create something where we could show what was possible to achieve with our data, but also what could not be achieved using the tools currently available to us.

Moving Beyond Spreadsheets

The FileMaker database[4] is broken down into different datasets or tables. The Artists table aptly holds information about the artist, such as birth and death dates

4. https://www.loom.com/share/70e206fbc170489ca509eb7fca99bc36?sid=41b60ae2-d5b5-4fb2-aa25-7b3954bda4c8

and locations, cities in which the artist lived in throughout their lifetime, and names of other persons with whom they may have been affiliated, such as a mentor or artist relative. A major advantage to the database was that the data could easily be queried. For example, one could show which artists were active in Florence in the year 1618 and of the seven artists entered into the database at the time, Artemisia Gentileschi and Nicolas Poussin were found to be in that location that year.

An Artworks table exists as well, where all the different artworks are listed. This table is where our amassed data comes into play, as it contains information about the artwork-like titles (including alternate titles, titles in different languages, etc.), when and where the painting was created, and its current whereabouts (whether that be a museum, within a private collection, or even if it has been lost or destroyed). Dropdowns were built in to incorporate that data squishiness and provide context wherever possible, for instance being able to differentiate an artist from their workshop, and specifying when dates may not be exact by using terms such as circa, unknown, after/before, etc. There are multiple supporting tables of information that feed data to these datasets. Ones that list the institutions, including their cities and countries, GIS coordinates, and pertinent contact information in order to reach out for pigment data and feedback.

Also included in the Artworks table is the pigment analysis section. This section was based on the original spreadsheets, splitting apart each of the different types of yellow pigments, allowing a statement about that pigment's presence to be chosen, and providing a citation section in order to trace where that information was obtained. This data can also be queried. For example, if one wanted to find out in which paintings Artemisia used the generic type of Lead Tin Yellow, the database returns the entries where that pigment has been confirmed to have been used. Furthermore, one can find out where she was using that pigment when she made those paintings. For Lead Tin Yellow, Artemisia was using that pigment primarily in Florence, Rome, and Venice. If a query was used instead to look for which paintings Naples yellow has been confirmed to have been used in, and compare that to where those paintings were created, one would find that Artemisia was using that pigment when she was in Naples and when she was in London.

Moving Forward

After sharing this iteration of the database to multiple colleagues, there were a plethora of suggestions as to how to make the database better. However, the main takeaway was the overwhelming interest and excitement in the project and the potential that it had. Therefore, something that seemed to start with a simple objective, such as creating a map from existing data points, has really morphed into something else entirely. While this project has started the wheels turning on what the possibilities could be, it has also highlighted numerous challenges that institutions, including the National Gallery of Art, face in order to create good data and have the ability to harness it to get something usable.

One of those challenges was as basic as a team roster. With different team members coming from different departments, trying to accommodate everyone's desired outcomes could be hard to grapple with and may lead to potential scope creep. For instance, in our project, colleagues in Conservation were very interested in being able to query overlapping data, such as when and where pigments are being used first, how that data was obtained and verified, etc. However, in Curatorial and the Center, their excitement was in visualizing the data, seeing the movement of those artists and how pigments were getting to the artists. While the Library and IT staff exist in the middle, trying to compile, standardize, and validate that data.

Another challenge was that everyone had their own definition of what constituted the term "good data." Among some data resources, one might find date formatting such as "c. 1615" or the "1550s," but spreadsheets and databases struggle to parse those as numbers. Misspellings also proved difficult (e.g. "Lourve"), as well as working with artwork whose titles could be referenced in a multitude of different languages (e.g. Judith Slaying Holofernes vs Giuditta che uccide Oloferne). We found that translating across different languages affected not only titles of artworks, as one would expect, but also the names of artists (e.g. Pietro Pablo Ruebens for Peter Paul Rubens), place names, and especially the names of pigments.

While one can mitigate some of these challenges through the user interface, for example limiting choice to a set vocabulary with dropdowns, there was a concern about introducing a different kind of limitation like the one experienced in the original spreadsheets where context was lost. With spreadsheets and databases, one needs to be careful, so it doesn't seem like a simple statement "x = y" is being made, when that statement is based on a much more complex formula. So, it can be said that "we *believe* this pigment is present in this artwork, because of *this* citation, which shows the XRF stratigraphy"; or that "we *think* so-and-so lived here this year because of a letter from within *this* archive citing the purchase of tools." How do we embrace that data squishiness? An honest skepticism needs to be built in about the data, while at the same time not paralyzing the ability to create or use it. We need to incorporate those levels of uncertainty by providing the context that surrounds that data.

While there are tools existing and emerging that could be used to tackle these very issues, it is sometimes a challenge to incorporate these new kinds of tools into existing workflows. Whether dealing with heightened security protocols, budgetary restrictions, or lack of technical expertise to implement new tools, there is a reason why the cultural heritage sector can be slow to adopt new technology in this burgeoning area.

But where there lie challenges, also lie opportunities. At the National Gallery of Art, our Chief Information Officer Rob Stein has championed the use of what is being called a "walled garden." This is a kind of internal-only development space, as well as a way of moving away from using one product for each project and moving toward having more cross-departmental and connected systems. While that seems incredibly daunting, it will hopefully allow the whole staff to take a fresh look at our way of doing things data-wise. The team can think through what we want to achieve with our data, and whether that can be accomplished with current resources, or whether we need to argue that what is needed is beyond those secured walls.

As we start this shift, having colleagues from the Library, the Center, from Curatorial and Conservation be able to introduce ourselves as, not just demanding customers of IT, but willing partners in trying out these new technologies and providing the expertise and knowledge base to support them, has been very well received. More often library staff and their expertise are starting to be brought into the discussion earlier on. Librarians are able to show that they are fantastic at data. They know about harnessing large quantities of it and do it every day in creating, migrating, and working with it. They know that the source and cleanliness of that data is key-especially for systems that are going to be museum-wide or those that become a resource available externally.

With projects like this, we have started the continual conversation and are working toward the goal of being able to keep cohesive collections of data that answer to the same standards. We will hopefully be able to feed those queries across disparate datasets, visualize our data, and make it usable for future endeavors that we currently cannot even imagine; not just from within the bubble of our own departments, but throughout the whole institution.

Rachel McPherson
Digital Projects Librarian
National Gallery of Art, Washington DC
Email: r-mcpherson@nga.gov
Mailing address: 2000 South Club Drive
Landover, MD, 20785