Implementation Guidance to Federal Agencies Regarding Enterprise Data and Source Code Inventories

Objective – The objective of this implementation guidance is to provide Federal Agencies the criteria for them to consider in developing a prioritized inventory of their data sets and models, and estimating the work necessary to make them discoverable and useable by the non-Federal AI R&D community.

Background – The Executive Order on Maintaining American Leadership in Artificial Intelligence "Directs the Heads of all agencies to review their data and models to identify opportunities to increase access and use by the greater non-Federal AI research community in a manner that benefits the community, while protecting safety, security, privacy and confidentiality. Specifically, agencies shall improve data and model inventory documentation to enable discovery and usability and shall prioritize improvements to access and quality of data and models based on the AI research community's user feedback."

Improving the discoverability and usability of Federal data is also consistent with the Cross Agency Priority Goal #2 of the President's Management Agenda which is leveraging data as a strategic asset.

To assist Agencies in scoping and prioritizing these improvements, the President has directed feedback from the public be solicited on needs for additional access to, or improvements in, the quality of Federal Data and Models as well as soliciting inputs from the Agencies on their known barriers limiting access to and quality of their data and models.

While feedback from the public is pending, an initial report has been completed with the inputs from various agencies to identify known barriers. This report categorized these known barriers for both data and models into three categories; discoverability, usability and governance. This initial implementation guidance to Federal Agencies is intended to address these barriers and for Agencies to utilize in:

- Prioritizing the data sets and models under their purview for enhancement.
- Assessing the level of effort needed to make necessary improvements in data sets and models, against available resources.
- Developing justifications for additional resources.

Prioritization Guidance for Improving Federal Data and Models – Federal Agencies should consider the following when prioritizing which of their data sets and models to improve for AI R&D discovery and usability:

- Input from the non-Federal AI R&D community from both the Federal Registry Notice and subsequent feedback.
- Data sets and source code that support the individual agencies goals for AI R&D as depicted on AI.gov.
    i. Transportation
    ii. Healthcare
    iii. Manufacturing

iv. Financial services
v. Agriculture
vi. Weather forecasting
vii. National security & defense

- Data sets and source code that address Fundamental R&D challenges in AI as discussed in The National Artificial Intelligence Strategic Plan: 2019 Update
- Principles and Practices of the Federal Data Strategy
- The cost (labor, technology, time) in the improvements necessary for data sets and models.

Guidance for Improving Discoverability of Agency Datasets for AI R&D

Currently Federal Agencies are required to provide comprehensive Enterprise Data Inventories with standard metadata[1] for harvesting by Data.gov that is consistent with DCAT and schema.org standards, so that search engines and tools like Google Dataset Search can index Federal data sets for discovery. To increase the discoverability of Federal datasets for AI Research & Development purposes, agencies shall inform their various offices/department heads associated with R&D that data developed (or procured) should be captured in Enterprise Data Inventories. In addition to the minimal metadata required for agency Enterprise Data Inventories, agencies shall follow the special considerations below, for datasets that may be useful to AI R&D.

Special Enterprise Data Inventory (data.json) Considerations for AI R&D Datasets				
Field	*Label*	*Definition*	*Required*	*AI R&D Guidance*
keyword	Tags	Tags (or keywords) help users discover your dataset; please include terms that would be used by technical and non-technical users.	Always	Agencies shall include the keyword of **"usg-artificial-intelligence"** for all datasets determined to support AI R&D. Datasets that specifically serve as training data for ML applications should additionally include a keyword of "**usg-ai-training-data,**" "**usg-ai-testing-data,**" and other keywords can be developed and used as appropriate, but

