Derived Metadata for Early 19C Illustrations: ACS Grant Final Report

Stephen W. Krewson

August 3, 2020

Process and deliverables

The "Deriving Basic Illustration Metadata" project has successfully concluded with the creation of a large and novel dataset of illustration metadata. The dataset was produced in four stages using two retrained convolutional neural networks as well as one standard model (InceptionV3).

The stages were as follows:

1. **[Find illustrated pages]** Identify all Google-digitized volumes published during the years 1800-1850 (inclusive). From this set of 500,013 volumes, use OCR metadata to find pages likely to contain illustrations. Classify each candidate page with a retrained CNN and keep pages labeled with `inline_image` and `plate_image`. My midpoint report describes this stage in greater detail and can be found here.
2. **[Extract illustrated regions]** Mask-RCNN models slide a window across an image input looking for regions that activate a target or "foreground" class. Regions that activate past a certain threshold are estimated to be regions of interest (ROIs). The rest of the image is considered to be the "background." Using a Mask-RCNN model retrained with annotated 19C page images (i.e. bounding boxes around illustrated regions), we extracted more than 2.5 million ROIs from the pages identified in Stage 1.
3. **[Generate lower-dimensional representations]** Comparing images programmatically requires lower-dimensional numerical representations. We used the InceptionV3 CNN to turn each ROI JPEG into a 1000-dimensional vector. This allows a distance metric to be applied pairwise to any two images in the dataset.
4. **[Build indices and visualize]** We used the Annoy library from Spotify to build an index of (approximate) nearest neighbors from the vector representations generated in Stage 3. This index allows calculation of the k most similar image vectors to an input image vector. Notebooks for using and building Annoy indexes have been provided.

The key deliverables of this projects are the following:

- A CSV file identifying all illustrated regions from HathiTrust volumes published between 1800 and 1850

- A nearest-neighbors index created from vector representations of these 2.5M regions of interest (ROIs)
- Sample notebooks for working with the metadata and index files

The metadata and index files are included in the project's Zenodo repository. The notebooks and project code can be found on GitHub. For more detailed usage information, consult the README files for these repositories.

Project assets

The following table gives a basic sense of the amount of data processed.

Project statistics	
Total volumes processed	183,553
Total candidate pages	1,922,602 (685+ GB)
Format = JP2	1,901,456
Format = TIFF	21,269
Label = `inline_image`	1,077,544
Label = `plate_image`	845,181
Total ROIs (JPEG)	2,584,888 (553+ GB)
Total ROI vectors (ndarray)	2,584,888 (15+ GB)

The following table describes the project files hosted on Zenodo. At this time, image assets are not publicly available.

File	**Description**
`google_ids_1800-1850.txt.gz`	A subset of the July 2019 Hathifile containing all volumes 1800-1850 that were digitized by Google. These volumes were the basis for all subsequent steps.
`hathi_field_list.txt`	The Hathifile column names
`stage1_fastai-retrained-cnn.pkl`	A convolutional neural network (CNN) retrained with 19C page images. Given a candidate page, returns the likeliest of 10 class labels. Only images labeled `inline_image` or `plate_image` were retained in Stage 1.
`stage2_mask-rcnn-bbox-weights.h5`	A Mask-RCNN model developed by Matterport. The model was retrained with page images for which illustrated regions were annotated with a bounding box. Given an image, the model predicts regions of interest (ROIs) that are likely to contain illustrations. These ROI bounding boxes are used to crop the input image.
`roi-vectors.tar`	1000-dimensional `numpy` arrays (`*.npy`) representing the cropped images (ROIs) from Stage 2. These vector representations were derived using InceptionV3, a standard image classification CNN. Their shape is `(1,1000)`.

