Data Readiness Questionnaire

(fill this out after the Use Case Assessment Worksheet)

Name & Dept: Jamie Mahowald, John W. Kluge Center

Use Case: Contrastive language-image pre-training (CLIP) for search and discovery in the Geography & Maps collection

Date:

1. Data: *Do you have data to support this use case? If so, describe any data that you can use as **target** data, i.e., data that an AI model can process for this use case. Do you have data that can **train** an AI model (data that already has the features you'd like the AI models to be trained on) and data that can be used to **verify** that the output of models are correct?*

Yes. The data that we use to fine-tune the existing CLIP model comes from 50,000 randomly selected caption-text pairs taken from the G&M collection. The caption is generated deterministically (i.e., no AI input) from G&M metadata for each selected item. Given the procedure for training a contrastive model, the same data can be used both to train and test the model. After fine-tuning, embeddings are calculated for all 562,842 items in the Division's 56,554 collections.

2. Composition: *What is represented in the data? Describe the language, time period, genre and other descriptive information about the intellectual content of the data.*

The maps used in fine-tuning span a wide range of dates, styles, and uses. The dates range from 1013 to 2018, focused around the beginning and middle of the 20th century to mirror the G&M collection. Genres include historical, decorative, insurance, topographical/geological, economic, transit, and nautical; they are made by a variety of sources, from artisans to insurance companies to government entities. They are primarily in English, though there are subsets in French, Italian, German, Chinese, and Latin, among others.

3. Compilation: *How was the dataset compiled? E.g., via API query or bulk download? When? With what tools or expertise?*

The dataset was compiled using a bulk download script I (Jamie Mahowald) created in Python. It reads from a series of CSV files provided by Rachel Trent that included IIIF and resource data, downloading each image and associating with it a caption generated deterministically from the resource's metadata accessed through the loc.gov API. Please see G&M resources for more information on data provenance.

5. Pre-processing: *Describe the steps and any transformations used to create the dataset. E.g., text from a digitized document that was OCR'd. When and with what tools was the data transformed? Was the data cleaned or normalized? If so, how?*

After loading the CSV files that include IIIF and resource information on each image, the script does the following:

- Initializes and manages a randomized list of indices for images to be processed based on the CSV file's length. It ensures that images which have already been processed in previous runs are not repeated.
- Constructs the caption for an image by processing its title and metadata fetched using the loc.gov API. It also checks if the title needs prefixing based on whether the item is a map and includes additional notes if available.
- Manages the download process in chunks of 500 using multiprocessing to improve efficiency. The multiprocessing pool is set with five parallel processes.
- Implements error handling to continue processing even when some downloads fail. Records the progress after each chunk to a file (last_chunk_index.txt) and logs each processed row to another file (row_log.txt).
- Measures the total execution time for downloads and logs this at the end.

4. 6. Data provenance: *Describe the relevant background on where the data comes from, why it was created, by whom, where, and when. Include any version information and if the data is used in other systems.*

Please see response to 3. Compilation.

8. Structure & Storage: *How is the data structured? E.g., in XML, CSV, unstructured text, etc. Does the structure follow any standards? If so, what are they? How and where is the data stored?*

After images have been downloaded to a directory using the previous script, the data is structured loosely into two list-type objects: one enumerates path names for each image, and the other enumerates captions corresponding to the path names. In accordance with PyTorch dataset standards, the two lists are then fed into a PyTorch DataLoader for fine-tuning. The data are stored on a Zenodo repository used for the rest of the project.

9. Characteristics, Patterns, Labels: *What are the characteristics or patterns the AI system will detect in the data? Describe the data elements the AI will predict or output for this use case?*

The AI is intended to recognize map features like age, purpose, publisher, geographic features, color, condition. The following is an instance of an image, its path name, and its associated caption:

ORISKANY FALLS, N.Y.

images/train/service_gmd_gmd380_g3804_g3804o_pm010990.jpg

A map of Oriskany Falls, N.Y, from 1891. Aerial view of village. LC copy imperfect: Torn, taped, fold-lined, faded, stained at lower edge. Includes index to points of interest.

The model can then recognize the features mentioned in the caption and progressively learn them over several epochs.

How were the patterns labeled? Are they naturally occurring, did experts label the patterns, or did unskilled or crowdsourced staff or volunteers label the data elements? What was the incentive structure for the labelers, if any.

The patterns are labeled using their IIIF identifier. We choose the IIIF identifier because it was easier to extract from it an object's resource identifier than vice versa and ensured uniqueness across objects.

Use this questionnaire individually, in workshops, or in groups to document how data can impact an AI system. The lack of training data is a common challenge in AI. In general, the more about the above elements you can documented about your data, the more ready it will be to support an AI use case.

Data Readiness Questionnaire

(fill this out after the Use Case Assessment Worksheet)

4. People: *Who is depicted in the data? Is there any PII in the data? Are people depicted in the data described in a potentially outdated or harmful way? Are the people depicted in the data aware their data will be part of an AI system?* Since the entire dataset is cartographic, neither people nor PII are depicted in any of the data.	7. Restrictions/Controls: *Who owns the data? What is the Copyright status of the data? Is the data restricted by privacy, confidentiality, license, or other terms?*	10. Sampling and known biases or imbalances: *What sampling method was employed to create the dataset, if any? What are the known biases in the dataset? E.g. language, geographic, demographic, historic.* The dataset was sampled randomly from the entire G&M digital collection with a large enough size to ensure a variety of languages, features, and intentions were represented. The dataset will skew toward larger collections (e.g., fire insurance maps, highway maps, certain historical collections), but the dataset is a large enough portion of the entire collection to ensure comprehensive representation.

Use this questionnaire individually, in workshops, or in groups to document how data can impact an AI system. The lack of training data is a common challenge in AI. In general, the more about the above elements you can documented about your data, the more ready it will be to support an AI use case.