**Project Proposal: Spanish Artists Dictionary dataset**

Pratt Institute

LIS 664 Programming for Cultural Heritage

Alexandra Provo & Diana Sapanaro

**INTRODUCTION**

The Spanish Artists Dictionary (SAD) is a reference source created by scholars at the Frick Art Reference Library. Originally a print publication, the dictionary was formatted as a Filemaker database in the early 1990s and made available through the Frick’s online research portal. The database consists of 5,186 records describing Spanish artists; these records include information about artist name, alternate names, dates (birth, death, and/or activity), field of artistic endeavor, bibliographic references, and Frick Photoarchive holdings.

The data from SAD’s three Filemaker tables has been made available for this project as CSV files.

**DATA ANALYSIS**

We would like to use Python to analyze several of the following aspects of the data in order to create visualizations:

* Artist gender
  + Count the number of female and male artists and create a new CSV.
* Region of Spain
  + Make new lists for each region that include which artists are from that region and create a new CSV.
* Earliest and latest dates (birth/death or activity)
  + Normalize the format of these dates
* Date qualifiers (what the distribution of different qualifiers is)
  + Generate lists of which artists have various qualifiers
* Alternate names
  + Count the number of alternate names for each artist
* Type of artist (watercolorist, miniaturist, etc)
* Photoarchive holdings (quantity of supply, classified, and unclassified holdings)

The data analysis portion of the project will help us to better understand how the factors under consideration intersect (for example, when the female artists described in the dictionary were active). The data will be formatted and cleaned for use in the visualization software Tableau.

**DATA CLEANUP**

We would also like to use Python to perform name matching and disambiguation. Cleaned data, as well as the new data generated during these processes, could then be inserted back into the Filemaker database.

**Problem**

From a SAD record, when a user clicks on one of the three links to a library catalog—the Frick’s library catalog (FRESCO), Arcade (the NYARC catalog), or Worldcat)—the search box that appears in the catalog is populated from a field in the Filemaker database labeled “LC Name.” However, many entries in SAD have incorrect or outdated names in the “LC Name” field, or they do not include a name in the “LC Name” field, in which case the name is populated from the main name field. This produces faulty results (i.e. Goya, Francisco, 1746-1828 ***not*** Goya, Francisco de, 1746-1828) (1181 vs. 2 hits in FRESCO).

**Name Disambiguation and Cleanup**

The proposed cleanup addresses the goal of ensuring that the name being searched when clicking the links to FRESCO, Arcade, or Worldcat matches the name in the FRESCO subject headings list.

The field name would be renamed “LC Name/FRESCO Subject Heading” (or something similar) to capture that the FRESCO subject headings are a mix of official LC and LC-compliant local names. Then, Python scripts would be used to establish matches between SAD and FRESCO subject headings. In other words, the list of names in SAD would be compared with the list of names in FRESCO. If a match is established, the name from the FRESCO subject headings list would be used to populate the “LC Name/FRESCO Subject Heading” field.

Following this, Python scripts would be applied to the remaining names (those that did not match to pre-existing FRESCO subject headings) to perform matching between SAD and LC authorities. Again, if a match is established, the name from the LC authority would be used to populate the “LC Name/FRESCO Subject Heading” field. Another option would be to match these names to VIAF records, which contain LC authority records as well as ULAN records, using Python and the VIAF API.

Finally, the stray names remaining after these matching sessions would be formatted as RDA-compliant names, using a combination of data cleaning software (such as OpenRefine) and Python scripts. These properly-formatted names would be moved from the “LC Name/FRESCO Subject Heading” field to a new field that could be named “RDA name” (or something similar). Putting these names in a separate field would ensure that they are not misinterpreted as official LC or FRESCO names, and instead are understood as local SAD names.

Another option that may be explored is Ecco!, a new name matching software tool developed by Linked Jazz. More information can be found at <http://linkedjazz.org/tools/ecco/> .

**DISCOVERABILITY**

Another project is to establish links to new sources to enhance discoverability of SAD, which in turn would increase traffic to the Library’s digital resources, research databases, and FRESCO.

Links to records from the Frick’s Digital Image Archive (such as <http://images.frick.org/PORTAL/IMAGEINFO.php?server=MTkyLjE2OC4xMC43Mg==&siteurl=&file=/Volumes/digitallab_xinet_5/NEH_grant/photographic_print/POST/folder-11/542_POST.tif>) would be inserted into records when applicable. This could be accomplished by checking the SAD list of names against the Digital Image Archive’s list of names. A list of links could then be generated either from the Digital Image Archive or by using Python to perform web scraping.

Additionally, a Python script would be written to check SAD names against names in DBpedia. This would establish which artists have a presence in Wikipedia and generate 1) a list of articles in which to insert links to SAD entries, and 2) a list of new stub articles to be created in Wikipedia.

**FUTURE PROJECT: DATA CONVERSION**

In addition to data analysis and cleanup, Python could be used to convert the SAD data from CSV to another format. After creating a metadata crosswalk and establishing which schemas to use, the data could be converted to RDF/XML as part of a project to make the dataset available as linked open data (LOD).