File	Description
`early-19C-illustrations_metadata.csv`	Each row of this summary table corresponds to one of the 2,584,888 ROI crops. The fields are: `htid`, `page_seq`, `page_label`, `crop_no`, `vector_path`.This is allows easy browsing of a given page on Hathitrust: `https://babel.hathitrust.org/cgi/pt?id=<htid>&view=1up&seq=<page_seq>` . The `crop_no` field reflects the possibility that a page could have multiple ROIs on it.
`early-19C-illustrations_full-index_list.txt.gz`	A list of all vector files used in the creation of the full-dataset Annoy nearest neighbors index. The order in this file provides the integer indices from `0`to `n-1` for each vector in the `.ann` index.
`early-19C-illustrations_full-index.ann`	A memory-mapped Annoy nearest neighbors index created from `early-19C-illustrations_full-index_list.txt.gz`. Should be accessed with `AnnoyIndex(1000, 'angular')` to match the dimension and metric parameters from when it was built. The index was built with 100 trees and is very fast for finding a reasonable number of neighboring vectors (<100 works well).
`pixplot-metadata_munroe-francis.csv`	A metadata file derived by searching `google_ids_1800-1850.txt.gz` for publishers (the `imprint` field) matching the firm "Munroe [and] Francis" (and variants). For the 360 matching `htids`, 1477 ROI crops existed. While the image data is not available through Zenodo, the CSV contains the `*.jpg` filenames that follow the project's conventions. The fields conform to those used by the PixPlot viewer: `filename` (path to image file), `label` (name of the subset: in this case, "munroe-francis"), `description` (volume title), `year` (volume year of publication).

The following notebooks are available via the project's code repository: https://github.com/htrc/ACS-krewson:

- `find_page_neighbors.ipynb` — Given a `htid` and page sequence number, return metadata for the `k` most similar images in the project dataset. This method is useful when browsing HathiTrust, since the `htid` and `page_seq` arguments are displayed in the viewer URL. Any input pages must have been processed by the project. Future methods for out-of-dataset inputs are planned (but these require extracting and vectorizing the input page image).
- `visualize_query.ipynb` — Query the Hathifile subset used by the project (e.g. search the `imprint` field for a particular publisher) and reformat the results into metadata that can be used by the PixPlot visualizer. Code for building a small nearest-neighbors index is also provided. In many cases, it is more useful to run similarity comparisons on small portions of the data — for instance, the illustration styles used by different publishing houses over time.

Discussion

The project was very successful in meeting its initial goal: a tabular report of all illustrated pages in a 50-year sample of HathiTrust. The next stage, innovative in its use of Mask-RCNN, was also a success. The cropped ROIs are reasonably accurate across a range of illustration types. Further training will only improve this processing pipeline.

Having a large corpus of illustrations opens up new questions for historians of print media. Consider what can be learned from looking at *all* illustrations commissioned by a publishing firm over a fifty year period. Illustrations from the Boston firm of Munroe and Francis (Figure 1) demonstrate the investment of 19C publishers in *series* of engravings, many of them relating to cultural and natural history. In some cases, access to a set of engravings was the impetus for commissioning the text of a book. Competition with other firms most likely had a decisive effect on illustration decisions.

I plan to visualize the Munroe and Francis illustrations in comparison with those from other regional publishers. Using both the JPEG and vector representations will reveal how different publishers carved out specialty subjects in their book lists (or perhaps closely tracked popular genres and forms). This research can, of course, be done by analyzing the titles and text of these volumes. But starting with the illustrations defamiliarizes the problem and allows new insights.

Unfortunately, the image files remain difficult to access and work with. As the project drew to a close, I was able to request small samples (as above) but the size of the images and copyright considerations continue to be barriers. Moreover, the computing resources necessary to derive the metadata for this project are prohibitive for individual researchers.

For computational work on historical illustrations to progress and sustain a community of researchers, digital libraries like HathiTrust will need to provide IIIF-style APIs for downloading bounding boxes from page scans. Ideally, ROI labels and vector representations will be able to be stored alongside the page assets. This way the quality of illustration estimates can be continually improved. Shared visual metadata would greatly assist a more scientific approach to validating vector representations and nearest-neighbors indexes, which are sensitive to parameterization and difficult to interpret.

My hope is that illustration locations and representations eventually become first-class objects, just like extracted text features.

