Reverse Engineering the Image Library: the feasibility of using deep learning to identify significance in a 35mm slide collection

This project will apply new techniques in computer science to investigate methods to extract data from digitized 35mm slides from the collection of the Columbia University Department of Art History. The collection consists of approximately 400,000 slides, collected from the 1940s through the early 2000s, covering topics in the fields of art history, archaeology, and anthropology. This project will conduct a trial with deep learning and optical character recognition software to automate the creation of a catalog of the slide collection. We seek a Sparks Grant of $21,835 to experiment with new algorithmic processing technologies to extract and analyze data from digital files, and create an open-source framework for automated processes of data analysis and classification for archival discover across humanities fields.

The obsolescence of the 35mm slide collection has been a prominent issue in visual resource collections for over a decade. As online image collections have grown, institutions have witnessed a precipitous drop in the use of their analog collections, and increased pressure to repurpose the spaces in which these collections are held. Many institutions have pursued a combination of deaccessioning and digitizing these analog resources. However, the tremendous amount of time and money required to manually digitize and catalog a full 35mm slide collection, especially when many of the images may be digitally available already, results in a process that is often rushed. Resources may be chosen for deaccession without systematic surveying, and the retention and digitization policies for legacy collections may be based on personal experience rather than analyzed data. Experimenting on a sample of our collection, we hope to create a tool to streamline the process of turning these legacy collections into accessible resources using automated, algorithmic processes.

This project will apply established techniques in computer science to develop new methods to automate and extract data from digitized collections. We seek to complete a trial with deep learning and optical character recognition software to determine what kind of data we can extract from images and text on digitized 35mm slides, exploring what kinds of databases could be created from our results, and experimenting with image detection technologies to identify and sort digitized collections. Beginning October 1, 2017, we will manually select a representative sample of approximately 1000 slides from our collection, and make a manual catalog of information from the slides to use as a control. We will research the specifications of available open-source deep learning software packages, with preference given to packages with Python interface for prototyping, such as Theano, Tensorflow, and Singa, and a new computer workstation will be purchased for parallel processing. In November, we will begin preprocessing the images, determining what information can be extracted from the files, and creating test databases. For the next two months, we will perform deep learning and performance tests on the sample set of slides, evaluate our results to optimize the process, and run further tests. We hope to achieve a general evaluation of the method and feasibility of implementing these evaluated technologies for automated processing of digitized 35mm slide collections. Results will be presented in a white paper in February.

This project will produce a white paper with an assessment of methods of automation including preparatory cataloging, configuration, and experimentation, and statistical results of different software tests. The paper will explain procedure, report on issues, and describe suggestions and advice for using deep learning for similar endeavors across the humanities. Our hope is to foster similar experiments for the multitude of large 35mm legacy collections. A successful assessment of methodology for automated data extraction and classification will make these tools accessible for other institutions to apply to their large-scale digitization projects. Through the dissemination of the white paper produced, we also hope to initiate methodological competition among scholars so that an optimal method can be determined to save these valuable collections while minimizing effort.

The total cost for the project will be $21,835. At the start of the project, the Digital Curator at the Media Center for Art History will travel to two peer institutions, Harvard and Yale, to gather information on the metadata standards used in the cataloging of their 35mm slide collections. Travel to New Haven and Boston, plus one night of lodging, will cost $380. We will purchase a new computer components for parallel processing for the project at a cost of $1,500. A graduate student in Computer Science and an undergraduate student in Art History will be hired as student casual employees. The Computer Science student will work 12 hours per week researching, investigating, and testing methods for digital extraction and classification with deep learning software, and will be paid $40/hour, totaling $7,680. The undergraduate student assistant will work 10 hours per week to assist with selecting and digitizing the slides to be processed, and will help create a manual catalog for use in testing the efficacy of the computational methods. The undergraduate assistant will be paid $20/hour, totaling $3,200. Fringe benefits are calculated at a cost of $887 and total indirect costs will be $8,188.