Data Readiness Questionnaire

(fill this out after the Use Case Assessment Worksheet)

Name & Dept: Aisaiah Pellecer, Ilayda Dogan, Shauna-Kay Harrison — OCIO Digital Innovation Division

Use Case: Object detection and Segmentation

Date: 6/25/2024

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?*

Target Data: Images from the Library of Congress' Free to Reuse collection were used as target data. Using the Faster R-CNN Object Detection and EfficientSAM Segmentation models, objects within an image scene will be extracted. Both "target data" and "use case data" will be used interchangeably in this assessment to refer to the images pulled from the API.

Pre-trained Model Datasets: Because pre-trained computer vision models are being used, no training or validation data will be handled on this end. Since this is a multi-step process that transforms the target data, it is critical that the datasets used to train the models are described as well. The Faster R-CNN model is trained on the Microsoft COCO dataset and the EfficientSAM on the SA-1B dataset.

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.*

For this demonstration, images (photos and prints) from the Free to Use collection were requested via the LOC API. They have been filtered to only include those related to the Washington D.C./DMV area. Other than this focus on D.C., no constraints were set on the time period, genre, or other aspects of the images from the data package.

Regarding the data used for model training, the MS COCO dataset, used to train the Faster R-CNN model, represents common objects in context. This dataset contains 2.5 million labeled instances in 328k. It aims to categorize objects within images into those easily identifiable by a 4-year-old.

The EfficientSAM model was pre-trained on the SA-1B dataset, which consists of 11 million licensed and privacy-protecting images and 1.1 billion segmentation masks. Images were sourced from a third-party provider, and there is no publicly available image metadata.

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?*

Target Data: Images from the Free to Use collection underwent no pre-processing. Besides the location criteria, images were selected based on the highest resolution available from the LOC API.

Pre-trained Model Datasets:

- **MS COCO Dataset:** Images compiled from Flickr underwent filtering stages that removed iconic images of categories. Images with rich contextual relationships were used to direct the compilation and preprocessing. In addition, the data pipeline and processing extensively utilized Amazon Mechanical Turk, an independent contractor service.
- **SA-1B Dataset:**Images in this dataset were compiled and processed using a 'Data Engine.' As described in the **Segment Anything paper**, the creation employed a model-in-the-loop dataset annotation approach; essentially, the data engine iterates between using the Segment Anything Model to assist data collection and incorporating newly collected data. Moreover, images were resized (shorter side is 1500 pixels), and to ensure privacy protection, faces and license plates used within the dataset were blurred. Further information can be found in the **Dataset Card.**

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?*

Target Data: The images were pulled using the Library of Congress' API, structured in JSON format.

Pre-trained Model Datasets: None of the training datasets were stored or directly used in this project. The documentation points to the use of JSON and COCO run-length encoding (RLE) format.

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 system is composed of two parts: Object Detection and Segmentation. The system will leverage the COCO Dataset to detect objects within its scope of 91 object categories. The segmentation process will isolate the item to generate the mask, using box coordinates provided by the object detection model.

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.

In the COCO dataset, annotations were crowdsourced via Amazon Mechanical Turk. There are no notes regarding compensation. As for

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)

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

Target Data: The target dataset was compiled using the LOC JSON/YAML API and comprises all digitized selections from the Free to Use and Reuse Sets that are available through the API. The dataset itself was created by AVP and its contributors, LC Labs, and the LoC Prints and Photographs Division.

Pre-trained Model Datasets: The training datasets were not downloaded directly, but the models leveraged in this exercise were pre-trained on them.

- **MS COCO Dataset:** Images for this dataset were compiled from Flickr. The documentation does not specify bulk download or API query.
- **SA-1B Dataset:** The images for this data originate from a third-party source. Masks in this dataset were generated by the Segment Anything Model (SAM) on 11 million open-world images; masks were verified by human ratings and various other 'data engine' stages. There are no class labels on the masks.

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*

Target Data: The images were curated by the Library of Congress staff. All of the images are believed to be part of the public domain. The collection is actively growing, with new sets added every month.

Pre-trained Model Datasets:

- **MS COCO Dataset:** The dataset was published by Microsoft and created by associated researchers (COCO Consortium) in 2014. It was created to support computer vision by defining common objects.
- **SA-1B Dataset:** The dataset was created by the FAIR team of Meta AI, April 2023.

the SA-1B Dataset, masks were inferred automatically by the SAM. No labels were created.

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?*

Target Data: In the use case images, there is a wide range of individuals depicted within photos and prints. There is no documentation regarding privacy protection or PII in the images. The individuals depicted in the image data are unaware that data about them will be a part of the AI system. In addition, it is important to note that there is a content advisory acknowledging harmful terminology in historical materials.

Pre-trained Model Datasets: As mentioned previously, the images compiled in the pre-training datasets are from open-world data. It is unlikely that individuals in the datasets are aware that data about them will be part of an AI system. Regarding PII, the SA-1B dataset is privacy-protected.

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?*

Target Data: In regards to the restrictions and controls of the data, the only data used directly within the repository will be images from the Free to Use and Reuse Collection. As the name of the data packages and its documentation state, the content has been designated by the library as either part of the public domain, has no copyright, or has been cleared by the copyright holder for use.

Pre-trained Model Datasets Licenses:

- **MS COCO Dataset**: The annotations in this dataset along with this website belong to the COCO Consortium and are licensed under a Creative Commons Attribution 4.0 License.
- **SA-1B Dataset:** Masks are given in the COCO run-length encoding (RLE) format, and do not have classes. By downloading the datasets you agree that you have read and accepted the terms of the **SA-1B Dataset Research License**.

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.*

Target Data: No comments can be made regarding the Free to Use collection images, as it comprises a broad selection of public domain/free images.

Pre-trained Model Datasets:

- **MS COCO Dataset:** As mentioned previously, rich contextual images were used to direct the sampling. General imbalances exist regarding the number of object category instances (e.g., thousands of images portraying humans compared to a mere hundred of less common items like a hairbrush). Analysis by **Zhao et al.** highlights the dataset's skew towards lighter-skinned individuals and male individuals.
- **SA-1B Dataset:** Given the model-in-the-loop dataset annotation, or 'data engine', involved in the creation of the dataset, biases may exist in the dataset.

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.