**Data Processing Plan Template**

*This template is provided to help partners and vendors understand the documentation and planning requirements for processing Library of Congress data in the context of experimentation and research. After an experiment is approved or awarded and before any data processing tasks are performed, vendors and/or partners shall submit an initial draft of this template to the Library for review and discussion. A final version of the template shall be delivered after the data has been processed with all of the relevant information completed. Each distinct data set that is used in an experiment will require a unique data processing plan.*

**Section A: General (required)**

|  |
| --- |
| **A1: Goals of experiment. (consult Library/task order)** |
| *Fill in based on the Library of Congress Statement of Work or Task Order.* |
| **A2: Describe the scope of the intended workflow or pipeline. (consult Library/task order)** |
| *Fill in based on the Library of Congress Statement of Work or Task Order.* |
| **A3: Data delivery format and specifications for data generated in the experiment. (consult Library/task order)** |
| *Fill in based on the Library of Congress Statement of Work, Task Order or directive.* |
| **A4: Description of intended use** |
| *Please describe how the data will be used in the experiment.* |

**Section B: Data Documentation (required)**

Please fill out a complete chart *for each existing dataset under consideration for use in the experiment.* All experiments must have Sections A and B filled out. If the experiment involves machine learning or other artificial intelligence, Section B3 and Section C must also be filled out.

|  |  |
| --- | --- |
| **B1: Description of Dataset** | |
| a) Title of dataset | *Please create a name for the dataset, or use the current name if the dataset represents an existing collection or dataset.* |
| b) Composition   1. Please describe the dataset’s technical composition, including file type, content type, number of items, and relative size. 2. Please describe the language, time period, genre and other descriptive information about what intellectual content the dataset contains. 3. Please also include relevant background context about the composition of the dataset. For example, a dataset may be organized as a single spreadsheet containing metadata about a collection or it may be a series of folders containing images derived from a particular source. |  |
| c) Provenance   1. Where did the information in this dataset originate? Please include relevant links where possible. 2. Include any version information if available. |  |
| d) Compilation methods   1. How is/was this dataset compiled, when, and by whom? 2. Please include technical details of how the dataset is/was compiled, e.g. loc.gov API query, bulk download. |  |
| e) Preprocessing steps   1. (How) has this dataset been classified, cleaned or otherwise prepared for the experiment? 2. How was material selected for inclusion or exclusion in the dataset? 3. Is the data organized according to a schema, content standard, or other standard? If yes, which one? |  |
| f) Potential risks to people, communities and organizations & strategies for risk mitigation:   1. What potential risks or harms could result to people, communities and organizations from processing this dataset in the experiment? (For example: searchable access to individual names and places could expose personal identifying information of private citizens.)    1. How will the experiment team mitigate these risks? (For example: the team will select data that is over 125 years old to include in the experiment.) | |
|  | |
| g) How will the experiment team address outdated or potentially offensive terms or elements of data that may be harmful if encountered by human users? | |
|  | |
| h) Copyright, licensing, rights, and/or privacy restrictions   1. Describe in sufficient detail limitations on any intellectual property or privacy or other restrictions that will affect the Library’s (or the public’s) subsequent use of any data processed. |  |

*Will the dataset be used in conjunction with machine learning or artificial intelligence processes? If yes, please fill out all of section C and section D.*

**Section C: Documentation of a dataset for machine learning or artificial intelligence processes**

|  |
| --- |
| 1) Please describe the purpose of this dataset with relation to the ML/AI workflow. Explicitly address if it is being used as training, validation, or test data. |
|  |
| b) For training data:  1) if the model is pre-trained, describe the data on which it was trained;  2) if the model will be fine-tuned, outline the data involved in this process;  3) if the model is being trained from scratch, outline the plan for creating training data. |
|  |
| c) If creating training data using **volunteers or paid participants (e.g. via crowdsourcing),** please describe the workflow and incentive structure. |
|  |
| d) If validating training data using **volunteers or paid participants (e.g. via crowdsourcing),** please describe the workflow and incentive structure. |
|  |
| e) Document any known gaps in the dataset, such as missing instances or forms of representation. Address possible sources of bias in the dataset resulting from these discrepancies.   1. Describe any steps taken to remediate or address gaps or bias in a dataset used in the ML/AI processing or the experiment overall. |
|  |

**Section D: Documentation of ML model (required for experiments involving machine learning or artificial intelligence)**

All experiments involving machine learning or artificial intelligence must complete the chart below for *any* models under consideration for use in the experiment.

|  |  |
| --- | --- |
| **C1: Machine Learning or Artificial Intelligence Model** | |
| a) Model Details |  |
| b) Intended use |  |
| c) Limitations |  |
| d) Copyright and licensing details for the model |  |
| e) Link to documentation |  |
| f) Predicted performance metrics (range) |  |
| g) Actual performance metrics |  |
| h) Audit schedule (how often and how many times will performance metrics be checked?) |  |
| i) Definitions of successful algorithmic performance. Specifically, performance evaluation factors and accuracy and performance results at each stage of the workflow and for each overall pass through the pipeline. | |
|  | |
| i) Workflow or pipeline description and diagram, including plans for conducting annotation and validation processes. Overview of supervised or unsupervised machine learning. | |
|  | |