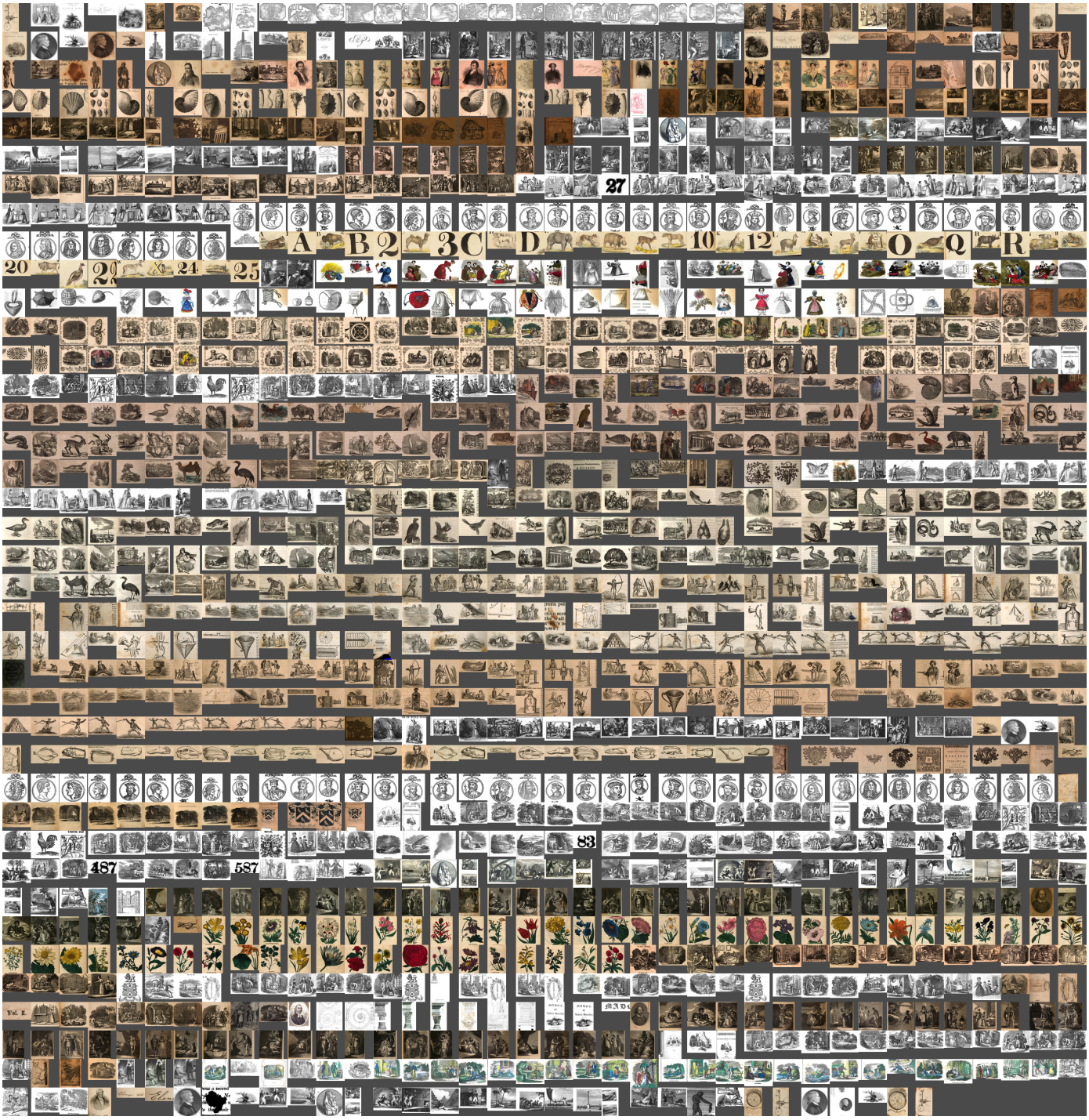

Figure 1. Montage of 1477 illustrated regions, selected by eaching for “Munro and Francis” and variants in the project Hathifile excerpt

Acknowledgements

I am grateful to Ryan Dubnicek and Boris Capitanu for their patience and expertise as this project stretched across a difficult year. They deserve the credit for this project succeeding in meeting its stretch goals, despite inconsistencies on my part. My thanks to Eleanor Dickson Koehl for perceptive questions about the project’s place in the wider world of DH research. I am thankful to Doug Duhaime for his advice re: vectorization and for bringing this grant to my attention. Damon Crockett’s `ivpy` package was invaluable for creating montages of images.