[1] https://project-open-data.cio.gov/v1.1/schema/

				check back on the Project Open Data site or resources.data.gov where a coordinated list of keywords will be maintained.
identifier	Unique Identifier	A unique identifier for the dataset or API as maintained within an Agency catalog or database.	Always	Consistent with existing guidance for this field and the National Science Foundation's May 2019 Dear Colleague Letter, it is highly recommended that this is a globally unique, resolvable, persistent identifier. This will aid in dataset citation, discoverability, and provenance between a derived product and source data. This is particularly important when logic is derived from machine learning using the source data.
rights	Rights	This may include information regarding access or restrictions based on privacy, security, or other policies. This should also serve as an explanation for the selected "accessLevel" including instructions for how to access a restricted file, if applicable, or explanation for why a "non-	If-Applicable	Agencies shall provide instructions for how researchers may access sensitive data.

		public" or "restricted public" data asset is not "public," if applicable. Text, 255 characters.		
dataQuality	Data Quality	Whether the dataset meets the agency's Information Quality Guidelines (true/false).	No	Agencies shall provide whether datasets explicitly indexed for AI R&D have met their agency's Information Quality guidelines.
isPartOf	Collection	The collection of which the dataset is a subset.	No	When providing disaggregate datasets, e.g. by year or state, agencies shall group these datasets into a master collection. Additionally, relationship between the dataset and bigger enterprise-wide data assets. An Entity-Relationship-Diagram (ERD) references are encouraged. - Describe if an 1:M or M:1 or M:M type data asset and include dimensions or fact type table
references	Related Documents	Related documents such as technical information about a dataset, developer documentation, etc.	No	Agencies shall provide any additional information related to the dataset to assist AI R&D. Examples would be Include Data Dictionaries, Entity

				Relationship Diagrams, Data Lineage maps

Guidance for Improving Discoverability of Agency Models for AI R&D

The AI Executive Order references source code as a mechanism for providing data and models. Code.gov allows for discoverability of federal source code by searching AI related keywords. To increase discovery for Federal AI R&D models, agencies shall inform their various offices/department heads associated with R&D that models being developed (or procured) should be captured in Source Code Inventories. To ensure agencies provide a comprehensive inventory of models for AI, agencies are encouraged to search their web properties as well as their code sharing and version control platforms for common search terms,[2] as well as use NGO databases such as Algorithmtips.org. In addition to the minimal metadata required for agency Source Code Inventories[3], agencies shall follow the special considerations for datasets that may be useful to AI R&D below:

Special Source Code Inventory (code.json) Considerations for AI R&D Models				
Field Name	*Data Type*	*Definition*	*Required*	*AI R&D Guidance*
tags	array	An array of keywords that will be helpful in discovering and searching for the release.	Always	Agencies shall include the keyword of **"usg-artificial-intelligence"** for all source code determined to support AI R&D. Other keywords can be developed and used as appropriate, but check the Project Open Data site or resources.data.gov where a coordinated

[2] Example search terms from Algorithmtips.org include: "Algorithm, Algorithmic, Automated analysis, Automated assessment, Automated calculation, Automated filtering, Automated grading, Automated ranking, Automated rating, Automated scoring, Automated simulation, Automated sorting, Automatic assessment, Automatic calculation, Automatic filtering, Automatic grading, Automatic ranking, Automatic rating, Automatic score, Automatic scoring, Automatic sorting, Calculating matrix, Calculating method, Calculating model, Computation, Computational, Computing, Grading calculation, Grading equation, Grading formula, Grading matrix, Grading method, Grading methodology, Grading model, Numerical rating, Predictive Analytics, Predictive modeling, Ranking calculation, Ranking equation, Ranking formula, Ranking matrix, Ranking method, Ranking methodology, Ranking model, Rating calculation, Rating equation, Rating formula, Rating matrix, Rating method, Rating methodology, Rating model, Scoring calculation, Scoring equation, Scoring formula, Scoring matrix, Scoring method, Scoring model, Statistical assessment, Statistical methodology, Statistical model, Statistical software"

[3] https://code.gov/about/compliance/inventory-code

				list of keywords will be maintained.
exemptionText	string or null	If an exemption is listed in the 'usageType' field, this field should include a one- or two- sentence justification for the exemption used.	No	Agencies shall describe how researchers may be able to access governmentWideReuse or exempt data
relatedCode	array	An array of affiliated government repositories that may be a part of the same project. For example, relatedCode for 'code-gov-front-end' would include 'code-gov-api' and 'code-gov-api-client'.	No	Agencies shall describe related models and code.

