Tian Zhong, Qianmian Gai
zhongt@usc.edu, qgai@usc.edu

Project Proposal

Domain & Goals

Domain: The team is planning to build a knowledge graph in the field of contemporary artists, their works and their social media. This knowledge graph will contain the information about the Artwork ID, Name of Artwork, Artist ID, Name of Artist, Social Media Page/Website(url), Artwork Styles/Genres, Year of the artwork, Artwork Price Range, Description, Year of Birth, Gallery, Artwork Size, and the Photo of Artwork.

The goal of this project is to utilize a knowledge graph and data visualization to draw connections and map relationships among works of art. The availability of large collections of digitized artworks can foster a deeper understanding of fine arts, ultimately supporting the spread of culture. When people are browsing for unfamiliar artists, the first thing they want to know is to find the basic information of the artists, the price for their works and their artistic background. What is more, people would like to explore art of similar genre and related artists. A Knowledge Graph that integrates a rich body of information about artworks, artists, genres, etc., in a unified structured framework, can provide a valuable resource for more powerful information retrieval and knowledge discovery tools in the artistic domain.

Datasets

This project will use at least the following four websites to construct the knowledge graph. First, we will collect the index of contemporary artists from *Wikiart*, and collect the information of artists' work (including dimension and price information) from *Artsy*. We will then link the artists with their *wikidata* profiles and *Instagram* accounts and/or their personal websites, and add those attributes to our knowledge graph. We will only focus on the photography artworks. However, since there are over 290,000 photography works on artsy, the size of our data source web page will exceed 4,000 easily. Also, due to the extremely large scale of the artworks, the team might only focus on the artworks created/presented in the 21st century.

Technical Challenge

Our project will solve technical challenges in Entity linkage, user interface design, large dataset operation and data visualization. After scraping the artworks and artists information, we will pair the linkage between the artists in the knowledge graph and their instagram using *The Record Linkage ToolKit (RLTK)*. The correctness of the linkage would be the most difficult challenge for the team, especially for the part linking the artists to their personal profiles. The team may have to use the search engine since there is no existing database for this part. *Neo4j database* would be used to process the large dataset. After the knowledge graph construction, we will design a website as a user interface using *React*, so that people can search for artists they are interested in. Query language *Cypher* will be used to retrieve data from the knowledge graph.