Bringing Dark Data to LUX:
Investigating Existing AI Tools for Museum Digitization Applications

Emma Adams

Abstract:
Artificial intelligence (AI) is embedded in the daily lives of many people. Ranging from assisting with writing to predicting weather and traffic patterns, AI has optimized many tasks essential to everyday life. But how can AI be utilized to optimize lengthy museum and archival processes to digitize dark data? This paper explores possible applications of existing AI pipelines and other emerging technologies to assist in the digitization of accumulated data in the Yale Peabody Museum collections. Specifically, we investigate the use of AI-powered optical character recognition (OCR), named entity recognition (NER), multimodal AI, and image-classification labeling.

Key Words:
Artificial Intelligence, Machine Learning, Digitization, Museums, Collections, Archives, GLAM

Introduction:
Employing AI and similar tools in the museum digitization process is of interest to the museum and archival world as it would allow for additional public access to large collections of data through online discovery portals and would increase the efficiency that data is produced for analytical purposes.

Marsh Paper OCR Text Extraction & Analysis
Many different applications of AI already exist, and many more are currently in development. One of the most promising forms of AI include optical character recognition (OCR) for handwriting-text recognition (HTR) purposes. Many different organizations have created their own versions of OCR technologies, including Google Cloud's *Vision AI*, AWS's *Textract*, Azure's *OCR Engine*, and *Transkribus*. To explore the applications of OCR, we utilized Google Cloud's Vision AI API to extract the handwriting from select letters from the O.C. Marsh Papers—a large collection of handwritten correspondence to Othniel Charles Marsh, former professor of paleontology at Yale College and former president of the National Academy of Sciences—and analyze the text for specific keywords, people, or places.

Using the Google Cloud platform, we were able to create a project to store relevant files, and use the Cloud Vision APIs to apply OCR and extract the UTF-8 text from a select subset of the O.C. Marsh papers. Specifically, applying the Cloud Vision OCR on a PDF or TIFF file results in a JSON file which contains the page annotations from each page of the inputted file. Using Python, we were then able to parse the JSON file and retrieve the text and respective page numbers, and output to a CSV file for further analysis. Using the resulting transcriptions, spaCy's named entity recognition (NER) was applied, extracting named entities (people, places, dates, etc.) from each page, and adding them to the CSV file.

Using Google Cloud Vision, we found the overall transcription of cursive, handwritten documents to be fairly accurate, spelling many of the words incorrectly, but capturing some of the words correctly. For example, the name *Marsh* in the letters were often captured as *March* by the Vision API. This is a common issue with the transcriptions of handwritten materials— the majority of the word will be correct with a few incorrect letters. Additionally, the Cloud Vision API is fairly simple to set up and use, and would be easy for someone with little programming or Python experience to use. The NER API developed by spaCy is also easy to install and use, and is completely free. The API was able to capture names, dates, places with high accuracy, but was still subject to the misspellings produced by Cloud Vision; for example, when Cloud Vision outputted *March* instead of *March*, the NER extracted *Marsh* as a named entity.

OCR technology is not yet perfect for handwriting, but it is still a useful tool for digitizing old documents and dark data. It greatly decreases the length of the transcription process, allowing more documents to be annotated. Using OCR along with a large-language model (LLM) might be useful to help output transcriptions that consider context and what spelling is most likely to be correct. Additionally, the outputted transcriptions and recognized entities will be useful in linking data (names, places, etc.) on public access portals, such as Yale University's LUX.

Specimen Drawer Multimodal AI Analysis

Another new innovation in the AI field is multimodal AI, which processes different forms of input (text, audio, images, etc.), and outputs various media and data in varying forms. For example, multimodal AI allows you to upload an image or document, ask a question about said

document, and output some sort of response in the form of an image, text, or other medium. To investigate the possible implications of multimodal AI, we utilized Google Vertex, Google Cloud's multimodal AI. Specifically, we used Google Vertex AI to examine various images of invertebrate paleontology specimens and extract the text from and generate descriptions of their respective labels.

Using the Google Cloud platform and Python, we were again able to create a project to store relevant files. Then we used Google Vertex AI, a multimodal AI, to examine a specific image file for an answer to a specific question or prompt. For example, to extract the text from a label on a specimen from an image and generate a description of said label, we would store the image in Google Cloud Storage and then prompt Google Vertex AI "Identify the label on the specimen in the image, ignoring the ruler at the bottom, and extract the label's text or numbers. Briefly describe the label's color, medium (paper, paint, etc.), and shape. Put in the format: label text - description." This, considering the image at the specified Uniform Resource Identifier (URI) and the specified prompt, then outputs a response (in the form: label text - description) which can be easily delimited and stored in a database.

The use of a multimodal AI works effectively for projects similar to this, in which you need individually specific yet standardized information from a large set of unique images. However, when prompting any multimodal AI service, including Google Vertext AI, we found that it is essential to be extremely meticulous when creating prompts. For example, you should specify the length of the response, the format, etc., otherwise it can vary from response to response. Using multimodal AI in similar ways can assist in the digitization of specimens and their labels and decrease the time and effort currently required.

Yale Peabody Museum Image Classification

AI also allows for computer vision, which is the methods used by computers to understand, process, and analyze digital images. With computer vision, AI can distinguish between images, placing them into various predefined categories using image classification. To explore the possible applications of image classification, we utilized Google's Teachable Machine, which creates a machine learning model which classifies images into various categories set by the user.

Using Teachable Machine, we created machine learning models which can distinguish if the image was taken before or after the 2019-2024 museum renovation and can determine which of the select exhibits an image depicts.

Using Teachable Machine, we were able to easily create a machine learning model for exhibit classification by creating a project, creating labels to represent select museum exhibits (Great Hall, Burke Hall, David Friend Hall, Dioramas), uploading images to each label, and training the model with a customizable number of epochs and learning rate. We utilized the same process for creating a model to differentiate between if an image was taken before or after the renovation. We were then able to download the usable model endpoints using Python, Tensorflow, and Keras to classify images into their various categories. Both models can even be run simultaneously using multiprocessing or processed based parallelism.

Using an AI tool to classify images will be helpful in reducing the amount of time spent categorizing images. Running a large quantity of images through the model at a time can allow for the creation of a database of images that can be easily searched. Additionally, the specific categories the image is predicted to be by the model can be embedded into the image as a metadata tag, allowing for more use of the image in future projects or digestion into in online discovery portals like LUX.

Moving Forward with AI

As an emerging technology, AI has many possible implementations into a museum or archival environment. From categorizing images to examining specimens, AI can be utilized to reduce the length of the processes used to digitize dark data. Specifically, many forms of AI, including speech recognition, OCR, computer vision, etc., can be combined with each other or other emerging technology to simplify digitization processes. However, AI is not yet perfect; it makes mistakes and is subjected to the bias of its creators and training data. Additionally, AI is not culturally aware, meaning it can generate insensitive descriptions of cultural artifacts of objects as it does not understand the human context behind them. While a turn-key AI system sounds ideal for some digitization projects, it is essential to still implement some human assessment of results in order to prevent biased outputs.