Guidance for Improving Usability of Federal Data and Models

Usability of Federal Data and Models is determined in the context of their utility to the non-Federal AI R&D community. Once a non-Federal users discovers a Federal data set or model applicable to them, the next step is an initial assessment of usability. Given the wide range of potential non Federal AI R&D use cases and the associated wide range of documentation and quality of data and models, it is difficult to standardize on criteria for usability that is applicable across all potential use cases. However, based on feedback from Agencies the following should be considered as initial factors to address in determining dataset and model useablity:

1. Ability to search by scope (as in type of algorithm, applications, sample size needed, etc)
2. Ability to link to instances of model with parameters, data used
3. Ability to label usability of AI tool level based on private sector or academia feedback.
4. Licensing of model.
5. Ability to make testing discoverable after the training model is built on training set.
6. Ability to label appropriate uses.
7. Ability to label safety, security and confidentiality protections.

As these 7 initial factors are not part of existing meta-data standards, Agencies are encourage to include in “readme” files associated with the data sets and models.

Strategy 8 in the National Artificial Intelligence Research & Development Strategic Plan: 2019 Update calls for establishing public-private partnerships to accelerate advance in Artificial Intelligence. In line with this strategy, Federal Agencies are asked to establish capabilities to continually solicit, collect and adjudicate user feedback on their data and models as well as implement capabilities to iteratively update data and models to address user feedback. In establishing these mechanisms, agencies should consider the metrics necessary to make resource

allocation decisions in enhancing their data and models as well as outcome based measurements to capture the impact of data and model enhancements to the non-Federal AI R&D community. Models should continually be streamlined by using newer technologies/processes that emerge in the industry, to keep pace with technology as well as improve accuracy. For example, especially important with Natural Language Processing, new libraries are being built and enhanced daily that can improve text mining outputs drastically.

Agency's initial focus should be on establishing feedback mechanisms for Federal data sets and models and to provide a data set "Versioning" capability so that it is clear to users what enhancement have been made to a dataset and model and their meta-data. This practice has been in place for some time for model development, though less so with data sets. The Data.gov team will develop and promulgate best practices, consistent with World Wide Web Consortium (W3C) Data Catalog Vocabulary (DCAT) vocabulary with respect to data set versioning. Code.gov and Data.gov will provide a discoverable "link" between Federal source code developed for AI R&D purposes and data used in the development on data.gov from their web-sites. For Natural Language Processing applications, Agencies are encouraged to use Application Programming Interfaces (APIs) to read from external Natural Language Processing (NLP) libraries to train the models for efficiency. As well as making the agency models available for user feedback to accommodate concepts pertinent to agency specific subject matter which might not be available in public NLP libraries.

Agencies should consider use cases from the non-Federal AI R&D community and implement in sprints as a way to make incremental progress on data and model discovery and usability. Sprints are the desired way to discover/create ecosystems of datasets and models for AI R&D purposes. Agencies should implement mechanisms to share lessons learned from the sprints they participate in across different teams.

Governance

The Office of the Federal Chief Information Officer will oversee the guidance associated with Federal data and model discoverability and usability as related to AI R&D, in consultation with the NSTC Machine Learning and AI Subcommittee (MLAI-SC). Activities will be tracked under Community Action 9 of the Federal Data Strategy, "Improving data resources for AI R&D". The OMB Data Council to be established as part of the Federal Data Strategy and the CDO Council to be established as required by the Foundations for Evidence Based Policy Act, will act as the bodies to share lessons learned and provide any future incremental guidance to agencies in improving discoverability and usability of Federal data and models for AI R&D, in consultation with the MLAI-SC.