Art Connections

Business Rules and DB Normalization

Bobb Shields, wde677

Data comes from different sources to describe museums or the art within them. A few large organizations released their collections’ data to the public in the last few years. Included in this set are the Modern Museum of Art, The Met, and The Carnage Melon Institute. Additionally, data sets for the largest museums in the world with their footprint sizes, and another set that attempts to identify all museums in the world.

The purpose of this database is to retain the source data as it is published, then to create views and joins such that its flattened as much as possible so that an API can simply search for art and return a standardized view across the multiple published formats from various museums.

A difficulty to this is the fact that all the different sources have different styles of primary keys:

|  |  |  |  |
| --- | --- | --- | --- |
| **Source Key** | **DB Table Name** | **Variable Name** | **Entity Type** |
| carn\_cmoa | carnige | id | Artwork |
| carn\_cmoa | carnige | artist\_id | Artist |
| carn\_teenie | carnige | id | Artwork |
| carn\_teenie | carnige | artist\_id | Artist |
| met | met\_objects | Object ID | Artwork |
| mod | mmoa\_artists | Artist ID | Artist |
| mod | mmoa\_artworks | Artwork ID | Artwork |
| mod | mmoa\_artworks | Artist ID | Artwork |
| mod | mmoa\_artworks | Width (cm) | Artwork |

Another issue is that the Carnegie institute released two .csv files, one for each location. It was identified during the data dictionary work that there should be a generated field added to identify which source so they can be combined into one table, being the exact same data format.

The Modern Museum of Art data is more normalized such that they release art and artist information separately, with the Met and Carnegie Institute publishing much the same data but artists are not separated. This deserves some analysis to see which museums have many pieces by the same artists and would benefit from normalization themselves, though that is out of scope.

Related fields across all sources for the Artist themselves:

* Artist Name, Nationality, some track birth and death years, and some have deeper portfolio information.

Related fields across all sources for the Art:

* Title, Description, Artist Note’s, Creation Date, Medium, Date Acquired, a URL link, the physical Dimensions, the Department and/or Physical Location, Culture, Period, How it was Acquired, and Catalogue numbers

Related fields for any given Museum:

* Name, Physical Address, Website URL, Year Established, Square footage / m2, and logically a description/purpose/mission statement but that is missing from these datasets.

There is no point to updating the .csv as they come from various museums, so we will write scripts to omit extraneous internal-to-the-specific museum data and transform related fields across the sources to a unified data structure that is fully normalized. A strong requirement for this project is to identify pieces of data that cannot be imported easily enough that simple character encoding issues can be targeted and resolved through a standardized process. An example:

A screenshot of a computer

Description automatically generated

Figure 1 - MMOA Artwork – Normalization Needed! And Character Encoding Issues

Thus the script that works on importing artwork metadata needs to break out descriptions into its own table and checking for duplicates before inserting a new one. That would fix the shown duplication issue of “One from a portfolio of fifty lithographs” but also “Untitled” or “N/A” responses would massively reduce duplication. Mediums and physical dimensions would benefit from this same normalization.

The import script should furthermore publish the import errors and fully print the encoding tables to a file for auditing purposes, then delete them from the database and update an import log table.

Examining the ERD, the relationships between entities looks as such: An artist makes portfolios that contain pieces of art. The art is located in a museum. The relationships should work to allow art from the same artist / in the same portfolio to be at different museums. Breaking out into the portfolio many-to-many linker table should allow normalization of data as found above in Figure 1.